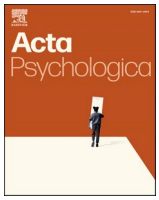
Acta Psychologica 241 (2023) 104073

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/locate/actpsy



Word frequency has similar effects in picture naming and gender decision: A failure to replicate Jescheniak and Levelt (1994)

Ruth E. Corps [a,*], Antje S. Meyer [a,b]

[a] Psychology of Language Department, Max Planck Institute for Psycholinguistics, The Netherlands
[b] Donders Centre of Cognition and Cognitive Neuroscience, Radboud University, The Netherlands

ARTICLE INFO

Keywords:
Word production
Word frequency
Repetition priming
Lexical access

ABSTRACT

Word frequency plays a key role in theories of lexical access, which assume that the word frequency effect (WFE, faster access to high-frequency than low-frequency words) occurs as a result of differences in the representation and processing of the words. In a seminal paper, Jescheniak and Levelt (1994) proposed that the WFE arises during the retrieval of word forms, rather than the retrieval of their syntactic representations (their lemmas) or articulatory commands. An important part of Jescheniak and Levelt's argument was that they found a stable WFE in a picture naming task, which requires complete lexical access, but not in a gender decision task, which only requires access to the words' lemmas and not their word forms. We report two attempts to replicate this pattern, one with new materials, and one with Jescheniak and Levelt's orginal pictures. In both studies we found a strong WFE when the pictures were shown for the first time, but much weaker effects on their second and third presentation. Importantly these patterns were seen in both the picture naming and the gender decision tasks, suggesting that either word frequency does not exclusively affect word form retrieval, or that the gender decision task does not exclusively tap lemma access.

1. Introduction

Speakers are faster to produce more frequent words (e.g., *dog*) than less frequent words (e.g., *stag*; e.g., Oldfield & Wingfield, 1965). Word frequency plays a key organising role in theories of lexical access in word production (Caramazza, 1997; Dell, 1986; Levelt, Roelofs, & Meyer, 1999; Roelofs, 1997), which assume that frequency effects occur as a result of exposure-related differences in the representation and processing of high-frequency (HF) and low-frequency (LF) words. But which components of word production are affected by word frequency? Although word frequency effects are well attested in the literature and are generally ascribed to properties of lexical representations, it is less clear which of these representations are frequency sensitive.

Lexical access is typically considered to be a staged process, involving the broad steps of identifying the concept to be expressed, the selection of the grammatical representation of the word (its lemma), and the retrieval of the corresponding word form (e.g., Indefrey, 2011; Levelt, 1989; Levelt et al., 1999). In some theories, processing is largely feedforward and so properties of words, such as their frequency, concreteness, or age of acquisition, can be ascribed to individual processing levels. In these theories, word frequency can be located at a single level, such as at the word form level (Jescheniak & Levelt, 1994). As a result, the ease of producing a word is affected by the ease of accessing its word form representation. Other theories, however, view lexical access as an interactive process, with bidirectional information flow between levels (e.g., Caramazza, 1997; Dell, 1986; Rapp & Goldrick, 2000) and (near) simultaneous access of different types of lexical representations (e.g., Strijkers & Costa, 2011). Under these theories, there is feedback between processing levels and so word frequency can affect multiple levels of processing (e.g., Kittredge, Dell, Verkuilen, & Schwartz, 2008).

In a seminal paper, Jescheniak and Levelt (1994) found three key results that located the word frequency effect exclusively at the word form level. First, frequency affected picture naming latencies, and this frequency effect was stable across three presentations of the same picture. Second, there was no robust evidence that frequency affected latencies in a gender decision task, which required retrieval of the picture's grammatical representation (its lemma, see below for details). In particular, frequency affected gender decision latencies on the picture's first presentation, but not on its second or third presentation. Third, there was robust evidence that frequency affected ease of accessing word forms during a homophone translation task, and this

* Corresponding author at: Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands.
E-mail address: Ruth.Corps@mpi.nl (R.E. Corps).

https://doi.org/10.1016/j.actpsy.2023.104073
Received 12 September 2022; Received in revised form 28 September 2023; Accepted 26 October 2023
Available online 8 November 2023

frequency effect was stable across three presentations. Thus, Jescheniak and Levelt concluded that frequency affected word form access only. These findings have been highly influential in the word production literature, with almost 700 citations in WebofScience and over 1300 citations in Google Scholar at the time of writing (20 September 2023). They have often been seen as support for a modular architecture of the production system, with serial stages of processing and little feedback between these stages.

Our paper concerns Jescheniak and Levelt's first two findings – in particular, the stability of the word frequency effect in picture naming and gender decision. The claim that frequency does not affect lemma access rests on finding a stable frequency effect during picture naming, but not during gender decision. But, as we discuss below, there is evidence that word frequency effects are unstable during picture naming, and often become smaller when pictures are presented multiple times. Additionally, there is experimental evidence suggesting that frequency does affect lemma access. As a result, we ran two experiments (each with three different tasks) designed to replicate Jescheniak and Levelt's pattern of results. To anticipate our findings, we found that word frequency affected response times only on the picture's first presentation in both picture naming and gender decision. These findings suggest that either lemma representations, implicated in both tasks, are frequency sensitive, or that participants activated word form representations in both gender decision and picture naming, which means that the two levels of representation cannot be separated with this combination of tasks. Either way, our results do not provide support for Jescheniak and Levelt's claim that word frequency does not affect lemma selection.

Before turning to our experiments, we describe Jescheniak and Levelt's study in more detail and discuss other studies that have investigated word frequency effects during lemma access and have tested the stability of these effects during picture naming.

1.1. Jescheniak and Levelt's study of word frequency effects

Jescheniak and Levelt (1994) conducted seven experiments designed to locate word frequency effects in a serial model of lexical access. In Experiment 1, participants studied a booklet to familiarise themselves with pictures of objects with HF and LF names and then named each picture three times as quickly and accurately as possible. Participants were faster to name HF (mean (M) = 649 ms) than LF (M = 711 ms) pictures, and this frequency effect was stable across three presentations. In Experiments 2 and 3, participants were not familiarised with the materials. In Experiment 2, they performed an object recognition task where they indicated whether a picture of an object matched a previously presented word. In Experiment 3, they produced a picture name after a delay. No word frequency effect was obtained in either experiment, suggesting frequency did not affect conceptual selection (e.g., Morrison, Ellis, & Quinlan, 1992) or speech motor planning (e.g., Balota & Chumbley, 1985). Thus, the frequency effect in the picture naming experiment arose during lexical access.

In subsequent experiments, Jescheniak and Levelt (1994) tested which levels of lexical access are affected by word frequency. Experiments 4 and 5 targeted the lemma level. In Experiment 4, participants were familiarised with the pictures (but not their names) and then performed a button-press task to indicate the grammatical gender of the picture names (either gendered [using the determiner *de*] or gender neutral [using the determiner *het*]). Participants were faster to determine the gender of pictures with HF names (M = 733 ms) than LF pictures (M = 769 ms). However, this frequency effect was present only on the picture's first presentation and disappeared on its second and third presentations. Jescheniak and Levelt suggested that this interaction reflected participants' accommodation to the task, rather than a robust frequency effect. In particular, they proposed that participants silently generated full noun phrases (i.e., determiner plus noun, as in *de hond, [the dog]*) on the picture's first presentation. Participants then determined the picture's gender by monitoring for the determiner in their inner speech, thus accessing the determiner and noun's word form. The frequency effect disappeared on subsequent presentations because participants had recently determined the noun's gender, and so they were more efficient and determined it without accessing the word form.

To test this proposal, participants in Experiment 5 named each picture twice before making gender decisions on the same pictures, again twice. Participants were not familiarised with the pictures before naming. In the naming phase, they either named the pictures using a bare noun (Experiment 5a) or a full noun phrase (Experiment 5b). This experiment tested whether the frequency effect in the gender decision task was eliminated after recent lemma access (Experiments 5a and 5b) or whether it was eliminated only after retrieval of the determiner and the word form (Experiment 5b only). In the naming phases of both experiments, participants were faster to name HF than LF pictures and this frequency effect was seen on both presentations of the materials (replicating Experiment 1). In the gender decision phase of Experiment 5a (after bare noun naming), participants were faster to determine the gender of HF than LF pictures on their first presentation, but crucially not on their second presentation. In contrast, there was no frequency effects during gender decision in Experiment 5b (after retrieving the determiner and the noun). Based on these results, Jescheniak and Levelt concluded that prior lemma access eliminated the word frequency effect during gender decision, and proposed that this effect was actually a recency effect.

There was evidence, however, that frequency affected word form selection. In Experiment 6, participants completed an oral English-Dutch translation task, which involved accessing the Dutch word form – participants had to retrieve the form of the Dutch word to produce it (see also Jescheniak, Meyer, & Levelt, 2003, Experiment 1). They translated each word three times. There were three types of Dutch targets: (1) homophones, which were low-lemma frequency words with a high-lemma frequency homophone (e.g., English *bunch* had to be translated into Dutch *bos*, which also means *forest*); (2) LF controls, which were low-lemma frequency words that did not have a homophone (e.g., *hok*, which means *kennel*); and (3) HF controls, which were high-lemma frequency words that matched the summed lemma frequency of the two members of homophone pairs (e.g., *hoek* [corner] with a frequency corresponding to the summed frequencies of *bos*[bunch] and *bos*[forest]). Before beginning the translation task, participants studied the probe words and their translations. The first step of the translation task involved recognising the visually presented English word. To account for this recognition process, response times were calculated as the difference between their latencies to decide whether the word denotated an animate or inanimate concept (a task completed a week after the translation task) and their translation times. Difference scores were larger for LF controls (M = 327 ms) than for homophones (M = 242 ms) and HF controls (M = 227). Importantly, difference scores for homophones did not differ from HF controls, suggesting that LF homophones (e.g., *bos* as *bunch*) inherited the frequency of their HF partners (e.g., *bos* as *forest*) and thus share a word form. These effects did not differ across the three presentations of the materials, suggesting there was a robust frequency effect at the word form level. Note, however, that other studies have failed to find evidence that pairs of homophones share a word form, undermining the logic of this experiment (e.g., Caramazza, Bi, Costa, & Miozzo, 2004; Caramazza, Costa, Miozzo, & Bi, 2001; Jescheniak et al., 2003).

Nevertheless, Jescheniak and Levelt provided evidence that word form, but not lemma, representations were sensitive to word frequency. This claim is primarily based on the finding that the word frequency effect was stable during picture naming (requiring both lemma and word form representations), but not during gender decision (requiring lemma representations only). In the next sections, we summarise research concerning the presence of word frequency effects at the lemma level and research concerning the stability of frequency effects during picture naming.

1.2. Frequency effects and lemma access

Jescheniak and Levelt found that gender decision latencies were not robustly affected by word frequency, thus concluding that lemma access was not frequency sensitive. However, other studies suggest word frequency does affect lemma access. For example, Navarrete, Basagni, Alario, and Costa (2006; Experiment 2) conducted a conceptual replication of Jescheniak and Levelt's gender decision experiment (Experiment 4) and found that participants were faster to judge the grammatical gender of pictures with HF than LF names. Importantly, and in contrast to Jescheniak and Levelt's results, this word frequency effect was stable across four repetitions. Furthermore, Finocchiaro and Caramazza (2006) tested whether noun frequency affected pronoun production latencies in Italian. Participants were shown written verbs in the infinitive (e.g., *portare*, which means *to bring*) followed by a picture of an object (e.g., a helmet, whose Italian name is the masculine noun *casco*, or a chair, whose Italian name is the feminine noun *sedia*) and produced an imperative sentence using the verb and a pronominal form of the object's name (e.g., the masculine word *portalo* or the feminine word *portala*, both of which mean *bring it* but differ in their grammatical gender). The authors found that participants were faster to produce the pronominal form when they were referring to a HF rather than LF noun, and this frequency effect was stable across three presentations. These findings suggest that word frequency affected lemma access, and this frequency effect was stable across multiple presentations of the same items.

In another study, Wheeldon and Monsell (1992) found that picture naming was facilitated by prior production of the picture name in response to a definition, but not by prior production of a homophone (e.g., *son*) of the picture's name (e.g., *sun*) in response to a definition. Importantly, LF words benefitted from this repetition more than HF words. The authors concluded that the word frequency effect arises from repetition, and this repetition effect occurs before word form selection. In particular, they suggested the effect is associated with the repeated activation of the word's lemma, or with the repeated activation of the link between the word's lemma and its word form (e.g., see Monsell, Matthews, & Miller, 1992). HF words benefitted from repeated presentation less than LF words because the activation levels of their lemmas or their lemma-to-word form mappings are closer to ceiling activation than LF words. Note that although these results suggest word frequency affects lemma access, it is also possible that the frequency effect arose during concept selection.

Some studies have also investigated the locus of word frequency in patients with aphasia. For example, Kittredge et al. (2008; see also Knobel, Finkbeiner, & Caramazza, 2008) found that aphasic patients were less likely to produce semantic and phonological errors when naming HF rather than LF pictures. These semantic errors canoccur at either the conceptual or the lemma level, but, importantly, they cannot occur at the word form level. Thus, these findings suggest that word frequency affects both the semantic and word form levels. Interestingly, Kittredge et al. also found that the effect of frequency on semantic errors was smaller than the effect on phonological errors (see also Nozari, Kittredge, Dell, & Schwartz, 2010). Together, these studies suggest word frequency effects are not isolated to word form representations.

1.3. The stability of word frequency effects in picture naming

Jescheniak and Levelt's claim that frequency affects word form access and not lemma access rests on finding a stable word frequency effect during picture naming but not during gender decision. Consistent with this finding, other studies have found stable frequency effects. For example, Levelt, Praamstra, Meyer, Helenius, and Salmelin (1998) found a stable frequency effect of around 40 ms across 12 presentations during picture naming in a pilot study. This pattern was not replicated when a different set of participants completed the same experiment in an MEG scanner – here, there was no word frequency effect at all, likely because naming times were much faster in the scanner. But the stable frequency effect was replicated when the same participants returned six months later and named the pictures again outside the MEG scanner. Additionally, Meyer, Sleiderink, and Levelt (1998) found a frequency effect of around 40 ms across 8 presentations during picture naming, but not during object recognition. Both of these studies used exactly the same materials as Jescheniak and Levelt (with the exception of one item, which was replaced in Meyer et al., 1998). In another study with German speakers Paucke, Oppermann, Koch, and Jescheniak (2015) found a frequency effect of around 60 ms over four presentations during picture naming.

The majority of these studies familiarised participants with the pictures before they named them. Other studies, however, have found that the word frequency effect is reduced with repeated presentation of the same items when participants are not familiarised with the materials For example, Griffin and Bock (1998; Experiment 1) had participants name HF and LF pictures three times and found a word frequency effect on the picture's first presentation (a 46 ms effect) and its second presentation (a 25 ms effect), but not on its third presentation (a 7 ms effect). Similarly, La Heij, Puerta-Melguizo, van Oostrum, and Starreveld (1999) found that the word frequency effect was larger in the first block of a picture naming experiment (first and second presentation; Experiment 1 M = 147 ms; Experiment 2 M = 78 ms) than in the second block (third and fourth presentation; Experiment 1 M = 100 ms; Experiment 2 M = 46 ms). Additionally, Wheeldon and Monsell (1992) found that participants were faster to name pictures when they had previously produced the same name in response to a definition, and LF words benefitted from this repetition more than HF words. Finally, Tsuboi, Francis, and Jameson (2021; see also Van Assche, Duyck, & Gollan, 2016) had participants complete word naming, word classification, and picture naming tasks in a training phase. When participants completed a test phase, including picture naming and word classification tasks, there was a larger priming effect for LF than HF words. Thus, there is some evidence that the word frequency effect is unstable during picture naming, inconsistent with Jescheniak and Levelt's (1994) findings.

1.4. The current study

In sum, much research has investigated the locus of the frequency effect during word production. Although a number of these studies have used similar paradigms to Jescheniak and Levelt (1994), they have not attempted to replicate the critical pattern – namely, a stable word frequency effect during picture naming but not during gender decision for the same materials.

This pattern is important because it indicates that lemma access (which is required for the gender decision task) is not frequency sensitive but word form access (which is required for the picture naming task) is. Furthermore, this pattern implies that there are manipulations that selectively affect one level of word planning (i.e., word form selection) but not the other (i.e., lemma access), suggesting the two levels of processing can be separated. This point is important because the distinction between the two levels has been repeatedly challenged in work on language production (e.g., Caramazza, 1997; Caramazza & Miozzo, 1998) and research in word comprehension typically assumes integrated word representations (e.g., Huettig, Audring, & Jackendoff, 2022). Thus, it is worth assessing whether the pattern observed by Jescheniak and Levelt can be reproduced. If it cannot, and word frequency effects decrease or remain stable across multiple presentations in both tasks, then we either have to conclude that frequency affects lemma access or that the two levels of representation cannot be separated with this combination of tasks. The latter might be the case if participants cannot determine the gender of a noun without covertly retrieving the form of the full noun phrase (e.g., Nickels, Biedermann, Fieder, & Schiller, 2015; Sá-Leite, Comesaña, Acuña-Fariña, and Fraga, 2023, for further discussion).

In the present paper, we report two experiments. The first was a conceptual replication of part of Jescheniak and Levelt's (1994) study

with new materials. We ran an object recognition task (Experiment 1a), intended to assess conceptual and perceptual processing; a picture naming task (Experiment 1b), intended to assess both lemma and word form selection; and a gender decision task (Experiment 1c), intended to assess lemma selection only. Following Jescheniak and Levelt, each picture was presented once in Experiment 1a and three times in Experiments 1b and 1c, and so we could determine whether word frequency effects were stable across multiple presentations of the same picture. Unlike Jescheniak and Levelt, we did not familiarise participants with the pictures or their names before the picture naming or gender decision task. Familiarisation is often used in picture naming studies to ensure that participants know what concepts are depicted in the pictures (for discussion see Collina, Tabossi, & De Simone, 2013; Gauvin, Jonen, Choi, McMahon, & de Zubicaray, 2018; Llorens, Trébuchon, Riés, Liégeois-Chauvel, & Alario, 2014). All of our pictures had high name agreement (at least 80 %), and so we expected participants to have no trouble identifying the concepts and selecting the expected pictures.

To preview our results, we found a word frequency effect on the picture's first presentation but not on its second or third presentation. We observed this pattern of results in both gender decision and picture naming, and so we did not replicate Jescheniak and Levelt. However, the difference in the results could be attributed to differences in the materials or the lack of familiarisation in Experiment 1. As a result, we conducted Experiment 2, which was a closer replication using the same pictures as Jescheniak and Levelt and an identical procedure, with the exception that the study was administered online.

We ran the experiments online because research suggests that web-based experiments can be successfully used to measure naming latencies and response times to visual stimuli with high accuracy (Fairs & Strijkers, 2021; Stark, van Scherpenberg, Obrig, & Abdel Rahman, 2023; Vogt, Hauber, Kuhlen, & Abdel Rahman, 2022). In both experiments, we recruited more participants than Jescheniak and Levelt to increase experimental power. They recruited 12 participants for each task and used 48 experimental items (24 HF, 24 LF). Prior to data collection for Experiment 1, we conducted a power analysis using simr (version 1.0.5) to determine our sample size. Since we recruited participants online, we used the condition means (HF = 1085 ms; LF = 1137 ms) and standard deviations (HF = 300 ms; LF = 312 ms) from Fairs and Strijkers' (2021) online study to simulate a random dataset of 40 participants with 34 HF and 34 LF pictures. With 40 participants, we reached a power estimate of 92.60 % (95 % confidence interval: 90.80, 94.15) for detecting a frequency effect of 52 ms. Note that this estimate was based on picture naming data. We recruited the same number of participants for the other tasks because the size of the word frequency effect in Navarrete et al. (2006; where frequency did affect gender decision times) was 82 ms on the picture's first presentation, 48 ms on its second presentation, and 53 ms on its third presentation. Thus, we assumed we had sufficient power to detect a word frequency effect during gender decision.

2. Experiment 1a: object recognition

In Experiment 1a, we used an object recognition task to determine whether recognition times were affected by the frequency of the object's name. Participants indicated whether or not a picture matched a previously presented word. Experimental items were chosen to elicit *yes* responses and filler items to elicit *no* responses. This task was designed to assess perceptual and conceptual processing.

2.1. Method

2.1.1. Participants

Forty native speakers of Dutch (27 females, 11 males; 2 NA; *M* age = 25 years) were recruited online using Prolific Academic, and participated in exchange for £6.87. Participants were randomly assigned to one of four stimulus lists. All participants lived in The Netherlands and had a minimum 90 % "satisfactory" rate of performance from prior assignments on Prolific. Participants reported no speech, reading, or hearing impairments. Ethical approval for the study was given by the Ethics Board of the Social Sciences Faculty at Radboud University. We discarded data from one participant because of a technical error, and so we analysed data from 39 participants.

2.1.2. Materials

We selected 68 pictures from the Dutch Bank of Standardised Stimuli (BOSS; Decuyper, Brysbaert, Brodeur, & Meyer, 2021), which is a database of coloured photographs of everyday objects (see Appendix A for a full list of items). All of these picture names were gendered (i.e., they were named using the definite determiner *de*). All descriptive statistics were taken from the BOSS database. Half of the pictures had HF names while the other half had LF names, and the two conditions differed in SUBTLEX frequency counts ($t(66) = 7.36$, $p < .001$). They also differed in Zipf frequency ($t(66) = 18.84$, $p < .001$), which is a logarithmic scale ranging from 1 (very low frequency words; frequencies of 1 per 100 million words) to 6 (very high frequency content words) or 7 (function words, pronouns, and verb forms like "have"; Van Heuven, Mandera, Keuleers, & Brysbaert, 2014). Zipf frequency is derived from the SUBTLEX frequency estimates. The two conditions were matched for object agreement ($t(66) = -1.43$, $p = .16$), which is a five-point rating of how well participants thought the picture represented its actual concept.

We also matched word prevalence, which is the percentage of the population who knows a particular word ($t(64) = 1.60$, $p = .11$). Research suggests that word prevalence has the highest correlation with lexical decision times after word frequency, and the prevalence effect is thought to have the same origin as the frequency effect (Brysbaert, Stevens, Mandera, & Keuleers, 2016). In particular, words known by many people are likely to be produced regularly, and are likely to have been encountered more often (and thus will be more frequent) than words known by fewer people. Thus, we may reduce any word frequency effects by matching the prevalence of the two conditions because the two variables overlap. However, research also suggests that frequency estimates (like SUBTLEX) do not load onto word prevalence during factor analysis (Brysbaert, Mandera, McCormick, & Keuleers, 2019), and so prevalence should be controlled when selecting stimuli. Furthermore, we wanted to ensure that any effects of word frequency occurred because words were used rarely, and not because they were words that participants did not know, especially since we did not familiarise participants with the picture names.

The two conditions were also matched for length in number of syllables ($t(66) = -1.89$, $p = .06$) and age of acquisition ($t(62) = -1.62$, $p = .11$). Where possible, we matched the word onsets of picture names in the two conditions to ensure that any differences in naming latencies could not be attributed to differences in detecting word onset. Twenty-six of the word onsets in the HF condition were matched to word onsets in the LF condition (54 names matched total; 79 %). For these experimental items, the word and the picture always matched and so participants were expected to respond *yes*. The names of all of these experimental items were gendered (i.e., used the determiner *de*).

In addition to the 68 experimental pictures, we selected 68 filler pictures. Half of these items had HF names (SUBTLEX $M = 29.83$; Zipf $M = 4.43$), while the other half had LF names (SUBTLEX $M = 0.69$; Zipf $M = 2.66$). Half of these picture names were gendered, while the other half were gender neutral (i.e., they were named using the definite determiner *het*). We did not match these conditions for any other variables because their function was to serve as *no* trials. Thus, the word and the picture did not match, and pictures were instead accompanied by a word that was semantically and phonologically unrelated to the picture's actual name. These words were the names of other filler pictures.

2.1.3. Design

Following Jescheniak and Levelt (1994), each of the 136 stimuli was presented to participants once. We created four pseudorandomised lists,

each containing 68 experimental items (requiring a *yes* response) and 68 filler items (requiring a *no* response), and 34 HF and 34 LF items from each of the experimental and filler items. We created the lists in such a way that: (1) experimental items were not immediately preceded by the presentation of a phonologically, semantically, or associatively related item; and (2) no more than five items of the same gender class were presented in adjacent trials. Participants were randomly assigned to one list.

2.1.4. Procedure

The experiment was administered online using Frinex (FRamework for INteractive EXperiments, a software package developed for running experiments by the technical group at the Max Planck Institute for Psycholinguistics). Participants were encouraged to complete the experiment in a quiet environment, away from any distractions such as phones or televisions. Each trial began with a fixation cross (+) presented in the centre of the screen for 500 ms. The word was presented 300 ms later, in lower case black Courier 32-point typeface. The word stayed on-screen for 2000 ms, and was then replaced with the target picture. Participants responded *yes* (M key on their keyboard) if the word and the picture matched, or *no* (Z key on their keyboard) if they did not. They had 2000 ms to respond. The next trial began 1500 ms later, either after the participant had responded or after the timeout.

At the start of the experiment, participants completed four practice trials (two stimuli from each of the frequency conditions) to familiarise themselves with the experimental procedure. Participants then began the main experiment. The 136 pictures were divided into two blocks of 68 trials, and participants could take a break after each block.

2.2. Data analysis

Participants' response times were measured from picture onset. Before analysis, we discarded 87 trials (1.65 %) where participants responded incorrectly (27 HF filler responses, 20 HF experimental responses, 16 LF filler responses, and 24 LF experimental responses), and 75 trials (1.42 %) where they did not respond at all (18 HF filler trials, 21 HF experimental trials, 19 LF filler trials, and 17 LF experimental trials). Note that Jescheniak and Levelt (1994) replaced such responses with estimates. However, we analysed our data using linear mixed effects models, which can deal with different numbers of observations between groups, and so we discarded these trials. We focused our analysis on response times rather than error rates because errors accounted for such a small portion of the data.

Following Jescheniak and Levelt, we analysed the experimental trials, where the word and the picture matched and participants responded *yes* (2604 trials in total). We evaluated the effects of word frequency on response times using linear mixed effects models (Baayen, Davidson, & Bates, 2008) using the *lmer* function of the *lme4* package (version 1.1–26; Bates, Maechler, Bolker, & Walker, 2021) in RStudio (version 1.2.5042). Response times were predicted by Frequency (reference level: low vs. high), which was contrast coded (−0.5, 0.5) and centered. We initially fitted models using the maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013), but a model including random effects for Frequency by-participants produced a singular fit error, likely because it explained zero variance. Thus, we removed this random effect from our analysis.

We report coefficient estimates (b), standard errors (SE), and t-values for each predictor. We assume that a t-value of ±1.96 or greater indicates significance at the 0.05 alpha level (Baayen et al., 2008). The raw data and analysis scripts are available at: https://osf.io/tw8hs/

2.3. Results and discussion

On average, participants responded 744 ms after picture onset. Participants were faster to recognise pictures with LF (M = 730 ms) rather than a HF names (M = 765 ms; b = 34.90, SE = 13.81, t = 2.53). This result runs contrary to what we would expect if objects with HF names were faster to recognise than objects with LF frequency names, as, for instance, reported by Kroll and Potter (1984). Inspection of the average object recognition times for each item showed that the effect was not driven by a few items (see Appendix B for the by-item means). To take this unexpected and unexplained finding into account, we included object recognition time as a fixed effect (i.e., a covariate) in all further analyses using these stimuli.

3. Experiment 1b: Picture naming

In Experiment 1b, we tested whether participants would be faster to name pictures with HF than LF names, and whether this word frequency effect would be stable across multiple presentations of the same picture.

3.1. Method

3.1.1. Participants

Forty-six native speakers (23 females, 20 males, 3 non-binary; *M*age = 27.22 years) were recruited using the same procedure as Experiment 1a. Participants received £5.15 for completing the study. None of these participants took part in Experiment 1a. Data from seven participants were discarded because their audio failed to record or was unintelligible. Thus, we analysed data from 39 participants.

3.1.2. Materials, design, and procedure

Experiment 1b used the same materials as Experiment 1a and a similar procedure. Following Jescheniak and Levelt (1994), each of the 136 stimuli was presented to participants three times, and so participants saw a total of 408 pictures. We created four pseudorandomised lists. Each list contained 68 experimental items and 68 filler items, with 34 HF and 34 LF items from each of the experimental and filler items. We created lists in such a way that: (1) an experimental item was not immediately preceded by a phonologically, semantically, or associatively related item; (2) no more than five items of the same gender class were presented in adjacent trials; and (3) repeated presentations of an individual item were separated by at least 20 trials. Participants were randomly assigned to one list.

Each trial started with a fixation cross (+) presented in the centre of the screen for 500 ms. The target picture was presented after a 300-ms blank interval. Participants had 2000 ms to name the picture before it disappeared and the next trial began automatically. If they named the picture before the end of the 2000 ms, they could click a “Volgende” (or “next”) button on-screen to begin the next trial. The next trial began 1500 ms later, either after the timeout or after participants had pressed the button. Each picture was preloaded at the start of the trial and audio recording began only once the image was presented. Thus, we ensured there were minimal delays in image presentation once the trial started.

Participants checked their microphone was working by creating a test recording (they could say whatever they wished). They then listened to the audio playback to ensure that they could clearly hear themselves. If there were problems with the audio, they were instructed to refresh the page, move closer to their microphone, and ensure they had enabled microphone permissions. Participants then completed four practice trials (two stimuli from each of the frequency conditions) to familiarise themselves with the experimental procedure, and then began the main experiment. The 408 pictures were divided into three blocks of 136 trials, and participants were given the opportunity to take a break after each block.

3.2. Data analysis

Picture naming times were measured from picture onset, and were manually measured in Praat by trained Dutch speakers. Before analysis, we discarded 706 trials (4.43 %) because participants either did not provide an answer within the 2000 ms time limit, or because the audio

file was corrupt and we could not determine what the participant had said. Following Jescheniak and Levelt (1994), we focused our analysis on the experimental items, which were items designed to elicit *yes* responses in Experiment 1a (7669 trials in total). We discarded 753 trials (9.82 %) where participants named the picture other than expected, produced a nonspeech sound, a disfluency, or an utterance repair (431 HF filler trials, 298 HF experimental trials, 431 HF filler trials, and 455 LF experimental trials). This left us with 6916 trials for analysis.

We analysed the data using the same procedure as in Experiment 1a, but we included Presentation (1, 2, or 3) and its interaction with Frequency as fixed effects in the analysis. We did not include the interaction between Frequency and Presentation in the random effects structure because doing so resulted in a singular fit error. Given that participants were slower to recognise HF than LF pictures in Experiment 1a, we also included each picture's average Object recognition time as a fixed effect in our analysis.

3.3. Results and discussion

On average, participants responded 941 ms (Fig. 1) after picture onset. Participants were faster to name pictures with HF ($M = 930$ ms) than LF names ($M = 961$ ms; $b = -108.31$, $SE = 26.75$, $t = -4.05$). We also found a significant effect of Presentation – participants' naming latencies decreased with each presentation (presentation 1 $M = 1025$ ms; presentation 2 $M = 920$ ms; presentation 3 $M = 894$ ms; $b = -68.80$, $SE = 5.42$, $t = -12.69$). There was also a positive relationship between Object Recognition time and picture naming latencies ($b = 0.57$, $SE = 0.21$, $t = 2.72$).

We also found an interaction between Frequency and Presentation ($b = 27.43$, $SE = 8.33$, $t = 3.29$). We followed up this interaction by fitting separate models to the first, second, and third presentation of each picture. In each of these models, naming latencies were predicted by Frequency. We included Object recognition time when analysing the picture's first presentation, but not when analysing the second or third presentation – in these latter cases, the participant has already recognised the object on its first presentation and the picture is familiar to the participant. We used the maximal random effects structure, including Frequency as a by-participant random effect. These analyses showed that participants were faster to name pictures with HF than LF names on their first presentation ($b = -94.36$, $SE = 26.70$, $t = -3.53$), but not on their second ($b = -11.26$, $SE = 16.14$, $t = -0.69$) or third presentation ($b = -17.62$, $SE = 14.10$, $t = -1.25$). Although our overall frequency effect, of 31 ms, is smaller than Jescheniak and Levelt's (62 ms), it is worth noting that the frequency effect on the picture's first presentation (76 ms) is larger than the overall frequency effect reported in that study, showing that we elicited a strong frequency effect.

In sum, participants were faster to name pictures with HF than LF names, suggesting word frequency affects lexical selection. But this frequency effect occurred only on the picture's first presentation – on the second and third presentation, participants were just as fast to name HF as LF pictures. Our findings are inconsistent with Jescheniak and Levelt (1994), who found that the frequency effect was stable across multiple presentations during picture naming. We return to this difference in the Interim Discussion.

4. Experiment 1c: gender decision

In Experiment 1c, we used a gender decision task to test whether lemma selection was affected by word frequency. In particular, participants decided whether a picture's name was gendered (a *de* word in Dutch) or gender-neutral (a *het* word in Dutch). They made these decisions three times for each picture. The interaction between word frequency and presentation is crucial in this experiment. Jescheniak and Levelt (1994) found that frequency affected gender decision times on a picture's first presentation, but not on its second or third presentation. They proposed that this effect was a recency effect, and concluded that word frequency did not affect lemma access. However, this conclusion rests on finding a stable frequency effect in picture naming, which we did not find in Experiment 1b.

4.1. Method

4.1.1. Participants

Forty native speakers of Dutch (24 females, 13 males, 1 non-binary, 2 NA; *M*age = 24.87 years) were recruited using the same procedure as Experiment 1a. They participated in exchange for £5.15. None of these participants took part in Experiments 1a or 1b.

4.1.2. Materials, design, and procedure

Experiment 1c used the same materials as Experiment 1b and a

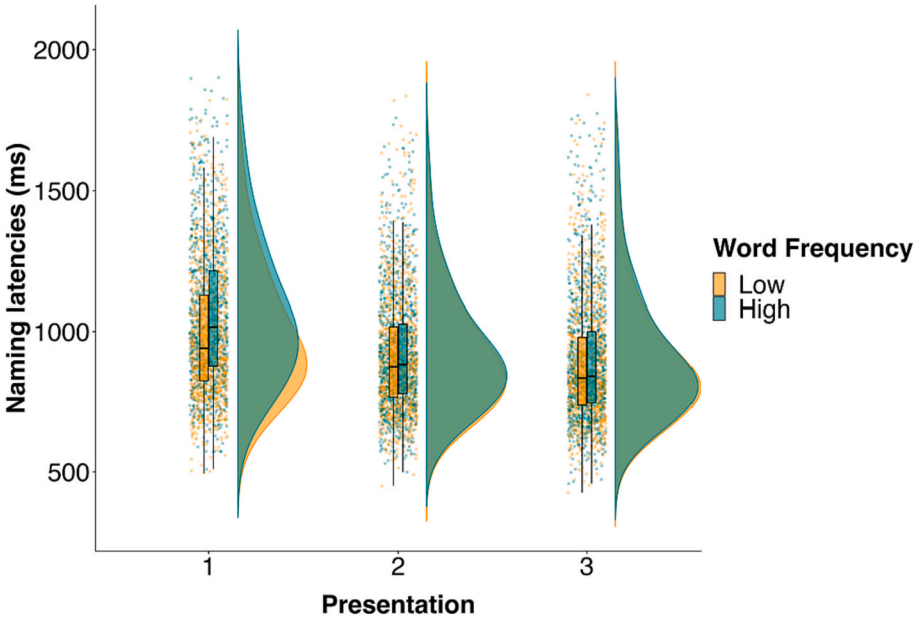


Fig. 1. Distribution of naming latencies (ms) for pictures with high and low frequency names on each of the three presentations in Experiment 1b. Individual dots show individual datapoints for each Frequency condition on each Presentation.

similar procedure. The only difference was that participants did not name the pictures; instead, they pressed a button to indicate the gender of the picture name. In the instructions gendered and neuter nouns were introduced as *de*-words and *het*-words, respectively. If the picture name was a *de* word, participants pressed *M* on their keyboard. If it was a neuter word, they pressed *Z*. The keys (and their corresponding determiners) were displayed at the bottom of the screen each time a picture was presented, and so participants did not need to remember which keys to press. A new trial began 1500 ms after the participant responded or after an interval of 2000 ms.

As described above, the materials included 34 pictures with neuter names (all of them fillers), and 102 pictures with gendered names (34 fillers and all 68 experimental items. Thus, the probability of a *de* response was 75 %. This corresponds roughly to the proportion of *de* nouns in the language (e.g., La Heij, Mak, Sander, and Willeboordse, 1998).

4.2. Data analysis

Gender decision times were measured from picture onset. Before analysis, we discarded 1381 trials (8.46 %) where participants responded incorrectly (411 HF filler trials, 128 HF experimental trials, 731 LF filler trials, 111 LF experimental trials) and 385 trials (2.36 %) where participants did not respond within 2000 ms (94 HF filler trials, 61 HF experimental trials, 132 LF filler trials, 98 LF experimental trials). The incorrect trials represented a small subset of the experimental data, and so, as in Experiment 1a, we did not run an accuracy analysis. Following Jescheniak and Levelt (1994), we focused our analysis on the experimental trials, where all the objects were gendered (7991 trials in total). We analysed the data using the same procedure as Experiment 1b, except our random effects structure included by-participant and by-item random effects for Presentation only.

4.3. Results and discussion

On average, participants responded 958 ms after picture onset (see Fig. 2). Participants were faster to judge the picture's gender when it had a HF ($M = 943$ ms) rather than a LF name ($M = 975$ ms; $b = -95.34$, $SE = 24.12$, $t = -3.95$). This finding is consistent with Jescheniak and Levelt (1994), who found an overall frequency effect of 36 ms. We also found a significant effect of Presentation – participants' gender decision times decreased with each presentation (presentation 1 $M = 1125$ ms; presentation 2 $M = 932$ ms; presentation 3 $M = 823$ ms; $b = -152.08$, $SE = 7.85$, $t = -19.39$). As the average time taken to recognise the object increased, gender decision times also increased ($b = 0.54$, $SE = 0.18$, $t = 3.05$).

In addition, there was an interaction between Frequency and Presentation ($b = 24.55$, $SE = 7.78$, $t = 3.16$). We followed up this interaction using the same procedure as Experiment 1b, including Object recognition time as a fixed effect for the first presentation. We included Frequency as a by-participant random effect for models fitted to presentations 1 and 2, but not to presentation 3 because doing so resulted in singular fit error. These analyses showed that participants were significantly faster to make gender decisions for HF rather than LF pictures on their first presentation ($b = 79.09$, $SE = 26.29$, $t = -3.01$), but not on their second ($b = -26.30$, $SE = 22.67$, $t = -1.16$), or third presentation ($b = -11.55$, $SE = 17.06$, $t = -0.68$).

In sum, participants were faster to make gender decisions to pictures with HF than LF names. But, just like in Experiment 1b, this frequency effect was significant only on the picture's first presentation. These findings are consistent with Jescheniak and Levelt (1994), who found a similar interaction between frequency and presentation. They interpreted this effect as a recency effect, and concluded that word frequency does not affect lemma access. However, this interpretation rests on finding a stable frequency effect during picture naming, which we did not find in Experiment 1b. Instead, we found evidence for a frequency effect on the first presentation of the materials but not on the following presentations in both experiments.

To formally assess the similarity of the results of Experiments 1b and 1c, we conducted a combined analysis using a similar procedure to our individual analyses. In particular, response times were predicted by Frequency, Presentation, and their interaction with Experiment (reference level: picture naming vs. gender decision). Using the maximal random effects structure resulted in a singular fit error, and so the final model included by-participant random effects for Presentation and by-item random effects for the interaction between Experiment and Presentation. This analysis confirmed the results from the two individual experiments, showing effects of Frequency ($b = -14.71$, $SE = 4.71$, $t = -3.13$), Presentation ($b = -113.75$, $SE = 5.57$, $t = -20.42$), and an interaction between the two ($b = 4.70$, $SE = 1.42$, $t = 3.31$). Importantly,

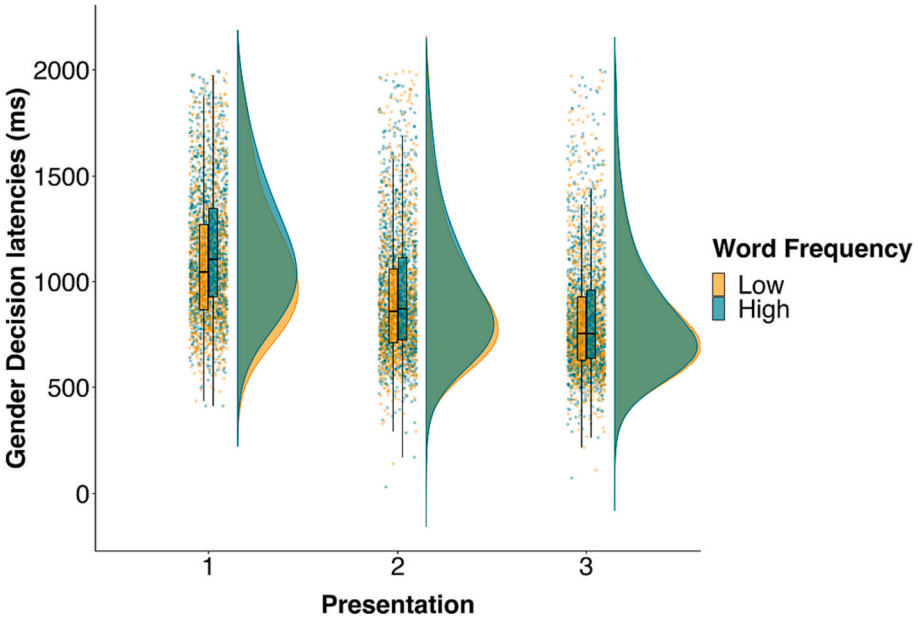


Fig. 2. Distribution of gender decision latencies (ms) for pictures with high and low frequency names on each of the three presentations in Experiment 1c. Individual dots show individual datapoints for each Frequency condition on each Presentation.

there was no two-way interaction between Frequency and Experiment ($b = 1.72$, $SE = 4.18$, $t = 0.41$) and no three-way interaction with Presentation ($b = -0.56$, $SE = 1.65$, $t = -0.34$), confirming that the word frequency effect was similar in the two tasks.

5. Interim discussion

In sum, in Experiment 1 we failed to fully replicate the results reported by Jescheniak and Levelt. In particular, participants were faster to name (Experiment 1b) and determine the grammatical gender (Experiment 1c) of pictures with HF than LF names. In both cases, this frequency effect occurred only on the picture's first presentation. Thus, there was an interaction between frequency and presentation in both tasks. There are many reasons why our results may differ from Jescheniak and Levelt's. First, we did not familiarise the participants with materials. Second, we used different items, which turned out to be slightly problematic because the HF items were unexpectedly recognised more slowly than the LF items. To address these issues, we ran Experiment 2, which featured a familarisation phase and used the same pictures as Jescheniak and Levelt's study and an identical procedure.

6. Experiment 2a: object recognition

6.1. Method

6.1.1. Participants

Forty-one native speakers (18 females, 22 males; *M*age = 28.28 years) were recruited using the same procedure as Experiment 1. All participants received £3.45.

6.1.2. Materials

We used the same materials as Jescheniak and Levelt (1994). In particular, we retrieved the original 48 experimental pictures from a picture database at the Max Planck Institute. All pictures were black and white line drawings of simple objects. Half of the pictures had LF names (mean token lemma frequency of 6.0 per million in the Celex database), while the other half had HF names (mean token lemma frequency of 150.7 per million). We also determined the SUBTLEX frequencies for these words. The HF words had an average SUBTLEX frequency of 93.62, which is higher than Experiment 1 (see Table 1). The LF words had an average of 3.85, which is similar to Experiment 1. The HF and LF picture names were matched for word length and word onset (see Jescheniak & Levelt, p. 828). The names of all these experimental items were gendered (i.e., used the determiner *de*). These trials required *yes* responses, and so the picture and word matched.

These experimental items were mixed with 48 filler items, which were all gender neutral (i.e., used the determiner *het*). Jescheniak and Levelt do not provide a list of filler items, and so we could not select the exact same items for the filler trials. However, we used the same selection criteria: we selected filler items that covered a wide frequency range and that belonged to similar semantic domains as the experimental items. Note that there was a limited set of *het* items in the picture database, and so there was likely some overlap between Jescheniak and Levelt's fillers and our own fillers. Importantly, we did not include the filler trials in the data analysis, and so any difference in the filler items are unlikely to explain any difference in the results between the two studies. The filler trials required *no* responses in the object recognition task, and so the picture and the word did not match. Instead, each picture was accompanied by a word that was semantically and phonologically unrelated to the picture's actual name. These words were the names of other filler pictures. We also selected ten practice pictures – one half with a *de* name and one half with a *het* name. *de* practice trials required a *yes* response while *het* trials required a *no* response.

Following Jescheniak and Levelt (1994), each picture was presented once. We created four pseudorandomised lists, each containing the ten practice items, followed by 48 experimental items (requiring a *yes* response) and 48 filler items (requiring a *no* response). We created lists in such a way that: (1) experimental items were not immediately preceded by the presentation of a phonologically, semantically, or associatively related item; and (2) no more than five items of the same gender class were presented in adjacent trials. Participants were randomly assigned to one list.

6.1.3. Procedure

We followed an identical procedure to Jescheniak and Levelt (1994, p. 830), with the exception that the experiment was administered online using Frinex. Each trial began with the presentation of a word, which was displayed in the centre of the screen in lowercase Times Roman 35-point typeface. Individual characters were separated by blank spaces. The word was displayed for 1000 ms. After a pause of 200 ms, the target picture was displayed in the centre of the screen. Participants responded *yes* (*M* key on their keyboard) if the word and the picture matched, or *no* (*Z* key on their keyboard) if they did not. The next trial began 1500 ms after a response was registered or after a 2000 ms timeout.

At the start of the experiment, participants completed ten practice trials to familiarise themselves with the experimental procedure. Participants then began the main experiment and were presented with the 96 test items.

6.2. Results and discussion

The data were analysed using the same procedure as Experiment 1a. Note that Jescheniak and Levelt (1994) analysed their data using by-participant and by-item ANOVAs, but we used linear mixed effects models because they allow us to account for by-participant and by-item variance in one analysis. All responses longer than 2000 ms and those deviating from a participant's and an item's mean by more than two standard deviations were coded as errors. Jescheniak and Levelt replaced these values with estimates, but it is not clear which estimates they used. Thus, we replaced these values with the upper limit. A total of 45 responses (1.14 %; 11 HF experimental trials, 11 LF experimental trials, 23 filler trials) were replaced and marked as errors because they were above both the by-participant and by-item upper limit. A further 90 responses (2.29 %; 22 HF experimental items, 29 LF experimental

Table 1
Maximum, minimum, means, and standard deviations of frequency measures (SUBTLEX and Zipf), name agreement, word prevalence, object agreement, syllable length, and age of acquisition for the high- and low-frequency pictures.

	High-frequency				Low-frequency			
	Maximum	Minimum	Mean	SD	Maximum	Minimum	Mean	SD
SUBTLEX	156.92	11.71	40.57	31.13	2.97	0.11	1.28	0.87
Zipf	5.20	4.07	4.51	0.28	3.48	2.14	2.98	0.38
Name agreement	100	84	93.12	5.15	100	82	94.27	5.84
Word prevalence	1.96	1.63	1.88	0.07	1.96	1.63	1.85	0.08
Object agreement	4.82	3.16	4.33	0.42	4.92	3.72	4.46	0.31
Syllable length	4	1	1.62	0.82	4	1	1.97	0.72
Age of acquisition	8.59	3.95	5.87	1.22	7.50	4.99	6.29	0.75

items, 39 filler trials) were replaced because they were above the by-item upper limit, and 136 (3.46 %; 26 HF experimental trials, 20 LF experimental trials, 90 filler trials) were replaced because they were above the by-participant upper limit.

Our analysis focused on the experimental trials, where the word and the picture matched and participants responded *yes* (1968 trials). We first evaluated the effects of word frequency on error rates with generalised linear mixed effects models (Baayen et al., 2008) using the *glmer* function of the *lme4* package. Error rates were predicted by Frequency (reference level: low vs. high), which was contrast coded (−0.5, 0.5) and centered. We initially fitted a model using the maximal random effects structure, but the model returned a singular fit error even when we used the simplest random effects structure and included only by-participant and by-item random effects, likely because there was little by-item variance. Importantly, however, there was no difference in error rates for the HF ($M = 7.21$ %) and LF ($M = 7.51$ %; $b = 0.05$, SE 0.17, $p = .79$).

We analysed response times using the *lmer* function. Again, we initially fitted models using the maximal random effects structure, but this model returned a singular fit error. As a result, we removed by-participant random effects for Frequency because it explained little variance. Thus, the final model included by-participant and by-item intercepts only. On average, participants responded 509 ms after picture onset. There was no significant difference in object recognition times for HF ($M = 503$ ms) and LF ($M = 515$ ms) pictures ($b = -11.60$, $SE = 13.55$, $t = -0.86$). These findings are consistent with Jescheniak and Levelt's (1994) results and suggest that frequency did not affect the speed or accuracy of object recognition for these materials.

7. Experiment 2b: picture naming

7.1. Method

7.1.1. Participants

We recruited 42 participants using the same procedure as in Experiment 1b (20 females, 21 males, 1 non-binary; *M*age = 31.62 years). We discarded data from one participant because their audio files were bad quality and difficult to annotate, and so we analysed the data from 41 participants. All participants received £5.30.

7.1.2. Materials and procedure

Experiment 2b used the same materials as Experiment 2a, but each picture was presented three times and so participants saw 30 practice pictures and 288 test pictures. As in Experiment 2a, we created four pseudorandomised lists. We created lists in such a way that: (1) an experimental item was not immediately preceded by a phonologically, semantically, or associatively related item; (2) no more than five items of the same gender class were presented in adjacent trials; and (3) repeated presentations of an individual item were separated by at least 20 trials (except for the practice trials, which were all presented at the beginning of the experiment). Participants were randomly assigned to one list.

Each trial started with a fixation point (*) presented in the centre of the screen for 200 ms. After a pause of 600 ms, the picture was presented in the middle of the screen. Participants had 2000 ms to name the picture before it disappeared and the next trial began automatically. In Jescheniak and Levelt's (1994) study, the picture disappeared once the microphone registered a vocal response. We could not follow this procedure because we could not implement a voice-key online. Instead, participants were given the opportunity to click a "Volgende" (or "next") button on the screen if they named the picture before the end of the 2000 ms interval and wanted to begin the next trial. Even though participants were given this option, they still preferred to await the end of the trial (70 % of trials). The next trial began 1500 ms after a timeout or after participants had pressed the button. Each picture was preloaded at the start of the trials to ensure there were no delays in image presentation once the trial started, as in Experiment 1b.

Before beginning the experiment, participants studied a set of instructions that emphasised both the speed and accuracy of their responses. In particular, they were told to respond quickly and to name the picture using the name presented to them during familiarisation. During this familiarisation phase, participants studied the pictures and their names. This phase was split across six pages, with the first five pages containing 18 pictures and the last page containing 17 pictures. Pictures were presented in alphabetical order in a 3 × 6 grid, with the picture's name presented beneath it. Participants could study the pictures as long as they wished, and pressed a button on-screen to view the next set of pictures. Once they reached the last set, they pressed a button on-screen to check their microphone was recording and begin the experiment.

As in Experiment 1b, participants checked their microphone was working by creating a test recording (they could say whatever they wished). They then completed 30 practice trials to familiarise themselves with the experimental procedure, and then began the main experiment. Participants were given the opportunity to take a break halfway through (i.e., after 144 test trials).

7.2. Results and discussion

Picture naming times were measured from picture onset using the same procedure as Experiment 1b. Following Jescheniak and Levelt (1994), we discarded (a) 252 trials (2.19 %) because participants either did not provide an answer within the 2000 ms time limit, or because the audio file was corrupt and we could not determine what the participant had said; (b) 643 trials (5.58 %; 96 HF experimental trials, 86 LF experimental trials; 461 filler trials) because participants named the picture other than expected; (c) 36 trials (0.31 %; 6 HF experimental trials, 9 LF experimental trials, 21 filler trials) because participants produced a disfluency, a non-speech sound, or their audio was and cut-off; and (2) 579 trials (5.03 %; 141 HF experimental trials, 145 LF experimental trials, 293 filler trials) because speech onset latency deviated from a participant's or an item's mean by more than two standard deviations. We focused our analysis on the experimental trials (5041 trials in total).

We analysed naming times using the same procedure as Experiment 1b, but we did not include Object Recognition Time as a fixed effect because there was no significant difference in object recognition times for HF and LF items in Experiment 2a. We included by-participant random effects for Frequency and Presentation and by-item random effects for Presentation. We did not include the interaction in the by-participant random effects because doing so returned a singular fit error.

On average, participants responded 819 ms after picture onset (Fig. 3). Participants were faster to name pictures with HF ($M = 808$ ms) than LF names ($M = 835$ ms; $b = -50.88$, $SE = 20.96$, $t = -2.43$). Note that our Frequency effect is smaller than Jescheniak and Levelt's (who found an average difference of 62 ms), but it was similar in size to Experiment 1b. We also found a significant effect of Presentation – participants' naming latencies decreased with each presentation (presentation 1 $M = 854$ ms; presentation 2 $M = 812$ ms; presentation 3 $M = 801$ ms; $b = -29.85$, $SE = 5.18$, $t = -5.76$). The Presentation effect in our study is very similar to the Presentation effect in Jescheniak and Levelt's study – they found an average difference in naming latencies of 60 ms from presentation 1 to 2 and 19 ms from presentation 2 to 3.

Consistent with Experiment 1b, we found an interaction between Frequency and Presentation ($b = 10.80$, $SE = 4.58$, $t = 2.36$). We followed up this interaction by fitting separate models to the first, second, and third presentation of each picture. In each of these models, naming times were predicted by Frequency and we used the maximal random effects structure. We did not find a significant Frequency effect on any individual presentations, but the interaction occurred because the Frequency effect was numerically larger on the picture's first presentation ($b = -46.14$, $SE = 24.05$, $t = -1.92$) than on the second ($b = -19.76$, $SE = 18.85$, $t = -1.05$) or third presentation ($b = -22.84$, $SE = 17.21$, $t = -1.33$). Thus, we again found an unstable word frequency effect during

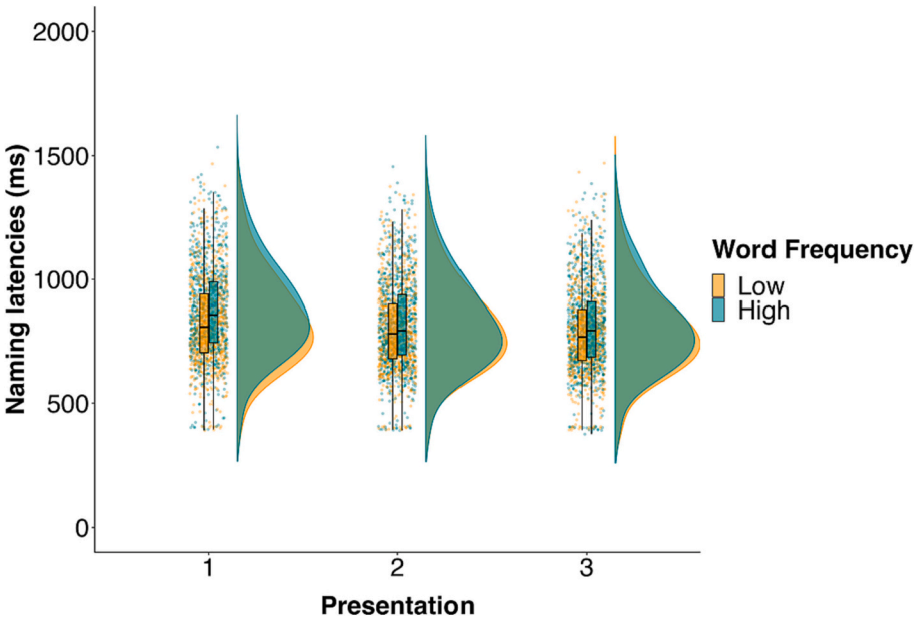


Fig. 3. Distribution of naming latencies (ms) for pictures with high and low frequency names on each of the three presentations in Experiment 2b. Individual dots show individual datapoints for each Frequency condition on each Presentation.

picture naming, now after familiarisation and with Jescheniak and Levelt's (1994) materials. Note that the interaction in Experiment 2b was smaller than Experiment 1b (a beta coefficient of 26.91), likely because participants were familiarised with the pictures and their names in Experiment 2b, thus reducing the impact of presentation.

8. Experiment 2c: gender decision

8.1. Method

8.1.1. Participants

We recruited 41 participants using the same procedure as Experiment 1c (17 females, 23 males, 1 non-binary; *M*age = 30.20 years). We discarded data from one participant because they did not provide any typed responses when they named the pictures after the main experiment, and so our analysis focused on data from 40 participants. All participants received £5.30.

8.1.2. Materials and procedure

Experiment 2c used the same materials as Experiment 2b and a similar procedure. The only difference was that participants did not name the pictures; instead, they pressed a button to indicate the gender of the picture name. If the picture name was a *de* word, then they pressed *M* on their keyboard. If it was a *het* word, then they pressed Z. The keys (and their corresponding determiners) were displayed at the bottom of the screen each time a picture was presented, and so participants did not need to remember which keys to press. Participants were instructed that each response would be required equally often. The next trial began 1500 ms later, either after the participant had responded or after a 2000 ms timeout.

Before beginning the experiment, participants studied the pictures. We followed the same procedure as Experiment 2b, but participants did not see the picture's names. After the main task, participants were presented with the pictures (one at a time) and were instructed to type the object's name. Thus, we could determine which picture names the participants actually used in the experiment (and based their gender decisions on). If participants provided a name other than expected, then the corresponding observations were excluded from statistical analysis.

8.2. Results and discussion

Gender decision times were measured from picture onset. Following Jescheniak and Levelt (1994), we discarded (a) 237 trials (2.01 %) because participants did not respond within the 2000 ms time limit; (b) 1283 trials (10.87 %; 144 HF experimental trials, 146 LF experimental trials, 993 filler trials) because participants responded incorrectly; (c) 798 trials (6.76 %; 127 HF experimental, 123 LF experimental, 548 filler trials) because participants named the picture other than expected; (d) and 467 trials (3.96 %; 128 HF experimental trials, 131 LF experimental trials, 208 filler trials) because gender decision times deviated from a participant's or an item's mean by more than two standard deviations. We focused our analysis on the experimental trials (4868 trials total).

We analysed gender decision times using the same procedure as in Experiment 1c, but we did not include Object Recognition Time as a fixed effect. We included by-participant and by-item random effects for Presentation only because including Frequency (and its interaction with Presentation) returned a singular fit error.

On average, participants responded 828 ms after picture onset (Fig. 4). As in Experiment 1c, participants were faster to judge the picture's gender when it had a HF ($M = 815$ ms) rather than a LF name ($M = 842$ ms; $b = -76.62$, $SE = 21.46$, $t = -3.57$). This finding is again consistent with Jescheniak and Levelt, who found an overall frequency effect of 36 ms. As in Experiment 1c, we found a significant effect of Presentation – gender decision times decreased with each presentation (presentation 1 $M = 878$ ms; presentation 2 $M = 823$ ms; presentation 3 $M = 788$ ms; $b = -47.12$, $SE = 4.62$, $t = -10.20$). Note that the effect of presentation in this experiment is smaller than in Jescheniak and Levelt's gender decision experiment (presentation 1 $M = 829$ ms; presentation 2 $M = 737$ ms; presentation 3 $M = 688$ ms), but both studies show a larger decrease in gender decision latencies from presentation 1 to presentation 2 than from presentation 2 to presentation 3.

Consistent with Jescheniak and Levelt (1994) and Experiment 1c, we found an interaction between Frequency and Presentation ($b = 25.13$, $SE = 7.20$, $t = 3.49$). We followed up this interaction by fitting separate models to the first, second, and third presentation of each picture. In each model, gender decision times were predicted by Frequency. We used the maximal random effects structure for the model fitted to presentation 2, but not for the models fitted to presentations 1 or 3 because doing so returned a singular fit error. Participants were faster to make

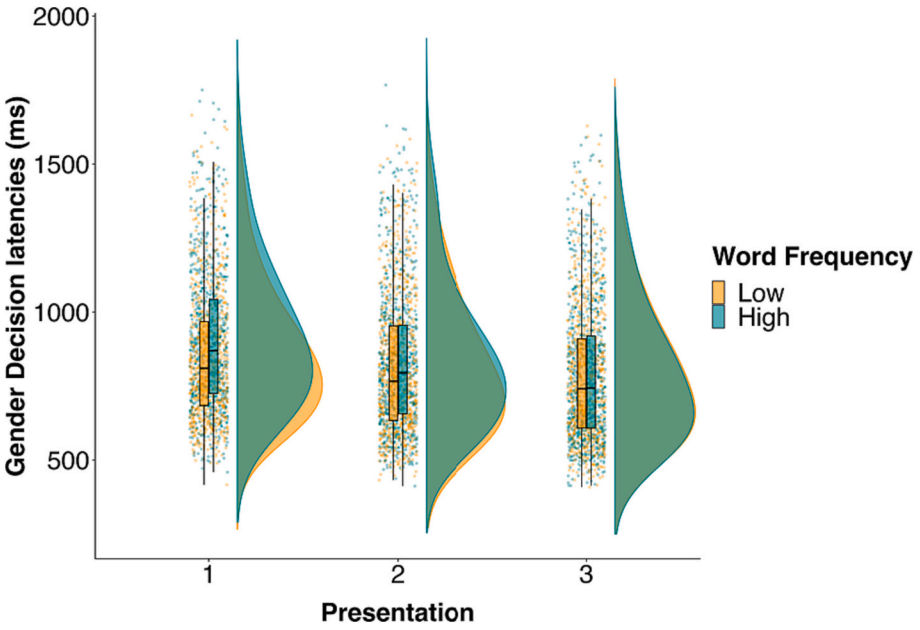


Fig. 4. Distribution of gender decision latencies (ms) for pictures with high and low frequency names on each of the three presentations in Experiment 2c. Individual dots show individual datapoints for each Frequency condition on each Presentation.

gender decisions for HF than LF pictures on their first presentation ($b = -60.65$, $SE = 24.16$, $t = -2.51$), but not on the second ($b = -18.30$, $SE = 17.53$, $t = -1.04$) or third presentation ($b = -5.74$, $SE = 16.24$, $t = -0.35$). Thus, we replicated Experiment 1c and found an unstable word frequency effect during gender decision.

We again assessed the similarity of the results of Experiments 2b and 2c by conducting a combined analysis using the same procedure as Experiment 1. Using the maximal random effects structure resulted in a singular fit error, and so the final model included by-participant random effects for Frequency and Presentation (but not their interaction) and by-item random effects for the interaction between Experiment and Presentation. This analysis confirmed the results from the two individual experiments, showing effects of Frequency ($b = -15.57$, $SE = 3.27$, $t = -4.67$), Presentation ($b = -40.43$, $SE = 3.69$, $t = -10.97$), and an interaction between the two ($b = 3.41$, $SE = 0.81$, $t = 4.19$). Importantly, there was no two-way interaction between Frequency and Experiment ($b = -2.74$, $SE = 3.97$, $t = -0.70$) and no three-way interaction with Presentation ($b = 2.03$, $SE = 1.45$, $t = 1.40$), confirming that the word frequency effect was similar in the two tasks.

9. General discussion

In two experiments, we conducted a conceptual and a closer replication of three tasks reported by Jescheniak and Levelt (1994) to investigate the locus of the word frequency effect during word production. In Experiment 1a, participants were unexpectedly slower to recognise HF than LF stimuli, and we controlled for this difference in Experiments 1b and 1c. In Experiment 2a, there was no difference in object recognition times for HF and LF pictures. Importantly, participants were faster to name (Experiments 1b and 2b) and determine the grammatical gender (Experiments 1c and 2c) of pictures with HF rather than LF names. In both cases, the frequency effect was larger on the picture's first presentation than on the second and third presentations. Thus, there was an interaction between frequency and presentation in both tasks.

The similarity of the results for the picture naming and gender decision tasks contrasts with the results of Jescheniak and Levelt, who found a stable word frequency effect during picture naming but not during gender decision. They suggested that the interaction between presentation and word frequency in the gender decision task reflected participants' accommodation to the task, rather than a robust frequency effect. In particular, they proposed that participants silently generated full noun phrases (i.e., determiner plus noun, as in *de hond*) on the picture's first presentation, and determined the picture's gender by monitoring for the determiner in their inner speech. This process involved access to the word's form, resulting in a word frequency effect. Participants became more efficient on subsequent presentations, deriving the gender of the nouns without accessing the word forms, and so no frequency effect occurred. This pattern contrasted with the stable word frequency effect observed in picture naming, which always required word form access. Consistent with common practice at the time, Jescheniak and Levelt did not statistically assess whether the two experiments did in fact yield different pattern of results. Based on the informal comparison of the pattern, they concluded that lemma access (required during gender decision) is not frequency sensitive, but word form access (required during picture naming) is.

We did not replicate the dissociation – a stable word frequency effect for picture naming and a short-lived frequency effect for gender decision – in our experiments. In fact, the similarity of the results in both picture naming and gender decision suggests that they tap into similar representations that are sensitive to word frequency. But it is impossible to say whether these are lemma representations, word form representations, or both. To elaborate, one possibility is that frequency affects lemma access, and so the pattern of results is similar in the picture naming and gender decision tasks because both tasks tap into these representations. This explanation is consistent with other experimental studies that have concluded that frequency affects access to lexical representations that capture the semantic and grammatical properties of words (e.g., Finocchiaro & Caramazza, 2006; Navarrete et al., 2006).

This explanation rests on the assumption that the gender decision task does not involve word form access, but purely involves lemma access. An alternative explanation is that both picture naming and gender decision tap into word form representations. Instead of relying on lemmas, which encode grammatical information, participants make gender decisions by silently generating full noun phrases and monitoring for the determiner in their inner speech, thus accessing the determiner and noun's word form. Jescheniak and Levelt (1994) proposed that participants adopted this strategy only on the picture's first presentation, but it

is possible that participants in our study used it on all presentations. Results from Starreveld and La Heij (2004) are consistent with this suggestion. In a series of picture-word interference experiments with Dutch speakers, they found that phonologically related distractors facilitated retrieval of gender information during determiner production (either *de* or *het*), determiner and noun production (e.g., *de kat* or *het boek*), and gender decisions, even after participants were familiarised with the materials. Similarly, Navarrete et al. (2006) found a word frequency effect that was stable across multiple presentations in a gender decision study conducted with Spanish speakers. Together with our results, these findings are consistent with Jescheniak and Levelt's claim that frequency affects word form selection. However, this interpretation implies that speakers activate both grammatical gender and word form information in tasks requiring access to grammatical gender information only.

Finally, it is possible that word frequency affects both lemma and word form retrieval, to the same or differing degrees. This view is consistent with approaches that assume unitary lexical representations (e.g., Huettig et al., 2022) or layered representations that rapidly co-activate each other (e.g., Strijkers & Costa, 2011; see also Zwitserlood, Bölte, Hofmann, Meier, and Dobel, 2018). It is also consistent with work involving patients with aphasia, which showed word frequency effects at multiple levels of production (e.g., Kittredge et al., 2008; Knobel et al., 2008). On the basis of our data, we cannot distinguish between these options. The important point is that we do not replicate Jescheniak and Levelt's findings that the frequency effect is more stable for naming than gender decision, and so we cannot conclude that frequency affects only word form access.

To determine whether word frequency affects word form access, lemma access, or both, further work is needed to establish which representations are activated during gender decision and, more generally, whether speakers can selectively activate a word's lemma without activating the associated word form. It is worth noting that there is much theoretical, computational, and neurobiological work supporting the distinction between lemmas and word forms (e.g., Indefrey, 2011; Indefrey & Levelt, 2004; Roelofs, 1992, 1997; see Kemmerer, 2019, for a review; but see Caramazza, 1997; Caramazza & Miozzo, 1998, for an alternative view). But to the best of our knowledge, there is only scant evidence, namely from studies of the tip-of-the-tongue (TOT) phenomenon, that speakers can access lemma information alone. For example, speakers in a TOT state cannot recall a particular word, even though they know the word, but they can sometimes recall syntactic information about that word, such as its grammatical gender (e.g., Iwasaki, Vigliocco, & Garrett, 1998; Vigliocco, Antonini, & Garrett, 1997), or whether it is a count or a mass noun (e.g., Vigliocco, Vinson, Martin, & Garrett, 1999).

But why do we find an interaction between word frequency and presentation in both picture naming and gender decision tasks, while Jescheniak and Levelt find a stable frequency effect during picture naming? This difference cannot be attributed to properties of the materials or the experimental procedure because we found the interaction in Experiment 1, with new materials, and in Experiment 2, which was a close replication of Jescheniak and Levelt's study with their materials. It is possible that Jescheniak and Levelt did not have sufficient power to detect the interaction between frequency and presentation during picture naming, given that they recruited only 12 participants. However, other studies have replicated the stable frequency effect using the same materials (e.g., Levelt et al., 1998; Meyer, Sleiderink, and Levelt, 1998). Thus, the word frequency effect observed in earlier studies was stable across repetition of the materials.

The most obvious difference between our study and Jescheniak and Levelt's is that we tested participants online rather than in the lab. Research from different groups has shown that web-based experiments can be used to measure naming and response latencies with high accuracy (e.g., Fairs & Strijkers, 2021). Similar results have been observed in our own group, using Frinex (He, Meyer, Creemers, & Brehm, 2021; Hintz, Dijkhuis, van 't Hoff, McQueen, and Meyer, 2020). Participants in Experiment 2 were slower to respond than participants in Jescheniak and Levelt's study, and this difference was more pronounced in picture naming (139 ms difference in average naming latencies in the two studies) than in the gender decision (77 ms difference) and object recognition tasks (70 ms difference). This difference in speed is consistent with earlier studies (e.g., Fairs & Strijkers, 2021). Importantly, however, this difference cannot explain the discrepancy between Jescheniak and Levelt's and our results. For the gender decision task, we replicated the main effect of word frequency (33 ms in Jescheniak and Levelt's study and 27 ms in our Experiment 1c), the main effect of presentation, and their interaction. For the naming task, we replicated the frequency effect on the first presentation of the materials and the presentation effect (60 ms from the first to third presentation in Jescheniak and Levelt's study and 53 ms in our study). These findings are important because they rule out the possibility that our experimental set-up was not sensitive enough to capture small effects or that participants were inattentive (see also Hintz, et al., 2020).

Having ruled out differences in the materials and procedure, we are left with the conclusion that participants in the two studies differed in how they carried out the naming task, such that word frequency and presentation yielded additive effects in one sample (in the lab) but an interaction in the other (online). There is a large literature on repetition priming in word production and comprehension, which shows that harder items (such as words with lower frequency) tend to benefit from repetition more than easier ones (such as words with higher frequency; e.g., Griffin & Bock, 1998; Forster & Davis, 1984), and our results are consistent with this literature. This finding can be explained in different ways. For example, LF words might benefit more from repetition than HF words because they are further away from a pre-defined production threshold, and any changes in activation (as a result of repetition) depend on the distance from this threshold (e.g., Levelt et al., 1999). Alternatively, weak links between units may be strengthened by repetition more than stronger links (e.g., Seidenberg & McClelland, 1989). Finally, our experiments may have tapped into episodic memory. Participants may have created a short-term episodic memory trace of the picture (and its name) the first time it was presented, which indicates that it has been recently used and makes it easier to retrieve again (e.g., Forster & Davis, 1984). Participants can use this episodic trace when they have to name or indicate the gender of the picture a second or a third time. Accessing this episodic trace does not involve accessing the word's lexical information, and so LF words are just as easy to retrieve as HF words, leading to an interaction between word frequency and repetition. But regardless of the exact mechanisms underlying our results, it is still unclear why our findings differ from Jescheniak and Levelt's. Further research is needed to investigated *when* and *why* the word frequency effect may be stable or unstable.

Our findings have important consequences for research using word frequency as a tool to test theories of lexical access. Jescheniak and Levelt's (1994) paper is a classic paper in psycholinguistics, and it has been cited over 1300 times, often by research using word frequency as an index of phonological encoding (e.g., Ferreira & Pashler, 2002; Graves, Grabowski, Mehta, & Gordon, 2007; Griffin & Bock, 1998). For example, Graves, et al. (2007) used fMRI to identify regions of the brain that were only related to word frequency, and that were thus likely to reflect word form selection. However, our findings do not convincingly support Jescheniak and Levelt's claim that frequency only affects word form selection, and so frequency should not be considered a pure index of phonological encoding.

In sum, we conducted a conceptual (Experiment 1) and close (Experiment 2) replication of part of Jescheniak and Levelt's (1994) influential psycholinguistic study, which investigated the locus of the word frequency effect during speech production. Participants were faster to name (Experiments 1b and 2b) and determine the grammatical gender (Experiments 1c and 2c) of pictures with HF rather than LF names. In both cases, this frequency effect was larger on the picture's

first presentation than on its second or third. This pattern contrasts with Jescheniak and Levelt, who found a stable frequency effect in picture naming but not in gender decision. The cross-task similarity of our results suggests that the tasks involve the same representations, and does not support Jescheniak and Levelt's conclusion that word frequency affects word form access but not lemma access.

Declaration of competing interest

The authors declare no conflict of interest.

Data availability

Raw data and analysis scripts can be accessed at https://osf.io/tw8hs/.

Acknowledgements

We thank Bilge Guney and Elif Yildiz for collecting data for Experiment 2 and assisting with annotations. I would like to acknowledge funding from the EPS: Experiment 2 was supported by a Small Grant from the Experimental Psychology Society, awarded to RC.

Table A1
Experimental and filler items used in Experiment 1. In Experiment 1a (object recognition), the word preceding the picture was always the picture's name for the experimental items. For the filler items, the word was different from the picture's name and is shown in a separate column.

Experimental items	Filler items	
Picture/word	Picture	Word
aap	nietjes	hondenhok
aardbei	beer	schort
stoel	bankje	hoed
stift	horloge	bril
arm	mes	doos
ananas	puzzelstukje	varken
tent	varken	schilderij
tulp	kruis	lamp
augurk	zandkasteel	nijlpaard
toren	nijlpaard	kruis
tomaat	masker	vingerhoedje
bank	kaarsje	springtow
baksteen	douchekop	zandekastel
trompet	kroon	bord
brug	koffiebonen	dromenvanger
bloemkool	hek	voetbal
boot	vlag	spinnenweb
bramen	trap	pet
pen	prikbord	nietmachine
pompoen	kasteel	kussen
broek	nietmachine	tennisracket
kluis	vogel	skibril
kersen	pet	rietje
pijp	vogelbekdier	brood
pleister	tennisracket	hek
computer	oor	douchekop
champignon	stokbrood	vogelbekdier
deuren	hoed	mes
deurklink	vingerhoedje	horloge
eikel	stopcontact	olifant
helikopter	usb-stick	nietjes
haarband	schort	Riem
koe	spiegel	maan
kapstok	rietje	masker
kip	wolf	fotolijstje
kiwi	voetbal	bladblazer
klok	bord	oog
klomp	oog	beer
motor	cadeau	veiligheidsspeld
mier	nagelknipper	kaarsje
kruiwagen	hoefijzer	mandje

(*continued on next column*)

Table A1 (*continued*)

Experimental items	Filler items	
Picture/word	Picture	Word
kogel	lamp	kroon
leeuw	skibril	paard
libelle	bladblazer	oor
pot	zwaard	cadeau
pauw	springtouw	zwaard
piano	bril	giraffe
peer	spinnenweb	podium
pizza	stoplicht	usb-stick
pinguin	blik	stopcontact
roos	doos	startkabels
rozijnen	giraffe	blik
schelp	kussen	wolf
sleutel	lieveheersbeestje	kasteel
sigaret	mandje	vogel
sinaasappel	brood	prikbord
zeehond	olifant	bankje
zeester	veiligheidsspeld	nagelknipper
tank	startkabels	vlag
telefoon	paard	stokbrood
dolfijn	dromenvanger	hoefijzer
knikker	maan	spiegel
prei	schilderij	dartbord
veter	dartbord	koffiebonen
koelkast	hondenhok	trap
slang	fotolijstje	puzzelstukje
ketting	podium	stoplicht
schouder	riem	lieveheersbeestje

Table A2
Experimental and filler items used in Experiment 2. In Experiment 2a (object recognition), the word preceding the picture was always the picture's name for the experimental items. For the filler items, the word was different from the picture's name and is shown in a separate column.

Experimental items	Filler items	
Picture/word	Picture	Word
arm	anker	gras
auto	bad	ei
bank	baken	kruis
bezem	bed	slot
bijl	beeld	hooi
bloem	been	zwaard
boom	blad	mes
boot	bord	oog
brief	bot	net
broek	brood	huis
deur	dak	hert
fles	dorp	scheermes
fluit	ei	anker
hark	eiland	gewei
harp	fornuis	nest
hond	geweer	kampvuur
kam	gewei	wiel
kano	glas	oor
kerk	gordijn	servet
krab	graf	pak
mond	gras	schip
muur	harnas	bord
neus	hart	bed
pauw	hert	blad
peer	hooi	beeld
rups	huis	web
schaar	kampvuur	bot
schoen	kanon	been
slak	kasteel	geweer
slee	kompas	brood
snavel	kruis	bad
spin	kussen	zadel
step	masker	glas
ster	mes	dak
stoel	nest	baken

(*continued on next page*)

Table A2 (*continued*)

Experimental items	Filler items	
Picture/word	Picture	Word
tafel	net	harnas
tang	oog	hart
tol	oor	kussen
trap	orgel	eiland
uil	pak	dorp
vaas	scheermes	kompas
vinger	schip	gordijn
vis	servet	masker
voet	slot	kanon
worst	web	kasteel
zaag	wiel	fornuis
zak	zadel	graf
zwaan	zwaard	orgel

Table B1
By-item mean response times (ms) for experimental items in Experiment 1a.

Picture	Frequency	Response time
aap	high	769
aardbei	low	688
ananas	low	737
arm	high	811
augurk	low	662
baksteen	low	680
bank	high	732
bloemkool	low	755
boot	high	749
bramen	low	767
broek	high	713
brug	high	797
champignon	low	665
computer	high	716
dueren	high	852
deurklink	low	788
dolfijn	low	652
eikel	high	698
haarband	low	795
helikopter	high	724
kapstok	low	730
kersen	low	724
ketting	high	912
kip	high	666
kiwi	low	717
klok	high	722
klomp	low	697
kluis	high	693
knikker	low	836
koe	high	832
koelkast	high	749
kogel	high	742
kruiwagen	low	679
leeuw	high	711
libelle	low	702
mier	low	716
motor	high	722
pauw	low	765
peer	low	656
pen	high	772
piano	high	684
pijp	high	777
pinguin	low	718
pizza	high	790
pleister	low	809
pompoen	low	703
pot	high	892
prei	low	745
roos	high	709
rozijnen	low	683
schelp	low	777
schouder	high	810
sigaret	high	802

(*continued on next column*)

Table B1 (*continued*)

Picture	Frequency	Response time
sinaasappel	low	664
slang	high	816
sleutel	high	740
stift	low	838
stoel	high	760
tank	high	682
telefoon	high	839
tent	high	742
tomaat	low	673
toren	high	745
trompet	low	705
tulp	low	712
veter	low	891
zeehond	low	702
zeester	low	687

References

Baayen, H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*, 390–412.

Balota, D. A., & Chumbley, J. I. (1985). The locus of word-frequency effects in the pronunciation task: Lexical access and/or production? *Journal of Memory and Language, 24*, 89–106.

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

Bates, D. M., Maechler, M., Bolker, B., & Walker, S. (2021). Lme4: Linear mixed- effects models using "Eigein" and S4 (R package version 1.1-26). Retrieved from http://CRAN.R-project.org/package=lme4.

Brysbaert, M., Mandera, P., McCormick, S. F., & Keuleers, E. (2019). Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods, 51*, 467–479.

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). The impact of word prevalence on lexical decision times: Evidence from the Dutch Lexicon Project 2. *Journal of Experimental Psychology: Human Perception and Performance, 42*, 441–458.

Caramazza, A. (1997). How many levels of processing are there in lexical access? *Cognitive Neuropsychology, 14*, 177–208.

Caramazza, A., Bi, Y., Costa, A., & Miozzo, M. (2004). What determines the speed of lexical access: Homophone or specific-word frequency? A reply to Jescheniak et al. (2003). *Journal of Experimental Psychology: Learning, Memory, and Cognition, 30*, 278–282.

Caramazza, A., Costa, A., Miozzo, M., & Bi, Y. (2001). The specific-word frequency effect: Implications for the representation of homophones in speech production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*, 1430–1450.

Caramazza, A., & Miozzo, M. (1998). More is not always better: A response to Roelofs, Meyer, and Levelt. *Cognition, 69*, 231–241.

Collina, S., Tabossi, P., & De Simone, F. (2013). Word production and the picture-word interference paradigm: The role of learning. *Journal of Psycholinguistic Research, 42*, 461–473.

Decuyper, C., Brysbaert, M., Brodeur, M. B., & Meyer, A. S. (2021). Bank of Standardized Stimuli (BOSS): Dutch names for 1400 photographs. *Journal of Cognition, 4*. https://doi.org/10.5334/joc.180

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review, 93*, 283–321.

Fairs, A., & Strijkers, K. (2021). Can we use the internet to study speech production? Yes we can! Evidence contrasting online versus laboratory naming latencies and errors. *PLoS One, 16*. https://doi.org/10.1371/journal.pone.0258908

Ferreira, V. S., & Pashler, H. (2002). Central bottleneck influences on the processing stages of word production. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*, 1187–1199.

Finocchiaro, C., & Caramazza, A. (2006). The production of pronominal clitics: Implications for theories of lexical access. *Language and Cognitive processes, 21*, 141–180.

Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*, 680–698.

Gauvin, H. S., Jonen, M. K., Choi, J., McMahon, K., & de Zubicaray, G. I. (2018). No lexical competition without priming: Evidence from the picture–word interference paradigm. *Quarterly Journal of Experimental Psychology, 71*, 2562–2570.

Graves, W. W., Grabowski, T. J., Mehta, S., & Gordon, J. K. (2007). A neural signature of phonological access: Distinguishing the effects of word frequency from familiarity and length in overt picture naming. *Journal of Cognitive Neuroscience, 19*, 617–631.

Griffin, Z. M., & Bock, K. (1998). Constraint, word frequency, and the relationship between lexical processing levels in spoken word production. *Journal of Memory and Language, 38*, 313–338.

He, J., Meyer, A. S., Creemers, A., & Brehm, L. (2021). Conducting language production research online: A web-based study of semantic context and name agreement effects in multi-word production. *Collabra: Psychology*, 7. https://doi.org/10.1525/collabra.29935

Hintz, F., Dijkhuis, M., van 't Hoff, V., McQueen, J. M., & Meyer, A. S. (2020). A behavioural dataset for studying individual differences in language skills. *Scientific Data, 7*. https://doi.org/10.1038/s415907-020-00758

Huettig, F., Audring, J., & Jackendoff, R. (2022). A parallel architecture perspective on pre-activation and prediction in language processing. *Cognition, 224*. https://doi.org/10.1016/j.cognition.2022.105050

Indefrey, P. (2011). The spatial and temporal signatures of word production components: A critical update. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00255

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition, 92*, 101–144.

Iwasaki, N., Vigliocco, G., & Garrett, M. F. (1998). Adjectives and adjectival nouns in Japanese: Psychological processes in sentence production. In D. J. Silva (Ed.), *Vol. 8. Japanese/Korean Linguistics* (pp. 93–106). Stanford, CA: Center for the Study of Language and Information.

Jescheniak, J. D., & Levelt, W. J. (1994). Word frequency effects in speech production: Retrieval of syntactic information and of phonological form. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 824–843.

Jescheniak, J. D., Meyer, A. S., & Levelt, W. J. (2003). Specific-word frequency is not all that counts in speech production: comments on Caramazza, Costa, et al. (2001) and new experimental data. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 29*, 432–438.

Kemmerer, D. (2019). From blueprints to brain maps: The status of the lemma model in cognitive neuroscience. *Language, Cognition and Neuroscience, 34*, 1085–1116.

Kittredge, A. K., Dell, G. S., Verkuilen, J., & Schwartz, M. F. (2008). Where is the effect of frequency in word production? Insights from aphasic picture-naming errors. *Cognitive Neuropsychology, 25*, 463–492.

Knobel, M., Finkbeiner, M., & Caramazza, A. (2008). The many places of frequency: Evidence for a novel locus of the lexical frequency effect in word production. *Cognitive Neuropsychology, 25*, 256–286.

Kroll, J. F., & Potter, M. C. (1984). Recognizing words, pictures, and concepts: A comparison of lexical, object, and reality decisions. *Journal of Verbal Learning and Verbal Behavior, 23*, 39–66.

La Heij, W., Mak, P., Sander, J., & Willeboordse, E. (1998). The gender-congruency effect in picture-word tasks. *Psychological Research, 61*, 209–219.

La Heij, W., Puerta-Melguizo, M. C., van Oostrum, M., & Starreveld, P. A. (1999). Picture naming: Identical priming and word frequency interact. *Acta Psychologica, 102*, 77–95.

Levelt, W. J., Praamstra, P., Meyer, A. S., Helenius, P., & Salmelin, R. (1998). An MEG study of picture naming. *Journal of Cognitive Neuroscience, 10*, 553–567.

Levelt, W. J., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences, 22*, 1–38.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, Massachusetts: MIT Press.

Llorens, A., Trébuchon, A., Riès, S., Liégeois-Chauvel, C., & Alario, F. X. (2014). How familiarization and repetition modulate the picture naming network. *Brain and Language, 133*, 47–58.

Meyer, A. S., Sleiderink, A. M., & Levelt, W. J. (1998). Viewing and naming objects: Eye movements during noun phrase production. *Cognition, 66*, B25–B33.

Monsell, S., Matthews, G. H., & Miller, D. C. (1992). Repetition of lexicalization across languages: A further test of the locus of priming. *The Quarterly Journal of Experimental Psychology Section A, 44*, 763–783.

Morrison, C. M., Ellis, A. W., & Quinlan, P. T. (1992). Age of acquisition, not word frequency, affects object naming, not object recognition. *Memory & Cognition, 20*, 705–714.

Navarrete, E., Basagni, B., Alario, F. X., & Costa, A. (2006). Does word frequency affect lexical selection in speech production? *Quarterly Journal of Experimental Psychology, 59*, 1681–1690.

Nickels, L., Biedermann, B., Fieder, N., & Schiller, N. O. (2015). The lexical-syntactic representation of number. *Language, Cognition and Neuroscience, 30*, 287–304.

Nozari, N., Kittredge, A. K., Dell, G. S., & Schwartz, M. F. (2010). Naming and repetition in aphasia: Steps, routes, and frequency effects. *Journal of Memory and Language, 63*, 541–559.

Oldfield, R. C., & Wingfield, A. (1965). Response latencies in naming objects. *Quarterly Journal of Experimental Psychology, 17*, 273–281.

Paucke, M., Oppermann, F., Koch, I., & Jescheniak, J. D. (2015). On the costs of parallel processing in dual-task performance: The case of lexical processing in word production. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 1539–1552.

Rapp, B., & Goldrick, M. (2000). Discreteness and interactivity in spoken word production. *Psychological Review, 107*, 460–499.

Roelofs, A. (1992). A spreading-activation theory of lemma retrieval in speaking. *Cognition, 42*, 107–142.

Roelofs, A. (1997). The WEAVER model of word-form encoding in speech production. *Cognition, 64*, 249–284.

Sá-Leite, A. R., Comesaña, M., Acuña-Fariña, C., & Fraga, I. (2023). A cautionary note on the studies using the picture-word interference paradigm: The unwelcome consequences of the random use of "in/animates". *Frontiers in Psychology, 14*, 1145884.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Stark, K., van Scherpenberg, C., Obrig, H., & Abdel Rahman, R. (2023). Web-based language production experiments: Semantic interference assessment is robust for spoken and typed response modalities. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01768-2

Starreveld, P., & La Heij, W. (2004). Phonological facilitation of grammatical gender retrieval. *Language and Cognitive Processes, 19*, 677–711.

Strijkers, K., & Costa, A. (2011). Riding the lexical speedway: A critical review on the time course of lexical selection in speech production. *Frontiers in Psychology, 2*. https://doi.org/10.3389/fpsyg.2011.00356

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*, 39–58.

Van Assche, E., Duyck, W., & Gollan, T. H. (2016). Linking recognition and production: Cross-modal transfer effects between picture naming and lexical decision during first and second language processing in bilinguals. *Journal of Memory and Language, 89*, 37–54.

Van Heuven, W. J., 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*, 1176–1190.

Vigliocco, G., Antonini, T., & Garrett, M. F. (1997). Grammatical gender is on the tip of Italian tongues. *Psychological Science, 8*, 314–317.

Vigliocco, G., Vinson, D. P., Martin, R. C., & Garrett, M. F. (1999). Is "count" and "mass" information available when the noun is not? An investigation of tip of the tongue states and anomia. *Journal of Memory and Language, 40*, 534–558.

Vogt, A., Hauber, R., Kuhlen, A. K., & Abdel Rahman, R. (2022). Internet-based language production research with ert articulation: Proof of concept, challenges, and practical advice. *Behavior Research Methods, 54*, 1954–1975.

Wheeldon, L. R., & Monsell, S. (1992). The locus of repetition priming of spoken word production. *The Quarterly Journal of Experimental Psychology, 44*, 723–761.

Zwitserlood, P., Bölte, J., Hofmann, R., Meier, C. C., & Dobel, C. (2018). Seeing for speaking: Semantic and lexical information provided by briefly presented, naturalistic action scenes. *PLoS One, 13*. https://doi.org/10.1371/journal.pone.0194762