The role of semantic encoding in production-enhanced memory: A registered report

Tanja C. Roembke and Rachel M. Brown
Cognitive and Experimental Psychology
Institute of Psychology, RWTH University, Aachen, Germany

Author Note

Rachel M. Brown https://orcid.org/0000-0002-4851-8429
Tanja C. Roembke https://orcid.org/0000-0003-3932-1488
We have no known conflict of interest to disclose.

Correspondence concerning this article should be addressed to:
Tanja C Roembke, RWTH University, Institute of Psychology, Jaegerstrasse 17-19, Building 6011, Room 207, 52066, Aachen, Germany. +49 241 80 96558
Email: Tanja.Roembke@psych.rwth-aachen.de

Abstract: 248 words
Manuscript (total length; excluding references): 9.983 words
Figures: 1
Tables: 1 (appendix)

Acknowledgments

We thank Maria Bruggaier and Alina Mummenhoff for her assistance in preparing stimulus materials.

Abstract

Words that are read aloud are recognized and recalled more accurately than words that are read silently (the production effect). The production effect is a robust memory phenomenon that has been found with a range of materials and manipulations. Nevertheless, mechanisms underlying the production effect are still unclear, possibly because speaking may engage different linguistic representations. A recent study reports that the production effect was reduced but not eliminated when semantic recognition was disrupted, suggesting a role of semantic encoding in the production effect. In line with this, we hypothesize that production increases spreading activation from proximate orthographic and phonological representations to more remote semantic ones. For bilinguals, activation may then also spread to orthographic and phonological representations in the different language, consistent with the idea that semantic representations are shared across languages. If production enhances semantic encoding in this way, the production effect should not only be reduced when semantic recognition is disrupted, but it should also persist when semantic recognition is favored. The goal of the proposed study is therefore to test this prediction in two experiments by manipulating how items are presented at recognition. We suggest that if production enhances semantic encoding of written words, then it should be possible to recognize these words later by their corresponding pictures or translations. Thus, we predict that a production effect should be observed even if recognition items are presented as pictures or translations, but it should reduce in this case if it relies on multiple linguistic representations.

Keywords: cognition, memory, production effect, language, encoding

Word count: 248

The role of semantic encoding in production-enhanced memory: A registered report

Performing a movement during encoding can improve how easily information is retained. For example, it is easier to remember a phone number after repeating it aloud (MacLeod et al., 2010), task instructions after enacting them (Allen et al., 2020) or an image after drawing it (Fernandes et al., 2018). One such well-documented memory phenomenon is the production effect: Words that are read aloud are recognized and recalled more accurately than words that are read silently (Conway & Gathercole, 1987; Gathercole & Conway, 1988; Hopkins & Edwards, 1972; MacLeod et al., 2010; Murray, 1965). In a typical experiment, participants first study a list of words, half of which they have to read aloud and half of which they have to read silently in a random order. In a later recognition test, a list of words is presented that includes words that were read aloud, silently or not at all. Participants are asked to indicate if they have previously seen a word or not. In this design, participants are found to be more likely to correctly recognize or recall a word if it was read aloud versus read silently. The production effect is a robust memory phenomenon that has been observed for words, sentences, and longer written texts (Forrin et al., 2012, 2014; Forrin & MacLeod, 2018; Icht et al., 2019; MacLeod, 2011; MacLeod et al., 2010, 2022; Mama & Icht, 2016; Ozubko et al., 2012; Ozubko & MacLeod, 2010). It has been shown when words are only mouthed silently (e.g., MacLeod et al., 2010), written (Mama & Icht, 2016) or sung (Quinlan & Taylor, 2013, 2019). It can be observed not only for long-term memory but also short-term memory (Saint-Aubin et al., 2021). A production effect can be observed within-subjects and, to a lesser extent, between-subjects (Fawcett, 2013; Fawcett & Ozubko, 2016).

Despite the production effect's observed replicability and generalizability, its underlying mechanisms are still being uncovered. One of the challenges in accounting for the effect may be that the act of speaking (oneself) has many possible associations in long-term memory, including the activation of a word's phonological (sound), orthographic (written form) and semantic (meaning) representations. Explanations for the production effect have often focused on whether distinguishing (i.e., distinctive) features of spoken items can improve encoding or retrieval (Conway & Gathercole, 1987; Dodson & Schacter, 2001; Fawcett & Ozubko, 2016; Gathercole & Conway, 1988; Hunt, 2003). The production effect has been simulated by formalized models of working memory and recognition memory that represent episodes as collections of features; these include the retrieving effectively from memory model (REM; Shiffrin & Steyvers, 1997), the multiple-trace simulation model (MINERVA 2; Jamieson et al., 2016), and the revised feature model (Cyr et al., 2022; Saint-Aubin et al., 2021). Within these models, memory traces for events (i.e., items) are stored as collections of features in short and long-term memory, and as the amount or integrity/quality of different features within a memory trace increases, the robustness and/or retrievability of that memory trace improves. Feature-enrichment may improve encoding by reducing confusability among items (Saint-Aubin et al., 2021), and it may improve retrieval by increasing the likelihood of matching a memory cue to the correct memory trace. An open question remains: what kind of features are included as a result of speaking? Some modelling approaches assume that speaking should only engage sensorimotor features, or "modality-dependent" features (e.g., Saint-Aubin et al., 2021; Wakeham-Lewis et al., 2022), while others include the possibility that speaking engages other linguistic features such as semantics, or "modality-independent" features (Jamieson et al., 2016), which has been supported by recent work (Fawcett et al., 2022). In this study we address this question by further investigating the contribution of semantic encoding in production-enhanced memory.

Spreading Activation and Production

When we comprehend or produce language, the words we perceive or produce have a range of representations or features, including written forms, sounds, or speech movements, any of which may be strongly associated with word meaning. These associations should allow any feature of a word to prime its conceptual referent. An explanation for such associative priming is provided by the notion of spreading activation, which is an important explanatory construct in many theories of memory and cognition (e.g., Collins & Loftus, 1975). In a typical spreading activation model, distinct concepts are represented as separate nodes and relationships among concepts are associative pathways between nodes (Balota & Lorch, 1986). Distance between nodes is a function of how strongly associated they are. When part of such a network is “activated” (for example when a word is read or remembered), activation automatically spreads along the associations between nodes to related areas in memory. As a result, related areas become more available for further cognitive processing, including easier retrieval. Previous work has suggested several properties of how information spreads within a memory network. For example, activation is more likely to spread from one node to another if they are strongly associated than when they are not (Lorch, 1982). Critically, activation spread is not limited to two directly associated concepts but rather may expand across multiple steps within a network (Balota & Lorch, 1986). Similarly, for bilinguals, activation may also spread to orthographic and phonological representations in the different language, consistent with the idea that semantic representations are (largely) shared across languages (de Groot, 1992; Glanzer & Duarte, 1971; Kroll & Stewart, 1994; van Hell & de Groot, 1998).

The assumption that activation can spread across multiple steps has been critical in accounting for many different memory phenomena: This includes short-lived semantic priming, where performance on a current trial is impacted by performance on previous trials with semantically related concepts. Semantic priming is often studied in the context of lexical processing: For example, it is easier to decide that the word CAT is a real word or not after having been presented the word DOG in a lexical decision task (McNamara, 1992). When first seeing DOG, activating spreads to related concepts such as CAT (another mammal that is often kept as a pet), which then leads to facilitation. Spreading activation has also been used to explain more complex memory phenomena such as performance on the Deese-Roediger-McDermott (DRM) task (e.g., Deese, 1959; McDermott, 1996; Roediger & McDermott, 1995). Here, participants are first presented with a list of highly related concepts (e.g., blanket, doze, slumber, bed, etc.) (Meade et al., 2007). Importantly, these lists are all semantically clustered around one concept (e.g., sleep) that is not actually included in the to-be-encoded list. Subsequently, participants are then tested whether they remember seeing previously presented words as well as the related critical lure (e.g., sleep). Participants typically recognize the lure word with high probability and confidence (Pardilla-Delgado & Payne, 2017). It has been argued that lures are recognized in the DRM task due to spreading activation in semantic memory at encoding: As a result, at retrieval, strongly activated lures may be misattributed to having occurred in the original list (Johnson et al., 1993; Meade et al., 2007). The DRM false memory task also been used to investigate the extent to which activation spread occurs across known languages. For example, a couple of studies investigated whether switching from one language at encoding to another at retrieval hurts memory (Suarez & Beato, 2021). It was found that even if the language of the presented items differed, participants falsely remembered non-presented foils at test

(Marmolejo et al., 2009; Miyaji-Kawasaki et al., 2004; Sahlin et al., 2005), suggesting that activation spreads between languages.

Spreading activation across multiple steps in a memory network was tested directly in a study by Balota and Lorch (1986). Here, stimuli materials consisted of triads, where two words were always directly related: For example, LION (Word 1) and TIGER (Word 2) are part of the same semantic category: big cats). The third word, however, was only directly related to one of the other two words. In our example, the word STRIPES (Word 3) is directly related to TIGER (Word 2; tigers have stripes) but not related with LION (Word 1; lions do not have stripes). Nevertheless, a connection between LION and STRIPES may still exist indirectly, via TIGER (e.g., STRIPES is related to TIGER which in turn is related to LION). This then allowed the researchers to test whether activation from word 1 did not only facilitate the activation of Word 2 but also Word 3. Interestingly, evidence for multi-step activation spread was observed in a primed speeded production task. Performance was speedier if participants were primed with a directly or indirectly related word versus as neutral one (e.g., BLANK). More recently, work with computational networks (e.g., Kenett et al., 2017) has corroborated the idea that network path lengths predict decision times for directly and more distantly related word pairs. Overall, these findings provide compelling evidence that activation can spread across multiple steps within an associative memory network. However, we do not yet know which other factors—beyond associative strength of two concepts—impact how far information spreads within a network. This is the knowledge gap we are addressing in this study.

A possibility is that language production influences the extent of spreading activation. It has been hypothesized that activation spreads during sentence production (Dell, 1986; Dell & Chang, 2014), where for example slips of the tongue are seen as evidence for activation spread resulting in the production of an incorrect word. When a word is read, activation is also assumed to spread from the word's orthographic representation to its semantic and phonological ones, (Coltheart & Rastle, 1994; Harm & Seidenberg, 1999; Seidenberg, 2005; Ziegler et al., 2008)(Coltheart & Rastle, 1994; Harm & Seidenberg, 1999; Seidenberg, 2005; Ziegler et al., 2008)(Coltheart & Rastle, 1994; Harm & Seidenberg, 1999; Seidenberg, 2005; Ziegler et al., 2008)thus accessing a word's sound and meaning (Coltheart & Rastle, 1994; Harm & Seidenberg, 1999; Seidenberg, 2005; Ziegler et al., 2008). How does this differ between reading silently versus reading out loud? In general, it appears that reading aloud and reading silently are related constructs that rely on shared developmental cognitive mechanisms (van den Boer et al., 2014). Having said that, previous research has often looked at perception and production separately, so that there currently is relatively little research on this. One possibility is that production increases spreading activation from orthographic and phonological representations of words, which may be more proximate or directly related to the visual input (especially in alphabetic languages), to more remote semantic ones, which may be more distally related to the word form. The notion that semantic representations are less directly or more distally related to production is suggested by evidence from reading development, where direct associations between orthography and semantics are acquired later than associations between orthography and phonology (e.g., Seidenberg, 2005; Ziegler et al., 2008). Thus, spreading activation may enable production to prime semantic representations of read-aloud items. If this is the case, could such activation spread (potentially via multiple steps) facilitate later retrieval? Consistent with this hypothesis, there is some research that suggests that students' text comprehension is improved if they read aloud versus read silently, especially for beginning and/or struggling readers (Robinson et al., 2019, but also see McCallum et al., 2004; Schimmel & Ness, 2017). In addition, in a recent study by Tsuboi et al. (2021), the impact of reading aloud vs. silently on priming was investigated (Experiment 2). They found that priming was increased for the read-aloud condition, consistent with the idea that activation spread is more extensive when reading aloud rather than reading silently.

If production enhances semantic encoding, the production effect should be reduced if semantic recognition is disrupted. Evidence for this comes from a recent study by Fawcett et al. (2022), where the impact of semantic encoding on the size of the production effect was investigated. As a variant of the standard paradigm, participants always saw two words simultaneously during the recognition test: the target word and a lure. In Experiment 1, the lure could be either a homophone (e.g., BEAR) of the target (e.g., BARE) or not (e.g., MERRY). In Experiment 2, the lure was a synonym (e.g., POISON) of the target (e.g., VENOM) or not (e.g., ETHICS). The production effect was found to be reduced with synonyms but not homophones. The results are consistent with the hypothesis that semantic encoding contributes to the production effect. This finding could be accounted for by the notion that production enhances the extent of spreading activation, and thereby enhances semantic encoding. Importantly, semantic encoding is here seen as contributing to the production effect, but not its sole cause, as the production effect can also be observed in the absence of existing full-fledged semantic representations (i.e., for nonwords; MacLeod et al., 2010).

The Current Study

If production enhances semantic encoding, the production effect should not only be reduced when semantic recognition is disrupted (as observed by Fawcett et al., 2022), but it should also *persist* when semantic recognition is favored. From a spreading activation perspective, if articulating a word (production) can prime the meaning (semantics) associated with that word (theoretically by increasing the extent of spreading activation), then this semantic priming may enable a production effect (memory enhancement for spoken compared to silently read words). We will investigate this explanatory hypothesis by examining whether the production effect can be observed when previously-articulated items match items at recognition on semantic features but not others (semantic recognition). Specifically, we test the novel prediction that the effect should be observed when a previously studied written word has to be later recognized as a picture or translation. In addition, we hypothesize that the production effect does not rely exclusively on semantic encoding. Articulating a word should not only prime its semantic features, but it should also prime its orthographic and phonological ones. If this is the case, the production effect should be greatest when recognition items are identical to those presented at learning (veridical recognition). We will investigate this additional hypothesis by comparing how production influences two types of recognition: semantic versus veridical item recognition.

Two experiments will test two independent groups of German-English bilinguals, all with German as a first language and English as a second language. Participants will study a list of German written words (learning task) and will subsequently be asked to recognize the words they had studied (recognition task). Each participant will complete both the learning task followed by a recognition task twice. In each of the two learning tasks, participants will be presented with written words in their first language (German) on a screen one at a time. They will read some of the words silently and they will read some words out loud. One of the learning tasks will be followed by a veridical recognition task, which will ask participants to recognize words presented in the same form as they were presented at learning (recognition targets will be the same written words that were presented at learning). The other learning task will be followed by a semantic recognition task. In Experiment 1, the semantic recognition task will present pictures that correspond to previously-studied words (targets) or pictures that do not correspond to any of the studied words (foils). In Experiment 2, the semantic recognition task will present written words in participants' second language (English) that are either translations of previously-presented German words (targets) or translations of words that were not studied (foils). In both experiments, participants will be asked to respond "yes" or "no" according to whether or not they recognize each word, picture, or translation.

Arguably, activation spread has to be more extensive for a production effect to be observed when recognizing translations (Experiment 2) compared to pictures (Experiment 1). That is, a match between a previously read word and an item at test may be more easily established for pictures because there is a more direct link between the item at encoding and recognition (e.g., the picture of a beetle and the word BEETLE). However, for a translation to be activated during encoding, activation has to travel further steps (from the L1 word to the semantic representation to the L2 word), unless one presumes additional direct connections between L1 and L2 words. In fact, such direct connections have been suggested in models of the bilingual lexicon. For example, one classic model, the Revised Hierarchical Model (RHM), proposes direct links between L1 and L2 words as well as separate links from L1 and L2 to shared semantic representations (Kroll & Stewart, 1994). As a result, activation during production could travel from L1 to L2 both via the direct word-to-word links as well as indirectly via semantic representations (or both). Interestingly, the RHM also presumes stronger links from L2 to L1 words than the other way around in non-simultaneous bilinguals where one language was acquired before the other. In the beginning of second language acquisition, a newly learned L2 word is linked with its L1 word before establishing direct links with the semantic representations and links from L1 to L2. Evidence for this view comes from the finding that backward translation (i.e., translating words from L2 to L1) is easier than forward translation (i.e., translating words from L1 to L2; Kroll & Stewart, 1994). The RHM would predict that activation flow is stronger from L2 to L1 words than the other way around. As such, presenting words in the L1 and then testing them in the L2 represents a more conservative test of activation spread via semantic representations (vs. direct links between L1 and L2 word representations) as a result of production.[1] We cannot exclude the possibility that some activation will also spread via direct links between words. This possibility is still consistent with the hypothesis that production increases activation spread, though would speak against the importance of semantic processing per se.

Based on the idea that the production effect should persist when items are matched in semantic but not other features at learning and recognition, we expect to observe a production effect (greater recognition accuracy for spoken compared to non-spoken words), both when recognition items are presented as pictures or translations (semantic recognition condition), and when recognition items match those at learning (veridical recognition condition: the same written words are presented at learning and recognition). In addition, based on the idea that the production effect does not rely exclusively on semantic encoding, we also expect the production effect to reduce in semantic recognition conditions relative to veridical conditions in which words are matched on multiple linguistic features. We will test the following specific predictions:

1) Experiment 1: Recognition accuracy will be greater for words that were spoken compared to those that were silently read (a production effect), both when participants recognize pictures (Prediction 1A) (semantic recognition condition), and

[1] Another argument for presenting words in the L1 during encoding (rather than L2) is that it will facilitate comparison of results across this study's experiments.

when participants recognize words in the same written form as they were presented at learning (Prediction 1B) (veridical recognition condition).

2) Experiment 1: The increase in recognition accuracy from having spoken compared to having silently read words (production effect: spoken > silent) will be larger when participants recognize the same written words (veridical recognition condition) than when they recognize pictures (semantic recognition condition) (Prediction 1C).
3) Experiment 2: Recognition accuracy will be greater for words that were spoken compared to those that were silently read (a production effect), both when participants recognize translations of the words they had studied (semantic recognition) (Prediction 2A), and when participants recognize the same words (in the same language) they had studied (veridical recognition) (Prediction 2B).
4) Experiment 2: The increase in recognition accuracy from having spoken compared to having silently read words (production effect: spoken > silent) will be larger when participants recognize the same words (in the same language) they had studied (veridical recognition condition) than when they recognize translations of words they had studied (semantic recognition condition) (Prediction 2C).

If we observe that the production effect is present (greater recognition accuracy for spoken words compared to silently-read words), both when participants are asked to recognize pictures (Exp. 1) or translations (Exp. 2) corresponding to the words they had studied, and when they are asked to recognize the same written words they had studied, this would suggest that the production effect persists when words that were studied can be recognized on their semantic features, and that production may influence semantic encoding. This outcome would be consistent with the idea of spreading activation: speaking (e.g., articulatory features) could engage modality-independent associations with semantic features, even if those associations are indirect (i.e. mediated by other, stronger associations such as motor-to-sensory associations). To support these interpretations, it would be additionally important to show that memory performance is above-chance in the semantic recognition conditions, particularly in the spoken condition, to help rule out the possibility that a large production effect was caused by poor overall memory performance in the semantic conditions (i.e., floor effects in the semantic silent condition, see below).

If we do not detect a production effect (contrary to our prediction) when participants are asked to recognize pictures or translations (semantic recognition), this would raise the possibility that production may have little or no influence on semantic encoding, but this interpretation would need to be more directly tested with further analyses. However, this outcome would strongly align with the assumption that speaking adds only (or mainly) modality-dependent features to memory traces, and not modality-independent features (such as semantic features) (e.g., Saint-Aubin et al., 2021; Wakeham-Lewis et al., 2022). If a production effect is not detected in the semantic conditions, and we observe overall reduced memory performance in the semantic conditions, this pattern would be predicted by a transfer-appropriate framework, where performance on any memory test is better if conditions at study and test match (Morris et al., 1977), as is the case for the veridical, but not the semantic conditions. Here too the theoretical interpretations are aided by additionally showing that memory performance in the semantic conditions are above chance, to rule out the possibility that a reduced or undetectable production effect was caused by overall poor memory in these conditions (see below).

If we observe, as we predict, that the production effect is present but decreases when participants recognize pictures (Exp. 1) or translations (Exp. 2) compared to when they recognize items that match those presented at learning, this would suggest that production may influence not only semantic encoding but other linguistic features as well. This outcome would align with spreading activation, and with memory models that assume that speaking can engage modality-independent features (e.g., Jamieson et al., 2016). This outcome could also be fit to attenuated or modified versions of alternative accounts. For example, speaking may engage modality-dependent features more strongly than modality-independent features; modality-dependent features may more easily bind to veridical stimulus features (written words) resulting in better veridical recognition. In addition, transfer-appropriate processing may modify retrieval success, such that memory can improve when there is some degree of similarity between processing at encoding and retrieval. These interpretations would be supported by additionally showing that memory performance is above-chance in the semantic conditions, at least in the semantic spoken conditions, to rule out that a reduced production effect is only due to a floor effect (see below).

If, contrary to our prediction, we do not detect a difference in the production effect as a function of how items are presented at recognition, it raises the possibility that semantic encoding may be sufficient for the production effect, but this would have to be examined with further analyses. However, as long as the presence of the production effect (spoken > silent words) can be shown in the semantic conditions, this outcome would still run counter to or call into question the assumption that speaking only engages modality-dependent features, and it would run counter to the idea of transfer-appropriate processing. Here too it will be helpful to additionally observe above-chance performance in the semantic conditions to help rule out floor effects as the cause of the production effect (see below).

Finally, if again contrary to our prediction, we observe a larger production effect when recognition items are presented as pictures or translations, this would suggest that production could enhance the encoding of semantic features relative to other linguistic features. This latter pattern would contradict the notion that articulation should have a selective effect on phonological encoding (encoding speech sounds) due to motor-phoneme associations (see Fawcett et al., 2022). In other words, this outcome would strongly contradict the assumption that speaking only engages modality-dependent features, as well as transfer-appropriate processing. These interpretations would further depend on ruling out a larger production effect due to floor effects in the silent condition (see below). This result would be comparable to a levels of processing effect, in which engaging semantic features improves memory (e.g., Jamieson et al., 2016). This outcome would also strongly align with memory models that assume that speaking can engage modality-independent features. As such, our study will not only have implications for research on the production effect but also on linguistic theories of production.

Quality Control

Two important quality checks are inherent to the experimental designs described above. First, the veridical conditions in each experiment serve as control or baseline recognition conditions, because they 1) implement the classic production effect paradigm using written words in participants' native language at both learning and recognition, and they 2) should minimize overall recognition difficulty compared to the semantic conditions. It is reasonable to assume that semantic recognition could be more difficult on average regardless of whether words were spoken or silently read compared to veridical recognition, because participants will be required to associate the learned stimuli (written words) to novel stimuli (pictures or translations). This presents the possibility of floor effects, or an artifact from the difficulty of the semantic condition. One possibility is that we may not be able to observe a production effect in the semantic conditions if participants are not able to match the pictures and translations to the words they saw during learning (e.g., they may remember that they saw the word “ostrich” but then mistake the picture of an “ostrich” to be an emu), or if this performance is highly variable. Another possibility is that if the semantic condition is too difficult, it may impair the silent semantic condition to such an extent that the production effect appears larger in the semantic condition, or it may be easier for speaking to enhance memory performance that is overall lower. To determine whether a decreased or increased production effect in the semantic condition would be due to floor effects, three strategies will be used. First, overall recognition in the veridical and semantic conditions will be compared (this would appear as a main effect in the ANOVA). If the semantic condition is too difficult, recognition for both spoken and silent conditions should be greatly reduced compared to those of the veridical condition, regardless of whether the production effect is present or not in the semantic condition. Second, to test for above-chance performance, post-hoc t-tests will be completed for each condition separately. If a production effect is observed in the semantic condition, above-chance recognition in at least the spoken condition will be taken as some evidence against an artifact or floor effect, while at or below chance performance in both spoken and silent semantic conditions will be taken as evidence for the presence of an artifact/floor effects, in which case any *change* in the production effect between semantic and veridical conditions will be interpreted cautiously. In either case, if the presence of a production effect is still observed in the semantic conditions, we would still not rule out the possibility that articulating words somehow promoted the ability to recognize items on the basis of their meaning, however difficult this might have been. However, if a relatively larger or smaller production effect is seen in the semantic condition, and all semantic conditions are at or below chance, we would attribute the production effect change to a possible floor effect. Finally, we will look particularly at whether performance in the spoken semantic condition is numerically at least as high as the performance in the veridical silent condition. Even if the semantic condition is overall more difficult, it should not be *too* difficult if participants can achieve at least the level of recognition accuracy as they do in the veridical silent condition (which we assume has the minimal conditions necessary for successful memory performance). In general, interpretation will be guided by the overall pattern observed across the conditions. Even if performance in all conditions is above-chance, we will not completely rule out floor effects if there is a pronounced performance drop between veridical and semantic conditions. Moreover, if numerically near-perfect performance is observed in veridical conditions (e.g., above 95%, particularly veridical spoken conditions) we will entertain the possibility of ceiling effects as a cause of a production effect change between veridical and semantic conditions (e.g., a decreased production effect in the veridical conditions). Finally, regardless of what we observe in the semantic conditions, if we observe a production effect in the veridical conditions (only Predictions 1B and 2B), we can assume that participants are able to demonstrate the classic production effect.

A second quality check is provided by our decision to examine two types of stimuli that can be matched to studied items on semantic but not other features (pictures or translations) across two different experiments with two independent groups of participants. If we observe a production effect in the semantic condition in each experiment, as we predict (Predictions 1A and 2A), this will further corroborate the assumption that articulation can promote semantic recognition, even if semantic recognition is more difficult overall than veridical recognition. If we observe a production effect in one semantic condition but not the other (e.g., only with pictures but not with translations), we will still have evidence that articulation may promote semantic recognition. Again, we will be able to see whether the condition in which a production effect was not detected was also overall more difficult compared to the veridical condition. Furthermore, if we do not detect any production effect in either semantic condition, even though it will not be possible to infer a lack of a production effect using our planned statistical approach (see below), the failure to detect the effect in two different groups with two different manipulations can potentially be of theoretical interest: it would raise the possibility that either the effect is not there or is at least greatly reduced when there are no features besides semantics available at recognition. This could call into question our hypothesis that articulation primes semantics via spreading activation. If we do observe this pattern, we will recommend further replications that additionally employ Bayesian statistics to examine the evidence for equivalence between speaking and silently reading in semantic recognition conditions.

Finally, our procedures for stimulus selection, participant selection, and measurement (see below) will provide further quality control. Our stimulus selection parameters will be aimed at increasing the likelihood that participants will be able to associate items at recognition with items presented at learning. For example, we will select stimuli for semantic conditions (pictures or translations) on the basis of how frequently previous participants associated particular words with particular pictures, or on the basis of the translated word's frequency in participants' second language. We will also select participants with a minimum level of second-language proficiency, to increase the likelihood that they can recognize words that are translated into their second language. In addition, participants will be recorded via a microphone on all trials of the learning tasks, to ensure that they speak and silently read as the tasks instruct.

Experiment 1

Method

Participants

To determine the planned sample size, a-priori power analyses were performed based on what we considered to be the smallest plausible effect size of a predicted interaction between the production effect (spoken > silent) and the semantic versus veridical recognition conditions in a 2 (production: spoken/silent) by 2 (recognition: semantic/veridical) within-subject ANOVA. We focused particularly on the expected size of the interaction effect, because we expect the interaction to have the smallest effect size among our predicted outcomes. We expect the main effect of production (spoken > silent) to be moderate to large. Previously-reported main effects can vary from about $\eta^2_p = 0.19$ (Kaushanskaya & Yoo, 2011) to $\eta^2_p = 0.28$ (Mama & Icht, 2016), to $\eta^2_p = 0.32$ (Fawcett et al., 2022) to $\eta^2_p = 0.38$—0.60 (Forrin et al., 2012; MacLeod et al., 2010; Ozubko et al., 2012) to $\eta^2_p > 0.60$ (Brown & Roembke, 2024; Cho & Feldman, 2016). Reported effect sizes for interactions between the production effect and other factors (e.g., language, delay between learning and test, blocked or interleaved speaking/silently reading) have been smaller, as would be expected (see for example, Cho & Feldman, 2016, $\eta^2_p = 0.18$; Fawcett et al., 2022, $\eta^2_p = 0.05$, Ozubko & MacLeod, 2010, $\eta^2_p = 0.14$; Ozubko et al., 2012, $\eta^2_p = 0.08$, Brown & Roembke, 2024, $\eta^2_p = .21$ and 0.06). We decided to base our estimate of the expected interaction effect size primarily on the results reported by Fawcett et al. (2022) because, among studies that report 2-way interactions involving the production effect, this is the only study we are aware of that also examines the interaction between the production effect and a manipulation of *semantic recognition*. In addition, because the particular task design is based on that of our previous work (Brown & Roembke, 2024) and differs from that of Fawcett et al. (2022), we also used our previous data as a second basis for estimating our expected effects.

We conducted a simulation-based power analysis using the SuperPower Shiny app (Lakens & Caldwell, 2021) with 10000 simulations, alpha = 0.05, assuming a common correlation of *r* = 0.5 among within-subject factors, and assuming a 2 (production: spoken/silent) by 2 (recognition: semantic/control) within-subject design and using the previously-reported means and SDs in each cell (Fawcett et al., 2022). This analysis suggested that 75 participants are needed to achieve 100% power for a production main effect in the ANOVA, 97% and 100% power for planned comparisons between spoken and silent conditions within semantic and control conditions, respectively, and 82% power for the 2-way interaction. In addition, to account for uncertainty in the previously-reported effect size estimates, we conducted an additional power analysis with adjusted cell means, based on the upper and lower confidence-interval boundaries around the means reported by Fawcett and colleagues (2022). We first computed the 95% confidence intervals around each cell mean from the reported standard errors. We then performed a second simulation-based power analysis using the largest values at the upper boundaries of the silent conditions, and the lowest values at the lower boundaries of the spoken conditions in their design, which yields a conservative estimate of the production effect in each recognition condition. The same standard deviations were used for each cell mean. Using alpha = 0.05, 10000 simulations, a common correlation of *r* = 0.5 among the factors, and the same 2 × 2 within-subject design, the analysis suggested that 75 participants would be needed to achieve 87% power for the 2-way interaction and 92% power for the planned comparison in the control condition (spoken > silent). However, the power for the main effect and for the spoken>silent planned comparison in the semantic condition decreased to 34% and 18%, respectively. It should be noted that the adjusted cell means using these upper and lower boundaries *reversed* the production effect (numerically) in the semantic condition, which may explain the reduced power. We then conducted the same power analysis with *N*=80, *N*=90, and *N*=100 (all other parameters were kept the same), and we observed similar levels of power: at least 85% for the 2 × 2 interaction and at least 90% for the planned comparison in the control condition, but low power for other analyses (about 37—44% for the main effect, and about 18—22% for the planned comparison in the semantic condition). Thus, increasing the sample size from 75 did not appear to substantially alter power for the conservative effect size estimate. In addition, we think a reversed production effect (silent>spoken) is unlikely, given the consistency of the effect across various replications. Thus, we deemed a sample size of *N*=75 to be sufficient to detect the plausible effect sizes (those based on Fawcett et al., 2022) for our predicted outcomes. Output of the above SuperPower Shiny app analyses can be viewed at: https://osf.io/z63am/?view_only=3f085646456f450398249501be24148d

In addition, to account for the differences in the task design between Fawcett et al. (2022), and our current proposal (e.g., different number of trials at test), and to more directly estimate the possible correlations between the dependent measure (d-prime) in different conditions, we conducted additional power analyses based on our previous work (Brown & Roembke, 2024). It should be noted that the semantic manipulation in Fawcett et al. (2022), was between-subjects, whereas our proposed manipulation will be within-subjects. Our previous study used a production effect task and experimental design that is the same as the one we are proposing here, including the same number of trials in different conditions, a 2 × 2 within-subject design, and d-prime scores as the dependent variable. We first ran a power analysis for *N*=75 with the same approach and parameters as described above, with the means and SDs in Fawcett et al. (2022), as well as the dependent measure correlation matrix we observed in our own data: this yielded 92.5% power for detecting an interaction between the semantic manipulation and the production effect. We then ran a power analysis with *N*=75 and the means, SDs, and correlation matrix from our previous data: this yielded 98% power for detecting an interaction between the production effect and the second factor of interest. We additionally ran this power analysis with different correlation sizes (instead of the observed correlation matrix): even with a correlation size of 0.2, N = 75 yielded >80% power. Thus, we take these additional power analyses to suggest that N = 75/experiment should be sufficient to observe a within-subject interaction between the production effect and a second factor of interest using our proposed task parameters (e.g., number of trials). Output from these power analyses can be viewed at Output of the above SuperPower Shiny app analyses can be viewed at: https://osf.io/z63am/?view_only=3f085646456f450398249501be24148d.

Neurotypical adults (aged 18-35, N=75) will be recruited from the student participant pool or the broader RWTH Aachen community. They will be compensated with course credit or 10€/hour. Participants will be German-English bilinguals with an at least medium proficiency in English (see inclusion criteria below). To verify participants' knowledge of German and English, we will administer the LexTALE word identification task (Lemhöfer & Broersma, 2012) in each language. All participants will provide written informed consent prior to participation. The study procedures described here have been approved by the internal ethics committee of the Institute of Psychology at RWTH Aachen University.

The participant inclusion criteria will be as follows:

1) Participants must report on a questionnaire that they are between 18 and 35 years of age.
2) Participants must report on a questionnaire that they are a native German speaker.
3) Participants must perform at or above a cutoff of 75% accuracy on the German LexTALE task.
4) Participants must report on a questionnaire that they speak English.

5) Participants must perform at or above a cutoff of 50% accuracy on the English LexTALE task.
6) Participants must report on a questionnaire that they have normal or corrected-to-normal vision and hearing.
7) Participants must report on a questionnaire that they are free of neurological disorders (specifically, any diagnosed language or learning disability).
8) An additional inclusion criterion will be based on task performance on the learning trials. Participants must perform at least 95% of all learning trials in each learning task correctly: a correctly-performed trial means that the participant correctly spoke the word into their microphone or they silently read the word while the microphone was recording, according to the instructions on a given trial, and the microphone must have successfully recorded the trial. This inclusion criterion will ensure that sufficient trials per participant adhere to the intended production effect manipulation and that the adherence to the manipulation can be verified by the experimenters

Materials

Stimuli consist of 200 high-frequency, concrete German nouns, each paired with a corresponding picture. The words and pictures were selected from the MultiPic word-picture database (Duñabeitia et al., 2018) on several criteria. This freely available picture database was developed to facilitate cross-linguistic work and has been normed for six European languages (including English and German). Because the same German words will be used in both Experiments 1 and 2, we selected stimuli on criteria that are relevant to both experiments (the word-picture pairings in Experiment 1, and the German-English translations in Experiment 2). First, all German words are 4 to 12 letters long. Second, only pictures with a modal name agreement of 75 or higher in

German were selected. This value indexes the percentage of participants who used the most frequent word to name a particular picture. Thus, all pictures used in this study have a 75% or higher level of agreement on the corresponding German word (e.g., WOMAN/FRAU or BRAIN/GEHIRN [English/German]). Third, all German words are non-cognates with English and will be distinct in written form from the English translation based on a normalized Levenshtein distance of below 0.5 (Schepens et al., 2012). Levenshtein distance is a metric to quantify string similarity, indicating the number of additions, subtractions and substitutions that have to be carried out to convert one string into another (Levenshtein, 1966). Levenshtein distance can be normalized with a simple formula that takes the maximum length of the words into account (Schepens et al., 2012), so that orthographic similarity can be more easily compared across different word lengths. Fourth and finally, all English translations of the German words have a minimum Zipf log frequency of 3.5 based on the British SUBTLEX database (see van Heuven et al., 2014). Zipf log frequencies are preferable to frequencies/million, as they allow for a more accurate comparison of frequencies across databases. In general, words with a Zipf log frequency above 4 are considered high-frequency and ones with a frequency below 3 are considered low-frequency (van Heuven et al., 2014). Applying these criteria resulted in approximately 250 possible stimuli in the MultiPic database. The final stimuli were selected from this pool based on qualitative judgments to minimize semantic overlap across pictures by the authors and research assistants that are representative of the participant sample. A full list of stimuli can be found in the appendix (Table A1). In addition to the word-picture stimulus pool used for the main experiment, additional items to be presented during a practice task will consist of the German words for the numbers 1 through 10.

Procedure

The experiment will be conducted online using Gorilla (gorilla.sc; Anwyl-Irvine et al., 2020). Participants will use their own computers to complete the study, and they will be required to use a laptop or a desktop with a microphone. All instructions will be presented in German, except for the instructions for the English LexTALE task (see below). After providing informed consent, participants will then be asked to test their microphone by making a short recording of their voice and playing it back. Participants will be asked to exit the experiment if the recording does not work. All vocal responses during the experiment will be recorded by the participant's microphone.

Participants will first complete two versions of a word-identification task (the LexTALE task): the first version will be in German and the second one will be in English. On each trial participants will be presented with a string of letters, and they will be instructed to decide whether it is an existing word in the respective language by pressing the "s" key if they think it is an existing word (even if they do not know its meaning) and "k" if they think it is not an existing word (or are not sure). Participants will be given 5 seconds to make a response on each trial. Each version of the task will include 60 trials (plus three additional warm-up trials that are discarded), half of which will be existing words and half of which will be non-words, presented in a pseudorandom order.

Participants will then complete a brief practice task to prepare them for the learning tasks of the main experiment. The aim of the practice task is to ensure that participants understand when to speak aloud and when to read silently during the learning tasks. Ten German number words (the German words for the numbers one through ten) will be presented capitalized one at a time in a pseudorandom order in either a blue or white font against a grey background. Participants will be instructed to speak out loud when the words are presented in blue, and to silently read the words presented in white. Each trial will begin with a blank screen shown for 500 milliseconds (ms), followed by the presentation of a word in either blue or white in the center of the screen. When the participant speaks the word aloud into their microphone, they will advance to the next trial as soon as the they indicate that their recording is completed or the maximum recording time of 2.5 seconds is reached. Each word will remain on the screen for the whole recording time.

The main tasks of the experiment consist of a learning task followed by a recognition task, both of which will be completed twice: once with a recognition task where pictures are presented instead of words (semantic recognition), and once with a recognition task where written words will be presented (veridical recognition). Each learning task will present written German words (words in the participants' first language). Thus, each participant will complete the learning and recognition task first with one version of the recognition task, and again with the other version of the recognition task, the order of which will be counter-balanced across participants. The procedure of the learning and recognition tasks will be modeled closely after MacLeod and colleagues (MacLeod et al., 2010) in order to replicate the classic production effect and to facilitate comparison between the present and previous results. For each participant, the 200 word-picture pairs will be randomly divided into two lists of 100 word-picture pairs each. One of the lists of 100 word-picture pairs will be designated for the learning and recognition tasks where pictures will be presented at recognition (the semantic condition), and the other list of 100 word-picture pairs will be designated for the learning and recognition tasks where words will be presented at recognition (the veridical condition). From each list of 100 word-picture pairs, 80 words (words only, without their pictures) will be randomly selected to be presented during the learning task in a given condition, and the remaining 20 word-picture pairs from each list will be used for the foils (either the words or the pictures) for the recognition task in a given condition. In both the semantic and veridical conditions, the 80 words to be presented at learning will be randomly divided into two sets of 40 words, such that 40 words will be presented in blue and 40 words will be presented in white during each learning task, in a random order. For the recognition task in each condition, 20 words will be randomly selected from each set of 40 words. This means that 20 words that were presented in blue, and 20 words that were presented in white, will function as targets in the recognition tasks. In the veridical condition, the 40 targets (words that were presented at learning) will also be presented during the recognition task, in the same written form as they were presented at learning. Likewise the 20 foils will also be presented among the targets, also in a written form. Thus, in the veridical recognition task, 60 written words in total will be presented (40 targets and 20 foils[2]) in a random order, and in a yellow font (as in MacLeod et al., 2010). In the semantic recognition condition, the pictures corresponding to the 40 targets (words that were presented at learning) will be presented at recognition. Likewise, the pictures corresponding to the 20 foils will also be presented among the targets. Thus, in the semantic recognition task, 60 pictures in total will be presented (40 targets and 20 foils) in a random order.

During the learning tasks (in both conditions), each trial will begin with a blank screen presented for 500 ms, followed by the presentation of a blue or white written word against a grey background in the center of the screen (Figure 1A). Participants will have been instructed to speak the word out loud into their microphone when they see a blue word, and to silently read

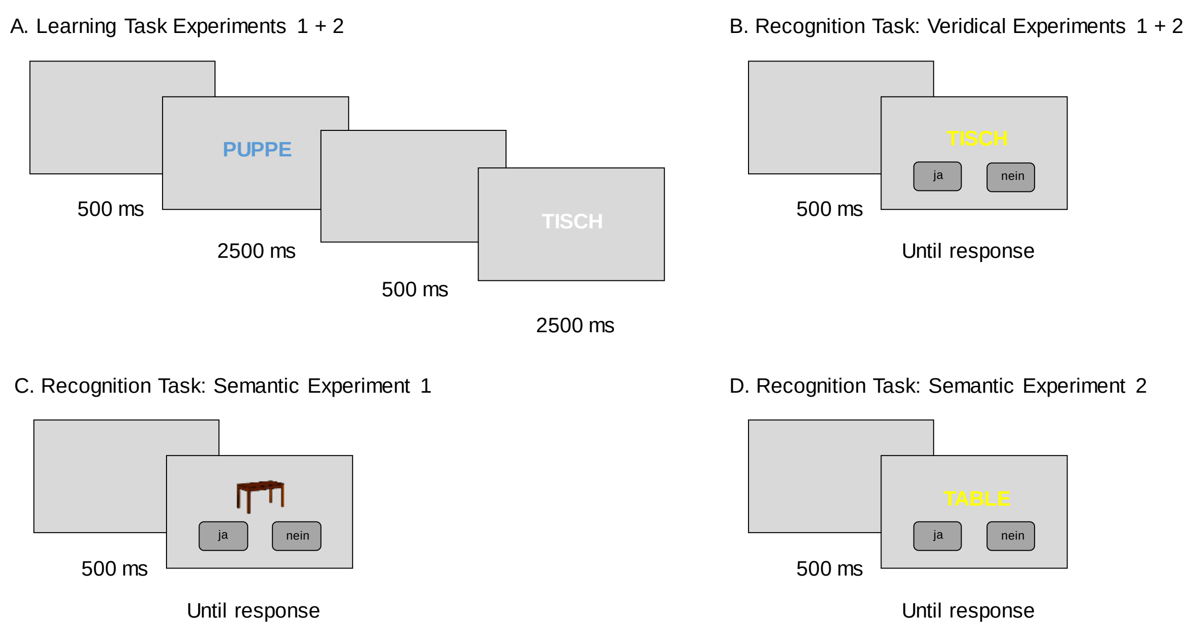
[2] It is a common design choice in production effect experiments that 2/3 of the items at recognition are old and only 1/3 of the items is new. This design was chosen for consistency with previous work (Brown & Roembke, 2024), but it may bias participants to respond “yes” (indicating a picture is old/has been seen before). Previous research suggests that the production effect can be observed independently of the exact make-up of the recognition task (c.f., MacLeod et al., 2022). Nevertheless, to our knowledge, there are currently no studies that directly compare whether proportion of foils impacts the production effect.

the word when they see a white word (they will have been told that when a word is presented in white they should not say anything out loud, not even in a whisper). Participants will also be told that they will be asked to remember the words later. The microphone will record on every trial. When the participant speaks the word aloud into their microphone, they will advance to the next trial as soon as the recording is completed. Each word will remain on the screen for a maximum of 2.5 seconds. Task compliance will be coded offline by research assistants by listening to each trial's recording; trials in which participants followed task instructions (e.g., remained silent on a read silently trial) will be coded as 1 and trials in which participants did not follow task instructions will be coded as 0. Read aloud trials in which the correct word was pronounced but cut-off will also be coded as 1. Each learning task will consist of 80 trials (one word per trial). Each learning task will then be followed immediately by a recognition task (either the veridical or semantic task), after a short set of instructions. During each recognition task, each trial will begin with a blank screen presented for 500 ms, followed by the presentation of either a yellow word (veridical condition; Figure 1B) or a picture (semantic condition; Figure 1C) against a grey background in the upper part of the screen along with two response buttons in the lower part of the screen, one labelled "Ja" (German for "Yes") and one labelled "Nein" (German for "No"). The word or picture will remain on the screen until the participant clicks with their mouse on one of the response buttons, after which the next trial will begin immediately. Each recognition task will consist of 60 trials (one word or picture per trial).

Participants will end the experiment by answering questions related to their language background (modeled on the LEAP-Q questionnaire; Kaushanskaya et al., 2020), demographic

Figure 1

Overview of trial procedures across conditions and experiments. Panel A depicts the learning task that was used in Experiments 1 and 2 (words in blue font should be read aloud; words in white font should be read silently). Panel B depicts the veridical recognition task used in Experiments 1 and 2. Panels C and D depict the semantic recognition tasks used in Experiments 1 and 2, respectively.



information (age, gender, handedness), and they will be asked to verify whether they are neurologically healthy. The entire experiment will last approximately 30 minutes.

Design and Analyses

Experiment 1 will employ a within-subject design with a 2 (Production: Spoken vs. Silent) by 2 (Recognition: Semantic (pictures) vs. Veridical (same words)) factor structure. The dependent variable will be recognition accuracy as indexed by d-prime scores (hit rate minus false alarm rate, both z-normalized), in order to account for possible response bias, similar to previously-reported procedures (Fernandes et al., 2018; Wammes et al., 2019). To account for the possibility of hit rates (the percentage of correct "yes" responses on the recognition test) and false alarm rates (the percentage of incorrect "yes" responses on the recognition test) with extreme values (0 or 1), which would result in infinite d-prime estimates, we will use the so-called log-linear rule (Hautus, 1995; Stanislaw & Todorov, 1999). In this correction, you first add 0.5 to the number of hits and false alarms and add 1 to the number of signal and noise trials, then calculate hit and false alarm rates. This correction has been found to result in less biased d-prime estimates than other possible corrections and is recommended to be used independently of whether extreme values are actually observed (Hautus, 1995; Stanislaw & Todorov, 1999). In addition, hit rates and false alarm rates will also be reported along with their confidence intervals, but they will not be submitted to statistical testing.

The predictions will be tested using frequentist statistics, as follows:

- Predictions 1A (spoken > silent: semantic condition) and 1B (spoken > silent: veridical condition) will be addressed first by looking for a main effect of the factor Production within a 2 (Production: spoken vs. silent) by 2 (recognition: semantic vs. veridical) within-subject ANOVA on d-prime scores, such that across levels of recognition, d-prime scores should be higher in the spoken condition compared to the silent condition (spoken > silent).
- Prediction 1A (spoken > silent: semantic condition): A planned paired-samples t-test will assess whether d-prime scores are higher in the spoken condition compared to the silent condition at a statistically-significant level in the semantic condition.

- Prediction 1B (spoken > silent: veridical condition): A planned paired-samples t-test will assess whether d-prime scores are higher in the spoken condition compared to the silent condition at a statistically-significant level in the veridical condition.
- Prediction 1C (spoken – silent veridical > spoken – silent semantic) will be assessed by looking for an interaction between production and recognition in the same 2 (production: spoken vs. silent) by 2 (recognition: semantic vs. veridical) within-subject ANOVA as above, such that the difference between d-prime scores for spoken and silent words (spoken > silent) should be greater in the veridical recognition condition than in the semantic recognition condition. In addition, planned spoken vs. silent comparisons via paired-samples t-tests on d-prime scores in each recognition condition will assess whether the difference between spoken and silent in the semantic condition is larger than this difference in the veridical condition. Thus, both 1) an interaction in the ANOVA and 2) a numerically larger spoken > silent effect size (as indexed by *Hedge's g*[3] effect size for correlated samples; Lakens, 2013) in the veridical condition are needed to support this prediction. Post-hoc t-tests in each condition separately will additionally assess whether d-prime scores are above-chance, in order to help assess possible floor effects.

Experiment 2

Method

Participants

[3] *Hedge's g* (corrected effect size) is considered to be less biased (especially for small samples) than Cohen's d (uncorrected effect size; Lakens, 2013), but Hedge's g and Cohen's d values are likely to be similar at the sample sizes collected here.

An independent group of participants (*N*=75) will be recruited from the same population as Experiment 1 and will be subject to the same inclusion criteria as stated above.

Materials

Stimuli will include the same 200 German words used in Experiment 1, and will additionally include English translations of each of the German words (200 German-English word pairs). Because words that are cognates in German and English will have already been excluded from Experiment 1, each English word will be different from its German counterpart in terms of sound and written form (normalized Levenshtein distance below 0.5).

Procedure

All aspects of the procedure will be identical to those of Experiment 1, with the following exceptions.

For each participant, the 200 German-English word pairs will be randomly divided into two lists of 100 German-English pairs each. One of the lists of 100 word pairs will be designated for the learning and recognition tasks where the English words will be presented at recognition (the semantic condition), and the other list of 100 word pairs will be designated for the learning and recognition tasks where the same German words will be presented at learning and recognition (the veridical condition). From each list of 100 word pairs, 80 words (only the German words) will be randomly selected to be presented during the learning task, and the remaining 20 word pairs from each list will be used for the foils (either the German words or the English words) for the recognition task. In both the semantic and veridical conditions, the 80 German words to be presented at learning will be randomly divided into two sets of 40 words (40 presented in blue, the other 40 in white, in a random order). From each set of 40 words, 20 will be randomly selected to function as targets in the recognition tasks. In the veridical condition, the

40 target German words will be presented during the recognition task, and 20 German words will be presented as foils. Thus, in the veridical recognition task, 60 German words in total will be presented (40 targets and 20 foils) in a random order. In the semantic recognition condition, the English translations corresponding to the 40 targets (words that were presented at learning) will be presented at recognition (Figure 1D), and the English translations of the 20 foils will also be presented among the targets. Thus, in the semantic recognition task, 60 English words in total will be presented (40 targets and 20 foils) in a random order.

Design and Analyses

Experiment 2 will employ a within-subject design with a 2 (production: spoken vs. silent) by 2 (recognition: semantic (translations) vs. veridical (same words)) factor structure. The dependent variable will be corrected d-prime scores (see Experiment 1). Hit rates and false alarm rates will also be reported along with their confidence intervals, but they will not be submitted to statistical testing.

The predictions will be tested using frequentist statistics, as follows:

- Predictions 2A (spoken > silent: semantic) and 2B (spoken > silent: veridical) will be addressed first by looking for a main effect of production within a 2 (production: spoken vs. silent) by 2 (recognition: semantic vs. veridical) within-subject ANOVA on d-prime scores, such that across levels of recognition, d-prime scores should be higher in the spoken condition compared to the silent condition (spoken > silent).
- Prediction 2A (spoken > silent: semantic condition): A planned paired-samples t-test will assess whether d-prime scores are higher in the spoken condition compared to the silent condition at a statistically-significant level in the semantic condition.

- Prediction 2B (spoken > silent: veridical condition): A planned paired-samples t-test will assess whether d-prime scores are higher in the spoken condition compared to the silent condition at a statistically-significant level in the veridical condition.
- Prediction 2C (spoken – silent veridical > spoken – silent semantic) will be assessed by looking for an interaction between production and recognition in the same 2 (production: spoken vs. silent) by 2 (recognition: semantic vs. veridical) within-subject ANOVA as above, such that the difference between d-prime scores for spoken and silent words (spoken > silent) should be greater in the veridical recognition condition than in the semantic recognition condition. In addition, planned spoken vs. silent comparisons via paired-samples t-tests on d-prime scores in each recognition condition will assess whether the difference between spoken and silent in the semantic condition is larger than this difference in the veridical condition. Thus, both 1) an interaction in the ANOVA and 2) a numerically larger spoken > silent effect size (*Hedge's g*) in the veridical condition than in the semantic condition are needed to support this prediction. Post-hoc t-tests in each condition separately will additionally assess whether d-prime scores are above-chance, in order to help assess possible floor effects.

Declarations

Funding: No funding was received for conducting this study.

Conflicts of interest/competing interests: The authors have no relevant financial or non-financial interests to disclose.

Ethics approval: All study procedures described above will be approved by the internal ethics committee of the Institute of Psychology at RWTH Aachen University, and they will be conducted in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

Consent to participate. All individual participants will give their informed consent prior to their inclusion in this study.

Consent for publication. All individual participants will give their informed consent to the use of their data for publication.

Availability of data and material: The raw, anonymized datasets will be made freely available on OSF.

Code availability: The custom R code used for the current study will be made freely available on OSF.

Authors' contributions: TCR and RMB conceptualized the study and wrote the manuscript.

Open Practices Statement: The experiments described here are described as a registered report. The data and code for all experiments will be made freely available on OSF.

References

Allen, R. J., Hill, L. J. B., Eddy, L. H., & Waterman, A. H. (2020). Exploring the effects of demonstration and enactment in facilitating recall of instructions in working memory. *Memory and Cognition*, *48*(3), 400–410. https://doi.org/10.3758/s13421-019-00978-6

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N. Z., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*, 388–407. https://doi.org/10.3758/s13428-019-01237-x

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 336–345.

Brown, R. M., & Roembke, T. C. (2024). Production benefits on encoding are modulated by language experience: Less experience may help. *Memory & Cognition*, *52*(4), 926–943. https://doi.org/10.3758/s13421-023-01510-7

Cho, K. W., & Feldman, L. B. (2016). When repeating aloud enhances episodic memory for spoken words: interactions between production- and perception-derived variability. *Journal of Cognitive Psychology*, *28*(6), 673–683. https://doi.org/10.1080/20445911.2016.1182173

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428. https://doi.org/10.1037/0033-295X.82.6.407

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(6), 1197–1211. https://doi.org/10.1037/0096-1523.20.6.1197

Conway, M. A., & Gathercole, S. E. (1987). Modality and long-term memory. *Journal of Memory and Language*, *26*(3), 341–361. https://doi.org/10.1016/0749-596X(87)90118-5

Cyr, V., Poirier, M., Yearsley, J. M., Guitard, D., Harrigan, I., & Saint-Aubin, J. (2022). The production effect over the long term: Modeling distinctiveness using serial positions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(12), 1797–1820. https://doi.org/10.1037/xlm0001093

de Groot, A. M. B. (1992). Bilingual lexical representation: A closer look at conceptual representations. *Advances in Psychology*, *94*(C), 389–412. https://doi.org/10.1016/S0166-4115(08)62805-8

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17–22. https://doi.org/10.1037/h0046671

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., & Chang, F. (2014). The p-chain: Relating sentence production and its disorders to comprehension and acquisition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*(1634), 20120394. https://doi.org/10.1098/rstb.2012.0394

Dodson, C. S., & Schacter, D. L. (2001). "If I had said it I would have remembered it": Reducing false memories with a distinctiveness heuristic. *Psychonomic Bulletin and Review*, *8*(1), 155–161. https://doi.org/10.3758/BF03196152

Duñabeitia, J. A., Crepaldi, D., Meyer, A. S., New, B., Pliatsikas, C., Smolka, E., & Brysbaert, M. (2018). MultiPic: A standardized set of 750 drawings with norms for six European languages. *Quarterly Journal of Experimental Psychology*, *71*(4), 808–816. https://doi.org/10.1080/17470218.2017.1310261

Fawcett, J. M. (2013). The production effect benefits performance in between-subject designs: A meta-analysis. *Acta Psychologica*, *142*(1), 1–5. https://doi.org/10.1016/j.actpsy.2012.10.001

Fawcett, J. M., Bodner, G. E., Paulewicz, B., Rose, J., & Wakeham-Lewis, R. (2022). Production can enhance semantic encoding: Evidence from forced-choice recognition with homophone versus synonym lures. *Psychonomic Bulletin and Review*, *29*(6), 2256–2263. https://doi.org/10.3758/s13423-022-02140-x

Fawcett, J. M., & Ozubko, J. D. (2016). *Familiarity, but not recollection, supports the between-subject production effect in recognition memory*. *70*(2), 99–115. https://doi.org/10.1037/cep0000089.supp

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The surprisingly powerful influence of drawing on memory. *Current Directions in Psychological Science*, *27*(5), 302–308. https://doi.org/10.1177/0963721418755385

Forrin, N. D., Jonker, T. R., & MacLeod, C. M. (2014). Production improves memory equivalently following elaborative vs non-elaborative processing. *Memory*, *22*(5), 470–480. https://doi.org/10.1080/09658211.2013.798417

Forrin, N. D., & MacLeod, C. M. (2018). This time it's personal: the memory benefit of hearing oneself. *Memory*, *26*(4), 574–579. https://doi.org/10.1080/09658211.2017.1383434

Forrin, N. D., MacLeod, C. M., & Ozubko, J. D. (2012). Widening the boundaries of the production effect. *Memory and Cognition*, *40*(7), 1046–1055. https://doi.org/10.3758/s13421-012-0210-8

Gathercole, S. E., & Conway, M. A. (1988). Exploring long-term modality effects: Vocalization leads to best retention. *Memory & Cognition*, *16*(2), 110–119. https://doi.org/10.3758/BF03213478

Glanzer, M., & Duarte, A. (1971). Repetition between and within languages in free recall. *Journal of Verbal Learning and Verbal Behavior*, *10*(6), 625–630.

https://doi.org/10.1016/S0022-5371(71)80069-5

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*(3), 491–528. https://doi.org/10.1037/0033-295X.106.3.491

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values ofd′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51. https://doi.org/10.3758/BF03203619

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 534–537. https://doi.org/10.1016/S0022-5371(72)80036-7

Hunt, R. R. (2003). Two contributions of distinctive processing to accurate memory. *Journal of Memory and Language*, *48*(4), 811–825. https://doi.org/10.1016/S0749-596X(03)00018-4

Icht, M., Bergerzon-Biton, O., & Mama, Y. (2019). The production effect in adults with dysarthria: improving long-term verbal memory by vocal production. *Neuropsychological Rehabilitation*, *29*(1), 131–143. https://doi.org/10.1080/09602011.2016.1272466

Jamieson, R. K., Mewhort, D. J. K., & Hockley, W. E. (2016). A computational account of the production effect: Still playing twenty questions with nature. *Canadian Journal of Experimental Psychology / Revue Canadienne de Psychologie Expérimentale*, *70*(2), 154–164. https://doi.org/10.1037/cep0000081

Johnson, M. K., Hashtroudi, S., & Lindsay, D. S. (1993). Source monitoring. *Psychological Bulletin*, *114*(1), 3–28. https://doi.org/10.1037/0033-2909.114.1.3

Kaushanskaya, M., Blumenfeld, H. K., & Marian, V. (2020). The Language Experience and Proficiency Questionnaire (LEAP-Q): Ten years later. *Bilingualism: Language and*

Cognition, *23*(5), 945–950. https://doi.org/10.1017/S1366728919000038

Kaushanskaya, M., & Yoo, J. (2011). Rehearsal effects in adult word learning. *Language and Cognitive Processes*, *26*(1), 121–148. https://doi.org/10.1080/01690965.2010.486579

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470–1489. https://doi.org/10.1037/xlm0000391

Kroll, J. F., & Stewart, E. (1994). Category inference in translation and picture naming: Evidence for asymmetric connections between bilingual memory representations. *Journal of Memory and Language*, *33*(2), 149–174. https://doi.org/10.1006/jmla.1994.1008

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*(NOV), 863. https://doi.org/10.3389/fpsyg.2013.00863

Lakens, D., & Caldwell, A. R. (2021). Simulation-based power analysis for factorial analysis of variance designs. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920951503

Lemhöfer, K., & Broersma, M. (2012). Introducing LexTALE: A quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods*, *44,* 325–343. https://doi.org/10.3758/s13428-011-0146-0

Levenshtein, V. I. (1966). Binary codes capable of correcting dele- tions, insertions and reversals. *Soviet Physics Doklady*, *10*(8), 707–710.

Lorch, R. F. (1982). Priming and search processes in semantic memory: a test of three models of spreading activation. *Journal of Verbal Learning and Verbal Behavior*, *21*(4), 468–492. https://doi.org/10.1016/S0022-5371(82)90736-8

MacLeod, C. M. (2011). I said, you said: The production effect gets personal. *Psychonomic Bulletin and Review*, *18*(6), 1197–1202. https://doi.org/10.3758/s13423-011-0168-8

MacLeod, C. M., Gopie, N., Hourihan, K. L., Neary, K. R., & Ozubko, J. D. (2010). The production effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *36*(3), 671–685. https://doi.org/10.1037/a0018785

MacLeod, C. M., Ozubko, J. D., Hourihan, K. L., & Major, J. C. (2022). The production effect is consistent over material variations: support for the distinctiveness account. *Memory*, *30*(8), 1000–1007. https://doi.org/10.1080/09658211.2022.2069270

Mama, Y., & Icht, M. (2016). Auditioning the distinctiveness account: Expanding the production effect to the auditory modality reveals the superiority of writing over vocalising. *Memory*, *24*(1), 98–113. https://doi.org/10.1080/09658211.2014.986135

Marmolejo, G., Diliberto-Macaluso, K. A., & Altarriba, J. (2009). False memory in bilinguals: does switching languages increase false memories? *The American Journal of Psychology*, *122*(1), 1–16.

McCallum, R. S., Sharp, S., Bell, S. M., & George, T. (2004). Silent versus oral reading comprehension and efficiency. *Psychology in the Schools*, *41*(2), 241–246. https://doi.org/10.1002/pits.10152

McDermott, K. B. (1996). The persistence of false memories in list recall. *Journal of Memory and Language*, *35*(2), 212–230. https://doi.org/10.1006/jmla.1996.0012

McNamara, T. P. (1992). Priming and constraints it places on theories of memory and retrieval. *Psychological Review*, *99*(4), 650–662. https://doi.org/10.1037/0033-295X.99.4.650

Meade, M. L., Watson, J. M., Balota, D. A., & Roediger III, H. L. (2007). The roles of spreading activation and retrieval mode in producing false recognition in the DRM paradigm. *Journal* *of Memory and Language*, *56*(3), 305–320. https://doi.org/10.1016/j.jml.2006.07.007

Miyaji-Kawasaki, Y., Inoue, T., & Yama, H. (2004). Cross-linguistic false recognition: How do japanese-dominant bilinguals process two languages: Japanese and english? *Psychologia*, *46*(4), 255–267. https://doi.org/10.2117/psysoc.2003.255

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519–533. https://doi.org/10.1016/S0022-5371(77)80016-9

Murray, D. J. (1965). Vocalization-at-presentation, auditory presentation and immediate recall. *Nature*, *207*(5000), 1011–1012. https://doi.org/10.1038/2071011a0

Ozubko, J. D., Hourihan, K. L., & MacLeod, C. M. (2012). Production benefits learning: The production effect endures and improves memory for text. *Memory*, *20*(7), 717–727. https://doi.org/10.1080/09658211.2012.699070

Ozubko, J. D., & MacLeod, C. M. (2010). The production effect in memory: Evidence that distinctiveness underlies the benefit. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1543–1547. https://doi.org/10.1037/a0020604

Pardilla-Delgado, E., & Payne, J. D. (2017). The Deese-Roediger-McDermott (DRM) task: A simple cognitive paradigm to investigate false memories in the laboratory. *Journal of Visualized Experiments*, *119*, e54793. https://doi.org/10.3791/54793

Quinlan, C. K., & Taylor, T. L. (2013). Enhancing the production effect in memory. *Memory*, *21*(8), 904–915. https://doi.org/10.1080/09658211.2013.766754

Quinlan, C. K., & Taylor, T. L. (2019). Mechanisms underlying the production effect for singing. *Canadian Journal of Experimental Psychology*, *73*(4), 254–264. https://doi.org/10.1037/cep0000179

Robinson, M. F., Meisinger, E. B., & Joyner, R. E. (2019). The influence of oral versus silent reading on reading comprehension in students with reading disabilities. *Learning Disability Quarterly*, *42*(2), 105–116. https://doi.org/10.1177/0731948718806665

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803–814. https://doi.org/10.1037/0278-7393.21.4.803

Sahlin, B. H., Harding, M. G., & Seamon, J. G. (2005). When do false memories cross language boundaries in English-Spanish bilinguals? *Memory and Cognition*, *33*(8), 1414–1421. https://doi.org/10.3758/BF03193374

Saint-Aubin, J., Yearsley, J. M., Poirier, M., Cyr, V., & Guitard, D. (2021). A model of the production effect over the short-term: The cost of relative distinctiveness. *Journal of Memory and Language*, *118*, 104219. https://doi.org/10.1016/j.jml.2021.104219

Schepens, J., Dijkstra, T., & Grootjen, F. (2012). Distributions of cognates in Europe as based on Levenshtein distance. *Bilingualism: Language and Cognition*, *15*(1), 157–166. https://doi.org/10.1017/S1366728910000623

Schimmel, N., & Ness, M. (2017). The effects of oral and silent reading on reading comprehension. *Reading Psychology*, *38*(4), 390–416. https://doi.org/10.1080/02702711.2016.1278416

Seidenberg, M. S. (2005). Connectionist models of word reading. *Current Directions in Psychological Science*, *14*(5), 238–242. https://doi.org/10.1111/j.0963-7214.2005.00372.x

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/BF03209391

Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*(1), 137–149. https://doi.org/10.3758/BF03207704

Suarez, M., & Beato, M. S. (2021). The role of language proficiency in false memory: A mini review. *Frontiers in Psychology*, *12*(April). https://doi.org/10.3389/fpsyg.2021.659434

Tsuboi, N., Francis, W. S., & Jameson, J. T. (2021). How word comprehension exposures facilitate later spoken production: implications for lexical processing and repetition priming. *Memory*, *29*(1), 39–58. https://doi.org/10.1080/09658211.2020.1845740

van den Boer, M., van Bergen, E., & de Jong, P. F. (2014). Underlying skills of oral and silent reading. *Journal of Experimental Child Psychology*, *128*, 138–151. https://doi.org/10.1016/j.jecp.2014.07.008

van Hell, J. G., & de Groot, A. M. B. (1998). Conceptual representation in bilingual memory: Effects of concreteness and cognate status in word association. *Bilingualism: Language and Cognition*, *1*(3), 193–211. https://doi.org/10.1017/s1366728998000352

van Heuven, W. J. B., Mandera, P., Keuleers, E., & Brysbaert, M. (2014). SUBTLEX-UK: A new and improved word frequency database for British English. *Quarterly Journal of Experimental Psychology*, *67*(6), 1176–1190. https://doi.org/10.1080/17470218.2013.850521

Wakeham-Lewis, R. M., Ozubko, J., & Fawcett, J. M. (2022). Characterizing production: the production effect is eliminated for unusual voices unless they are frequent at study. *Memory*, *30*(10), 1319–1333. https://doi.org/10.1080/09658211.2022.2115075

Wammes, J. D., Jonker, T. R., & Fernandes, M. A. (2019). Drawing improves memory: The importance of multimodal encoding context. *Cognition*, *191*, 103955.

https://doi.org/10.1016/j.cognition.2019.04.024

Ziegler, J. C., Castel, C., Pech-Georgel, C., George, F., Alario, F. X., & Perry, C. (2008). Developmental dyslexia and the dual route model of reading: simulating individual differences and subtypes. *Cognition*, *107*(1), 151–178. https://doi.org/10.1016/j.cognition.2007.09.004

Appendix

Table A1*:*

Overview of stimuli. Picture numbers indicate the corresponding picture from the MultiPic word-picture database (Duñabeitia et al., 2018) that was used for each word pair.

English (L2)	German (L1)	Picture No.
CHIN	KINN	6
LEG	BEIN	8
SHOWER	DUSCHE	9
LION	LÖWE	15
HUNTER	JÄGER	16
DEVIL	TEUFEL	17
TURKEY	TRUTHAHN	19
BIKE	FAHRRAD	23
BONE	KNOCHEN	24
ONION	ZWIEBEL	26
CROWN	KRONE	32
POCKET	HOSENTASCHE	39
PEAR	BIRNE	42
CLAW	KRALLE	44
DESERT	WÜSTE	47
SAW	SÄGE	52
SHADOW	SCHATTEN	55
ELBOW	ELLENBOGEN	56
STAMP	BRIEFMARKE	57
NEWSPAPER	ZEITUNG	61
BUTTON	KNOPF	64
ROOF	DACH	66
CANDLE	KERZE	68
TROPHY	POKAL	71
CHEESE	KÄSE	72
WAVE	WELLE	73
KEY	SCHLÜSSEL	78
KEYBOARD	TASTATUR	79
MEAT	FLEISCH	81
RIVER	FLUSS	86
SUITCASE	KOFFER	95
GREENHOUSE	GEWÄCHSHAUS	106
TABLE	TISCH	110
FRUIT	OBST	114
HIPPO	NILPFERD	115
FIRE	FEUER	116
HORSE	PFERD	117
CUT	WUNDE	118

MAZE	LABYRINTH	121
CHAIR	STUHL	122
CORNER	ECKE	124
PARACHUTE	FALLSCHIRM	126
DOLL	PUPPE	128
LEMON	ZITRONE	129
FRIDGE	KÜHLSCHRANK	138
TAP	WASSERHAHN	140
SUNFLOWER	SONNENBLUME	141
BIN	MÜLLEIMER	145
TRAY	TABLETT	151
SNAIL	SCHNECKE	152
TIE	KRAWATTE	161
COCONUT	KOKOSNUSS	163
ROAD	STRAßE	164
BELT	GÜRTEL	165
BACK	RÜCKEN	169
GOALKEEPER	TORWART	171
PARROT	PAPAGEI	172
TRIANGLE	DREIECK	181
WITCH	HEXE	184
ENGINE	MOTOR	186
JELLYFISH	QUALLE	193
SCAR	NARBE	207
SCALES	WAAGE	212
POLICEMAN	POLIZIST	213
BEDROOM	SCHLAFZIMMER	214
WALL	MAUER	227
CHEST	BRUST	231
PLATE	TELLER	234
NEEDLE	NADEL	235
EYE	AUGE	241
ROOTS	WURZELN	245
DENTIST	ZAHNARZT	246
BRAIN	GEHIRN	247
POTATO	KARTOFFEL	249
BUTCHER	METZGER	254
ISLAND	INSEL	260
RABBIT	HASE	263
HEEL	ABSATZ	266
ARROW	PFEIL	270
WIG	PERÜCKE	277
DRUM	TROMMEL	278
NECK	HALS	280

RHINO	NASHORN	282
FOREST	WALD	288
OSTRICH	STRAUß	299
SWORD	SCHWERT	301
WING	FLÜGEL	307
DICE	WÜRFEL	308
RAZOR	RASIERER	311
BARBER	FRISEUR	313
SHELL	MUSCHEL	316
OWL	EULE	323
TELEVISION	FERNSEHER	325
BRA	BÜSTENHALTER	331
RULER	LINEAL	336
PRESENT	GESCHENK	338
MIRROR	SPIEGEL	340
BOTTLE	FLASCHE	343
FOUNTAIN	BRUNNEN	344
PEPPER	PAPRIKA	348
LOCK	SCHLOSS	349
GOAT	ZIEGE	354
CAR	AUTO	358
KNIFE	MESSER	359
FAN	VENTILATOR	363
HEDGEHOG	IGEL	365
POT	TOPF	369
CAULIFLOWER	BLUMENKOHL	373
LOBSTER	HUMMER	378
STRAWBERRY	ERDBEERE	381
COIN	MÜNZE	383
CHAIN	KETTE	391
DUCK	ENTE	400
MOUNTAIN	BERG	407
FACTORY	FABRIK	412
CURTAIN	VORHANG	418
COMB	KAMM	426
GLOVE	HANDSCHUH	431
TEACHER	LEHRER	434
GIRL	MÄDCHEN	436
THUMB	DAUMEN	438
SAUSAGE	WURST	441
PINEAPPLE	ANANAS	442
PIG	SCHWEIN	446
BENCH	BANK	449
SCISSORS	SCHERE	453

WARDROBE	SCHRANK	458
GUN	PISTOLE	461
FLY	FLIEGE	462
RUBBER	RADIERGUMMI	464
JUDGE	RICHTER	466
SINK	WASCHBECKEN	478
SQUIRREL	EICHHÖRNCHEN	484
BASKET	KORB	487
SUN	SONNE	488
RUG	TEPPICH	491
HARBOUR	HAFEN	495
CIRCLE	KREIS	496
COFFIN	SARG	497
MUG	TASSE	498
STAIRS	TREPPE	500
BOOK	BUCH	505
FACE	GESICHT	510
SUBMARINE	U-BOOT	514
ROPE	SEIL	517
FOX	FUCHS	521
FENCE	ZAUN	530
TURTLE	SCHILDKRÖTE	531
LIGHTHOUSE	LEUCHTTURM	540
SHOE	SCHUH	541
BAT	FLEDERMAUS	547
SNAKE	SCHLANGE	549
PUMPKIN	KÜRBIS	551
APPLE	APFEL	552
LIGHTER	FEUERZEUG	558
CUCUMBER	GURKE	559
CATERPILLAR	RAUPE	561
SPOON	LÖFFEL	564
BELL	GLOCKE	565
VIOLIN	GEIGE	566
CASTLE	BURG	568
TOOTH	ZAHN	569
BEAK	SCHNABEL	575
TOAD	KRÖTE	584
ANT	AMEISE	585
IRON	BÜGELEISEN	586
BROOM	BESEN	590
TOWEL	HANDTUCH	591
LETTUCE	SALAT	592
WINDOW	FENSTER	597

CLOUD	WOLKE	599
CAT	KATZE	606
QUEEN	KÖNIGIN	614
AIRPORT	FLUGHAFEN	617
SOAP	SEIFE	619
TANK	PANZER	621
CHALK	KREIDE	622
DONKEY	ESEL	624
SCARF	SCHAL	628
SHEEP	SCHAF	636
COAT	MANTEL	644
AMBULANCE	KRANKENWAGEN	649
SPONGE	SCHWAMM	650
SOUP	SUPPE	651
PENCIL	BLEISTIFT	654
DRAWER	SCHUBLADE	661
DRESS	KLEID	664
FORK	GABEL	673
BOW	BOGEN	680
PEANUT	ERDNUSS	684
GOOSE	GANS	685
CHERRY	KIRSCHE	692
TREE	BAUM	693
HUG	UMARMUNG	694
EAGLE	ADLER	703
DOG	HUND	707
TROUSERS	HOSE	718
WAITER	KELLNER	726
GLASSES	BRILLE	733
FISHERMAN	ANGLER	740

1

Study Design Overview

Question	Hypothesis	Sampling plan	Analysis Plan	Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis	Interpretation given different outcomes	Theory that could be shown wrong by the outcomes
Does the production effect (higher recognition accuracy for words that were previously spoken out loud compared to words that were silently read) persist when items at learning and recognition match on semantic but not other features?	Experiment 1: Prediction 1A: Recognition accuracy will be greater for words that were spoken compared to those that were silently read at learning (a production effect) when participants are asked to recognize pictures corresponding to the words they had studied at learning. Prediction 1B: Recognition accuracy will be greater for words that were spoken	In each of two experiments, seventy-five German-English bilingual adults (18-35 years of age) with German as a first language and English as a second language will be recruited from the RWTH Aachen University community. See Participants section for	Experiment 1: 2 (Learning condition: Spoken vs. Silent) by 2 (Recognition condition: Pictures vs. Written words) within-subject ANOVA with d-prime scores as the dependent variable. Prediction 1A and 1B: A main effect of Learning condition (Spoken > Silent). Prediction 1A: A planned pairwise Spoken > Silent	A series of simulation-based power analyses using the SuperPower Shiny app (Lakens & Caldwell, 2021) with 10000 simulations and alpha = 0.05 were performed, based on a previously-reported 2 × 2 interaction in a study addressing a theoretically-similar research question (Fawcett et al., 2022). An initial power analysis using the previously-reported means and SDs in each cell of a 2	We could observe that the production effect is present (greater recognition accuracy for spoken words compared to silently-read words), both when participants are asked to recognize pictures corresponding to the words they had studied, and when they are asked to recognize the same written words they had studied, as we predict. This would suggest that the production effect persists when words that were studied can be recognized on their semantic features,	The notion that the production effect could be influenced by semantic encoding (via the influence of articulation on spreading activation) could fail to be supported by the findings, and could thus be called into question. Our results also have implications for whether modality-

	compared to those that were silently read at learning (a production effect) when participants are asked to recognize words presented in the same written form as they were at learning.	more details.	comparison when participants recognize pictures (in the Pictures condition). Prediction 1B: A planned pairwise Spoken > Silent comparison when participants recognize words presented in the same written form as they were at learning (in the veridical condition).	(production: aloud/silent) by 2 (recognition: semantic interference / control) suggested that 75 participants are needed to achieve 100% power for a production main effect, at least 97% power for each planned comparison, and 82% power for the interaction. Additional power analyses were preformed using the values at the upper 95% CI boundaries of the “silent” condition cell means and the values at the lower 95% CI boundaries of the “spoken” condition cell means (thus using the most	and that production may influence semantic encoding. This outcome would be consistent with the idea of spreading activation: speaking (e.g., articulatory features) could engage modality-independent associations with semantic features, even if those associations are indirect (i.e. mediated by other, stronger associations). If we do not detect a production effect (contrary to our prediction) when participants are asked to recognize pictures, this would raise the possibility that production may have little or no influence on semantic encoding, but this	independent and modality-dependent features are engaged during production and the role of transfer-appropriate processing in the production effect.

				conservative estimate of the production effect in each condition). These analyses suggested that 75 participants are sufficient to achieve 92% power for the planned comparison in the control condition, and 87% power for the interaction. Similar levels of power were shown for samples sizes of *N*=80, *N*=90, and *N*=100, thus we deemed *N*=75 to be sufficient (for more details on the sample size justification see the Participants section). Another series of power analyses were run using data from a previous study with a similar task design (Brown & Roembke, 2024).	interpretation would need to be more directly tested with further analyses. The absence of a production effect would be in line with transfer-appropriate processing enhancing memory performance when conditions match at encoding and retrieval (which is more the case in the veridical than the semantic conditions). This outcome would also strongly align with the assumption of some memory models that speaking adds only modality-dependent features to memory traces, and not modality-independent features (such as semantic features).	

				A power analysis with N=75 using the means and SDs in Fawcett et al., as well as the dependent measure correlation matrix observed in Brown & Roembke yielded 92.5 percent power for detecting an interaction between the production effect and the semantic manipulation. A power analysis with N=75 and the d-prime means, SDs, and correlation matrix from Brown & Roembke yielded 98% power for detecting an interaction between the production effect and the second factor of interest.		

				Note: CI=confidence intervals.		
As above	Experiment 2: Prediction 2A: Recognition accuracy will be greater for words that were spoken compared to those that were silently read at learning (a production effect) when participants are asked to recognize translations of the studied words into another language. Prediction 2B: Recognition accuracy will be greater for words that were spoken compared to those that were silently read at learning (a production effect)	Experiment 2: 2 (Learning condition: Spoken vs. Silent) by 2 (Recognition condition: Different Language vs. Same Language) within-subject ANOVA with d-prime scores as the dependent variable. Prediction 2A and 2B: A main effect of Learning condition (Spoken > Silent). Prediction 2A: A planned pairwise Spoken > Silent comparison when participants recognize translations of studied words (in the Different Language condition). Prediction 2B: A planned pairwise Spoken > Silent comparison when participants recognize words presented in the same language as they were at			We could observe that the production effect is present (greater recognition accuracy for spoken words compared to silently-read words) when participants recognize translations of studied words, and when they recognize the same words they had studied (in the same language), as we predict. This would suggest that the production effect persists when words that were studied can only be recognized on semantic features, and that production may influence semantic encoding. This outcome would be consistent with the idea of	As above.

	when participants are asked to recognize words presented in the same language as they were at learning.	learning (in the Same Language condition).		spreading activation: speaking (e.g., articulatory features) could engage modality-independent associations with semantic features, even if those associations are indirect (i.e. mediated by other, stronger associations). Contrary to our prediction, if we do not detect a production effect when participants recognize translations of the studied words, this would raise the possibility that production may have little or no influence on semantic encoding, but this interpretation would need to be	

				more directly tested with further analyses. This outcome would also strongly align with the assumption of some memory models that speaking adds only modality-dependent features to memory traces, and not modality-independent features (such as semantic features).	
Is the production effect greater when items at learning and recognition match on multiple linguistic features compared to only semantic features?	Experiment1: Prediction 1C: The increase in recognition accuracy from having spoken compared to having silently read words (production effect: spoken > silent)	Experiments 1 and 2: Predictions 1C and 2C: We will look for an interaction in the ANOVA for each experiment (above), and the same planned comparisons will assess the effect size of the difference between spoken and silent words in each recognition condition.		We could observe, as we predict, that the production effect decreases when participants recognize pictures (Exp. 1) or translations (Exp. 2) compared to when they recognize items that match those presented at learning. This	As above.

	will be larger when words are presented in the same written form at recognition compared to when recognition items are presented as pictures. Experiment 2: Prediction 2C: The increase in recognition accuracy from having spoken compared to having silently read words (production effect: spoken > silent) will be larger when words are presented in the same language at learning and recognition compared to when words are presented in a			would suggest that production may influence not only semantic encoding but other linguistic features as well. This outcome would align with spreading activation, and with memory models that assume that speaking can engage modality-independent features and could also be fit to attenuated or modified versions of alternative accounts. In addition, transfer-appropriate processing may modify retrieval success, such that memory can improve when there is some degree of similarity between processing at encoding and retrieval.	

	different language at recognition.			If, contrary to our prediction, we do not detect a difference in the production effect as a function of how items are presented at recognition, it raises the possibility that semantic encoding may be sufficient for the production effect, but this would have to be examined with further analyses. This outcome would also call into question the assumption that speaking only engages modality-dependent features, and the idea of transfer-appropriate processing. If, contrary to our prediction, we observe a larger	

				production effect when recognition items are presented as pictures or translations, this would suggest that production could enhance the encoding of semantic features relative to other linguistic features. This outcome would strongly contradict the assumption that speaking only engages modality-dependent features, as well as transfer-appropriate processing.	

Guidance Notes

- **Question**: articulate each research question being addressed in one sentence.
- **Hypothesis**: where applicable, a prediction arising from the research question, stated in terms of specific variables rather than concepts. Where the testability of one or more hypotheses depends on the verification of auxiliary assumptions (such as positive controls, tests of intervention fidelity, manipulation checks, or any other quality checks), any tests of such assumptions should be listed as hypotheses. Stage 1 proposals that do not seek to test hypotheses can ignore or delete this column.

- **Sampling plan**: For proposals using inferential statistics, the details of the statistical sampling plan for the specific hypothesis (e.g power analysis, Bayes Factor Design Analysis, ROPE etc). For proposals that do not use inferential statistics, include a description and justification of the sample size.
- **Analysis plan**: For hypothesis-driven studies, the specific test(s) that will confirm or disconfirm the hypothesis. For non-hypothesis-driven studies, the test(s) that will answer the research question.
- **Rationale for deciding the sensitivity of the test for confirming or disconfirming the hypothesis**: For hypothesis-driven studies that employ inferential statistics, an explanation of how the authors determined a relevant effect size for statistical power analysis, equivalence testing, Bayes factors, or other approach.
- **Interpretation given different outcomes**: A prospective interpretation of different potential outcomes, making clear which outcomes would confirm or disconfirm the hypothesis.
- **Theory that could be shown wrong by the outcomes**: Where the proposal is testing a theory, make clear what theory could be shown to be wrong, incomplete, or otherwise inadequate by the outcomes of the research.