Memory & Cognition
https://doi.org/10.3758/s13421-025-01768-z

What is the level of precision of phonological representations in working memory?

Marion Bouffier[1,2] · Robin Remouchamps[1] · Steve Majerus[1]

Accepted: 18 July 2025

Abstract
The precision with which verbal information is represented in working memory (WM) is a debated question. Some studies suggest that verbal WM precision is limited to abstract phonological levels of representation while other studies, using specifically designed paradigms, indicate that information can reach phonetic-level precision. The present study investigated at which linguistic level verbal WM operates by default, by probing memory for phonological versus phonetic information in a non-word WM paradigm. In three experiments, we presented non-word lists followed by a non-word probe, with negative probes differing from targets by a single phoneme. This phoneme was either a phonetic variant of the target (e.g., /t/ - /t*/), a phonologically close phoneme (e.g., /t/ – /d/) or a phonologically distant phoneme (e.g., /t/ – /v/). In the three experiments, we observed reliable rejection of negative probes differing by a phonologically distant phoneme, while rejection of negative probes differing by either a phonologically close phoneme or a phonetic variant was much less robust. This study shows that verbal WM preferentially involves phonological levels of representation, and with limited precision at this level.

Keywords Auditory-verbal working memory · Precision · Phonetics · Phonology

Introduction

While working memory (WM) capacity has been studied extensively in terms of the quantity of information that can be maintained (e.g., Cowan, 1995; Miller, 1956), the quality of WM representations as reflected by their precision has only been considered more recently, and this mostly in the visuo-spatial WM domain (e.g., Ma et al., 2014). The few studies that have investigated WM precision in the verbal domain have led to contradictory evidence, with some studies suggesting a high level of precision up to phonetic (subphonemic) levels of representation (Hepner & Nozari, 2019; Joseph et al., 2015). Other studies indicate a lower level of verbal precision limited to phonological levels of representation (Bouffier et al., 2022; Rhodes et al., 2019). The aim of the present study was to examine the base level of verbal WM precision by contrasting WM precision at the phonetic and phonological level for word-like stimuli. This was achieved by manipulating phonetic and phonological information at three levels:

(1) At the phonetic level via a subtle alteration within the initial phoneme, rendering this phoneme ambiguous,
(2) At the phonological level by manipulating the voiced versus voiceless parameter of the initial phoneme (e.g.,/p/vs./b/), and
(3) Again at the phonological level but this time with more distant contrasts (e.g.,/b/vs./k/).

Most models of WM consider its limitations in terms of the number of items that can be stored in WM (Cowan, 2001, 2010; Miller, 1956; Pashler, 1988). On the other hand, resource models, particularly in the visuo-spatial WM research field, suggest that resources are flexibly allocated to all presented stimuli with no strict limit upon the number of items that can be maintained (Bays et al. 2009; see, however, Mızrak & Oberauer, 2021, for a related proposal in the verbal WM field). Instead of being either recalled or forgotten, items will be stored with variable levels of quality or precision. These assumptions have been derived from studies in the visuo-spatial WM domain using continuous reproduction

✉ Marion Bouffier
marion.bouffier@uliege.be

[1] University of Liège, Liège, Belgium

[2] Neurobiology of Concepts Expression (NoCE) Research Unit, University of Geneva, Chemin des mines, 9, 1202 Geneva, Switzerland

Published online: 13 August 2025

paradigms (Wilken & Ma, 2004), in which participants visualize a memory sequence and are then invited to adjust a probe item to match it with a feature of a target item from the memory sequence (e.g., Bays et al., 2009; Gorgoraptis et al., 2011; Oberauer & Lin, 2017; Zokaei et al., 2011). The amount of deviation between the target item and the reproduced item allows to determine the degree of visual WM precision. Precision has been found to decrease monotonically with increasing set size, and to be highest for final list items (i.e., recency effects).

Few studies have addressed, directly or indirectly, WM precision in the auditory-verbal domain, and those that did have led to conflicting results. One line of studies used paradigms close to the visual WM reproduction paradigms mentioned above, but which are rather artificial as regards naturalistic processing of oral language stimuli. Joseph et al. (2015) investigated phonetic levels of WM precision by presenting lists of syllables composed of a vowel and the final consonant /d/. After the sequence, participants had to match a random probe vowel to a target vowel from the sequence by using a rotative dial reproducing vowels in a continuous manner. Hepner and Nozari (2019) used a similar procedure for the reproduction of consonants instead of vowels. Both studies showed similar principles to those for visual WM precision, with precision decreasing monotonically with increasing set size and higher phonetic precision for final list items. These findings suggest that phonetic information can be maintained in verbal WM at a reasonable level of precision. At the same time, Joseph et al. (2015) showed greater clustering of reproduced vowels around particular phonemes with increasing memory load, suggesting that while WM representations can be contained in a non-categorical manner in WM when memory load is low, they progressively become more categorical (i.e., phonological) as memory load increases. In a more indirect manner, a number of studies addressed the precision with which auditory-verbal information is coded by using implicit memory paradigms. Toscano and colleagues (2010) measured evoked potential responses in an oddball paradigm for sequences of repeating natural versus phonetically edited words. They observed perceptual N1, but also post-perceptual P3 components when a phonetically deviating stimulus occurred, indicating that the participants had implicitly memorized the preceding repeating phonetic information. Similar results have also been observed for so-called visual-world paradigms in which participants were invited to fixate different pictures (McMurray et al., 2009) or cartoon characters (Brown-Schmidt & Toscano, 2017) while hearing phonetically ambiguous (with edited voice onset time (VOT), or along "he-she" continua) target words that described the visual scene. Disambiguating information then occurred at the end of the sentence. The authors observed that subsequent fixation latencies to the correct target were proportional to the distance of the onset phoneme to the target endpoint. This effect lasted over several syllables (Connine et al., 1991; Falandays et al., 2020). In sum, these latter studies show that a phonetic level of information can be maintained implicitly in the context of speech perception paradigms.

On the other hand, Rhodes et al. (2019) found no evidence that verbal information is maintained at a fine-grained phonetic level even in implicit memory situations. Like some of the earlier mentioned studies, this study also measured evoked potentials in an oddball paradigm. Participants were presented with two series of /t/ sounds while watching a silent movie. The two series of sounds had two different mean VOT values, while the deviant sound was the phoneme /d/. The authors hypothesized that if fine-grained phonetic information was contained in memory traces, the amplitude of the surprise response (as measured with the mismatch negativity response) should vary as a function of the deviation between the mean VOT of the sound series and the deviant sound. However, this was not the case, suggesting that fine-grained phonetic information might not be contained in implicit memory.

Overall, the implications of these findings for the level of precision in verbal WM are difficult to determine as most of these studies have focused on implicit memory rather than explicit, verbal WM. On the other hand, the studies that have focused on verbal WM used rather artificial paradigms in which the participants' attentional focus was directly oriented towards phonetic levels of representations. Thus, the question of the spontaneous, natural level of verbal WM precision (i.e., phonetic vs. phonological) remains unanswered. One recent study (Bouffier et al., 2022) examined precision for more naturalistic verbal WM paradigms but focused exclusively on phonological levels of processing, by presenting to-be-memorized word lists followed by a probe word that differed either minimally or maximally at the phonological level from a target word in the memory list, with phoneme overlap varying between 25% (e.g., "Smog" vs. "Plug"), 50% (e.g., "Mask" vs. "Dusk"), and 75% (e.g., "Rest" vs. "Test"). Precision was measured with regard to the extent to which participants were able to discriminate between target words and negative probe words as a function of their level of proximity. The authors showed that rejection performance gradually decreased as the number of shared phonemes between the target word and the negative probe word increased.

The present study examined the phonetic versus phonological level of precision that characterizes the maintenance of verbal information in WM by "default" (i.e., when attention is not explicitly directed towards subphonemic levels of speech processing). In three experiments, we used standard probe recognition WM tasks (Bouffier et al., 2022) and stimuli as close as possible to natural speech stimuli, without explicitly directing the participants' focus to phonetic levels of processing. We then varied the phonetic

versus phonological proximity of the probes relative to the target items in order to determine at which level of precision speech stimuli are maintained in WM. All items were word-like non-words obeying native language phonotactics in order to make them sound as natural as possible. We used non-words rather than familiar words in order to restrict WM processing to sound-based levels of maintenance, allowing us to assess the maximal limits of spontaneous phonetic versus phonological levels of precision. All experiments used the same task design and examined the same research question, Experiment 2 being a replication study of Experiment 1 but with the additional control of hearing status of the participants, and Experiment 3 aiming to replicate the results of the two first experiments with a new set of stimuli.

Experiment 1

We used an auditory probe recognition task including memory lists of five non-words followed by either a positive probe or any of three types of negative probes. We manipulated the number of overlapping dimensions between the probe and the target items; indeed, earlier work has shown that consonants might not be maintained in memory as entities, but along different articulatory dimensions (Wickelgren, 1966). The phonetically different negative probes (Phonetic contrast probes) were created by temporally reducing parts of the phonetic signal of the initial consonant of the target stimulus, rendering the probes and targets phonetically different, based on a procedure used by Majerus and Lorent (2009). For the phonologically close negative probes (Close phonological contrast probes) we opposed phonemes of the same phonemic contrast, at the level of their voicing parameter (e.g.,/p/ vs./b/). Finally, for the phonologically most distant negative probes (Distant phonological contrast probes) used in this experiment, the probe differed from the target by a single phoneme involving at least two phoneme-family differences (e.g.,/b/vs./ʃ/) or a multi-feature single-family change (e.g.,/t/ vs./k/; Garnier et al., 2018). Note that these contrasts were more variable than in the two other conditions, the main goal of this condition being to present stimuli that varied along more dimensions than the voicing parameter manipulated in the two other conditions. Note that contrasts such as /t/ versus /k/ can still be considered closely related in other languages (Schweppe et al., 2011). If non-words are maintained mainly at a phonological level of precision, we expect to observe significantly lower rejection accuracy for the Close phonological contrast probes than for the Distant phonological contrast probes and close-to-chance-level performance for Phonetic contrast probes. If non-words are maintained also at a phonetic level of resolution, then we should observe overall above-chance rejection accuracies for all three probe conditions, with nevertheless a decreasing gradient of recognition performance from Distant phonological contrast to Close phonological contrast to Phonetic contrast conditions, as a function of the increase of phonetic differences between the three conditions. Finally, we administered a minimal pair discrimination task at the end of the experiment in order to ascertain that the phonetic differences between stimuli were perceivable under minimal WM demands.

Method

Participants

Thirty young adults from the university community (14 women, age: mean = 25.87 years, SD = 3.85 years; number of years of education: mean = 14.77, SD = 2.19) participated in the study (see *Data analysis* section for statistical justification of sample size). They were recruited via ads posted on social media or via word of mouth. They were all monolingual French speakers, had no history of drug or medication abuse, and did not suffer from neurological or psychiatric disorders, or learning disabilities. They gave their informed consent prior to participating in the study. The study was approved by the ethics committee of the University of Liège (project number: 2016/358).

Materials and procedure

Experimental task

Stimuli One hundred and nine non-words were created using the non-word generator from the lexical database Lexique (New et al., 2004). All non-words were composed of four phonemes, starting with a consonant. Distant phonological contrast, Close phonological contrast, and Phonetic contrast stimuli were created by drawing 28 stimuli (i.e., ten stimuli for Phonetic contrast trials and nine stimuli for Close phonological contrast and Distant phonological contrast trials, respectively), and creating a counterpart differing only by its initial phoneme. For Close phonological contrast and Phonetic contrast trials, this counterpart differed on its voicing parameter, and for Distant phonological contrast trials, the counterpart differed on multiple phonetic features and phonological categories. Two Phonetic contrast pairs, one Close phonological contrast pair, and one Distant phonological contrast pair served as practice pairs only. We also controlled for phonological neighborhood density, which is the number of real words differing from a given non-word by only one phoneme via addition, deletion, or substitution. To calculate the number of neighbors, we used the Levenshtein distance (Levenshtein, 1966), which indicates the distance between two character arrays including addition, substitution, or deletion. The Levenshtein distance was computed by comparing

the non-words from our set with the words from the Lexique database. We considered as phonological neighbors the number of words with a Levenshtein distance of 1. In our set, the number of neighbors ranged between 0 and 13 (mean = 2.73, SD = 2.35), indicating an overall small neighborhood density, with only one non-word having a neighborhood density greater than 10. For Close phonological contrast and Distant phonological contrast trials, target and probe non-words were matched for their number of phonological neighbors (Bayesian independent t-test: $BF_{01} = 2.97$). For targets, the mean number of phonological neighbors was 2.00 (SD = 2.03), while for probes, the mean number of phonological neighbors was 1.94 (SD = 2.14). Target and probe items were further matched for their number of syllables, the maximum number of syllables in a given non-word being 2. All non-words started with a consonant, and only this onset consonant was manipulated. The reason for manipulating only the onset consonant was to avoid perceptual interactions between the manipulated phonetic feature(s)/phoneme and its within-stimulus position, stimulus beginnings being perceptually most salient (Beckman, 1998, 2013). All non-words were legal with respect to French phonotactic rules. The full list of stimuli and their associated metrics are available in Table S9 in the Online Supplementary Material (OSM) section.

Creation of phonetically modified non-words We edited the waveform of the ten selected stimuli using Praat (Boersma, 2001), by locating the onset of the plosive or fricative signal of the initial voiceless consonant on the spectrograms and by cutting out an average of 71% of the duration of the signal (note that in the two retained stimuli starting with a plosive and followed by the semi-consonant/w/, part of the onset of the semi-consonant was also cut). The phonetically modified stimuli were submitted to ten participants who were not part of the experimental sample for a forced-choice phoneme-identification task in order to confirm that the stimuli were perceived as ambiguous (i.e., leading to inconsistent identification) and distinct from the initial phoneme, thereby increasing the likelihood that the initial voiceless consonant was now perceived as its voiced counterpart. We also created a discrimination task of minimal pairs consisting of the ambiguous stimulus and its voiced/voiceless counterpart. This task was administered to ten other participants not part of the experimental sample. This was an iterative process: stimuli that did not directly lead to the expected identification pattern were further edited until they were identified in an ambiguous manner while being perceived as distinct from their two counterparts. The list of these stimuli as well as the final amount of reduction of initial consonant signal is displayed in Table 1.

For all list conditions, the non-words were assigned to five non-word lists. Each list contained four monosyllabic and one disyllabic non-word. In order to minimize interference, we avoided the situation where a given probe item or its phonetic/phonological variant could occur in the three lists preceding the trial in which the probe was used, and a given non-word could not occur on two consecutive trials. Fifty percent of all trials were followed by a positive probe (48 trials). Sixteen of these trials contained a phonetically modified (Phonetic contrast) non-word as target-probe stimulus. Negative trials were made up of three types, also with 16 trials for each type (i.e., Phonetic contrast trials, Close phonological contrast trials, Distant phonological contrast trials). The number of trials in the Close phonological contrast and Distant phonological contrast conditions followed the number of Phonetic contrast trials.

A subset of 56 stimuli served as the critical target/probe pairs in the experimental task, including the trials with phonetically modified stimuli. The same stimuli were used for target-probe pairs in the positive condition and negative probes. Non-words appearing in target-probe pairs occurred three to six times. Eighty-nine stimuli were used as filler items and occurred three to four times as non-targets across the different lists (see Table S9 in the OSM). The serial position of the target item within the memory list was counterbalanced across trials.

Table 1 Natural and phonetically modified stimuli

Natural			
Voiced	Voiceless	Modified stimulus	Reduction of onset noise (in %)
vorve	**forve**		
/vɔʁv/	/fɔʁv/	/f*ɔʁv/	86.24
douac	**touac**		
/dwak/	/twak/	/t*wak/	59.29
jueppe	**chueppe**		
/ʒɥɛp/	/ʃɥɛp/	/ʃ*ɥɛp/	65.4
vadre	**fadre**		
/vadʁ/	/fadʁ/	/f*adʁ/	77.67
juncle	**chuncle**		
/ʒœ̃kl/	/ʃœ̃kl/	/ʃ*œ̃kl/	69.71
jolf	**cholf**		
/ʒɔlf/	/ʃɔlf/	/ʃ*ɔlf/	68.98
doiffe	**toiffe**		
/dwaf/	/twaf/	/t*waf/	77.14
domlant	**tomlant**		
/dɔ̃lɑ̃/	/tɔ̃lɑ̃/	/t*ɔ̃lɑ̃/	74.6
zisc	**cisc**		
/zisk/	/sisk/	/s*isk/	57.14
gesgue	**chesgue**		
/ʒɛzg/	/ʃɛzg/	/ʃ*ɛzg/	73.99

The stimuli/s*isk/and/ʃ*ɛzg/were practice trials

Procedure Participants were presented with a total of 96 auditory five non-word lists, all starting with a 3-2-1 countdown (600 ms) written in white on a black background. After the countdown, the screen remained blank during the auditory presentation of the sequence, with each non-word presented at a rate of one non-word per second. The sequence was followed by a 3-s retention interval; a retention interval of 3–4 s is commonly used in probe recognition tasks (Atkins & Reuter-Lorenz, 2008) and aims at discarding purely sensory memory for most recent items (Crowder & Morton, 1969), or "sensory-matching" processes (Monsell, 1978). After this delay, a question mark appeared, followed 500 ms later by the auditory presentation of the probe non-word. Participants had to determine whether the probe non-word had been in the list by pressing the S key on the computer keyboard for "Yes" and the L key for "No." The keys were respectively marked with an "O" for "Oui" ("Yes") and with an "N" for "Non" ("No"). Note that participants were also asked about the certainty of their response at the end of each trial but these responses were not further analyzed in the context of this study given that they were not associated with a specific hypothesis. Figure 1 provides an overview of the different trial types.

Participants were given the following instructions: "You are going to participate in a study aiming to investigate verbal memory. During this experiment, you are going to hear lists of non-words: non-words are sounds that do not exist, but that sound like words from the French language. Every list will be followed by a brief interval during which you will have to maintain the non-words in your head. After this interval, you will see a question mark, and then you will hear an isolated non-word. You will be invited to indicate as fast as possible whether this non-word was in the list. You can answer 'Yes' by pressing the key marked with an 'O,' and 'No' by pressing the key marked with an 'N.' Be careful, some isolated non-words will be very similar to non-words from the list, without being identical." There were nine

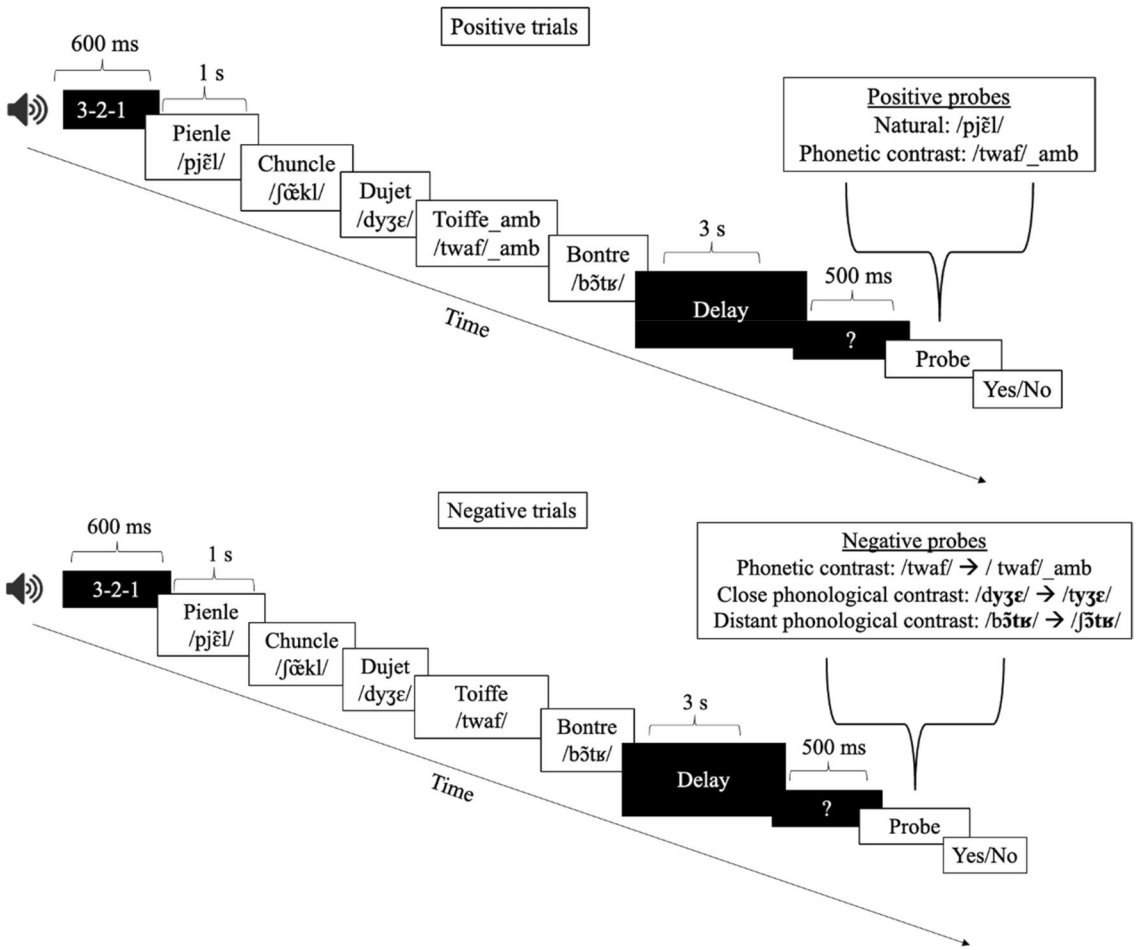

Fig. 1 Overview of the different trial types. Phonetic contrast non-words = Non-words with a phonetically modified initial phoneme; Close phonological contrast non-words = Non-words with a close initial phoneme; Distant phonological contrast non-words = Non-words with a distinct initial phoneme

practice trials prior to starting the task. The phonetically modified stimuli in these trials did not occur in the experimental task. Participants were given response feedback after each practice trial, but not in the experimental task. There was no limit in allowed response time. We created three fixed-list presentation orders. The stimuli were recorded by a native female French speaker, were converted to .wav files, and were presented via Grado SR60 headphones connected to the computer. The task was implemented using OpenSesame (Mathôt et al., 2012). We retained raw accuracy rates and mean reaction times (see OSM for the latter) for each condition. In order to maximize the reliability of data as a function of serial position, we pooled data for the first/second and fourth/fifth position. Hence, the number of trials for mid-list items was reduced as compared to the other positions, and caution is needed in the interpretation of position effects. We also separated the analysis of the positive condition (i.e., "yes" trials) from the analysis of the negative conditions (i.e., "no" trials), given that our aim was to focus on the effect of the negative trials on the ability to discriminate between two different stimuli (i.e., correct rejections). Note that the positive trials were divided into two conditions, namely the trials with phonetically modified probes and the trials with natural probes. We further conducted d' analyses (Macmillan & Creelman, 2005) in order to assess the impact of probe type on the participants' ability to distinguish between the target and the probe items. We calculated the hit rate in the positive condition minus the false alarm rate for each negative condition, which provided three d' categories: hits – false alarms for (1) Phonetic contrast trials, (2) Close phonological contrast trials, and (3) Distant phonological contrast trials.

Discrimination task

We administered a minimal pair discrimination task in order to check that the participants could perceptually discriminate between the different probe-target pairs of the experimental task when not needing to be maintained in WM. There were 20 identical pairs, ten Phonetic contrast non-word pairs, eight Close phonological contrast non-word pairs, and eight Distant phonological contrast non-word pairs, taken from the experimental task (including the two Phonetic contrast pairs from the practice trials). The Phonetic contrast non-word pairs occurred each twice (once with each counterpart) in the discrimination task in order to estimate discrimination ability for these highly close stimuli in a most reliable manner. The task was implemented in OpenSesame (Mathôt et al., 2012). The trials started with a white fixation cross displayed on a white background, followed after 500 ms by a non-word pair recorded at a rate of one stimulus per second. Immediately after the presentation of the pair, a question mark appeared, and participants had to determine as fast as possible whether the two stimuli were identical by pressing the S key for "Yes" or the L key for "No"; the S key was marked with an "O" for "Oui" ("Yes"), and the L key was marked with an "N" for "Non" ("No"). They were given the following instructions: "During this task, you are going to hear pairs of non-words. After each auditory presentation, you will be invited to indicate as fast as possible whether these non-words were identical. For some pairs, the differences will be very subtle. Hence, it is important to listen carefully to each non-word. If the non-words are identical, you have to respond 'Yes' by pressing the key marked with an 'O.' If the non-words are different, you have to respond 'No' by pressing the key marked with an 'N'."

The experimental and the discrimination tasks were administered in a single testing session lasting approximately 1 h. Participants were placed in a quiet room. The discrimination task was always administered after the WM precision task in order not to familiarize the participants with the stimuli before the WM task, which might have biased awareness towards the type of contrasts manipulated in the WM task (Moore et al., 2005). The tasks were administered at a comfortable listening level.

Data analysis

A Bayesian statistical approach was used in order to assess the impact of phonetic and phonological proximity on WM precision. Contrary to frequentist statistics, Bayesian statistics allow us to determine the strength of the evidence both against and in favor of the null hypothesis (Clark et al., 2018; Kruschke, 2011; Lee & Wagenmakers, 2013; Wagenmakers et al., 2018). The likelihood ratio of a model is given by the Bayes factor (BF); the best-fitting model is the one with the highest BF. BF_{01} indicates evidence in favor of the null hypothesis, while BF_{10} indicates evidence in favor of the alternative hypothesis. According to Jeffreys (1998), a $BF \geq 3$ is considered to indicate moderate evidence, a $BF \geq 10$ is considered to provide strong evidence, a $BF \geq 30$ is considered to provide very strong evidence, and a $BF \geq 100$ is considered to provide decisive evidence. The model with the highest BF is selected if it is at least three times more likely than the next-best model. If this is not the case, the most parsimonious model is selected. Bayesian repeated-measures ANOVAs were performed using the JASP statistical package with default prior settings (JASP Team, 2019, Version 0.11.1).

Statistical sensitivity (Bayesian equivalent of statistical power tests) of our experimental design was assessed with the BFDA package implemented in R (Schönbrodt & Stefan, 2018). The BFDA package provides simulations of specific statistical tests allowing us to determine the sample sizes and associated BF values as a function of a priori defined effect sizes. When planning the present study, we used an indicative effect size of Cohen's $d = 0.6$ based on a previous

study investigating the impact of phonological distractors in probe recognition tasks (Hamilton & Martin, 2007); note that a similar effect size was also observed by Bouffier et al. (2022). This analysis showed that for a paired t-test for pairwise comparisons of conditions-of-interest, the minimal sample size needed for reaching a minimal level of evidence ($BF_{10} > 3$) in favor of the effect in 90% of simulated samples was N = 30.

The main focus of the analysis was the raw recognition and rejection accuracies. These measures were analyzed by further taking serial position into account, allowing for a more fine-grained analysis of WM precision as a function of the position of the target non-word in the list. In a second step, d' analyses were conducted in order to complement the accuracy analyses, by providing information about memory discrimination performance. Reaction time analyses, although not the main focus of this study, were also conducted and can be found in the OSM section.

Results and discussion

Experimental task

Accuracy and d' values

We first assessed response accuracy in terms of proportions of correct responses (i.e., hits in the positive conditions and rejection accuracy in the negative conditions) as a function of probe type and target serial position (see Fig. 2). Descriptive statistics can be found in the OSM section.

A Bayesian repeated ANOVA on rejection accuracy for negative trials showed that the model with the strongest evidence included the probe type factor (Phonetic contrast, Close phonological contrast, and Distant phonological contrast trials) and the interaction between probe type and serial position. This model was 8,995.30 times more likely than the next highest model including the probe type factor alone (see Table 2a). Since there was no evidence in favor of an effect of the serial position factor ($BF_{10} = 0.33$), this variable was added to the null model for allowing accurate estimation of the evidence for the model including the interaction with this variable. The probe type factor was explored using Bayesian paired t-tests showing an overall advantage for Distant phonological contrast relative to Close phonological contrast ($BF_{10} = 4.21 \times 10^{+6}$) and to Phonetic contrast ($BF_{10} = 856553.00$) trials; there was evidence for an *absence* of a difference for the Close phonological contrast and Phonetic contrast trials ($BF_{01} = 5.09$). When conducting Bayesian one-sample t-tests, we observed no conclusive evidence in favor of above-chance performance for Phonetic contrast trials ($BF_{10} = 0.45$) and Close phonological contrast trials ($BF_{10} = 0.76$), while accuracy for Distant phonological contrast trials was well above chance ($BF_{10} = 2.99 \times 10^{+12}$). In addition, the interaction between probe type and serial position was explored by running pairwise comparisons between the probe type conditions and serial position. When comparing the Close phonological contrast and Phonetic contrast trials as a function of serial position, we observed evidence for an absence of a difference for positions 1-2 ($BF_{01} = 4.46$), but this evidence was inconclusive at positions 3 and 4-5 ($BF_{10} = 1.11$ and $BF_{10} = 1.16$, respectively), with

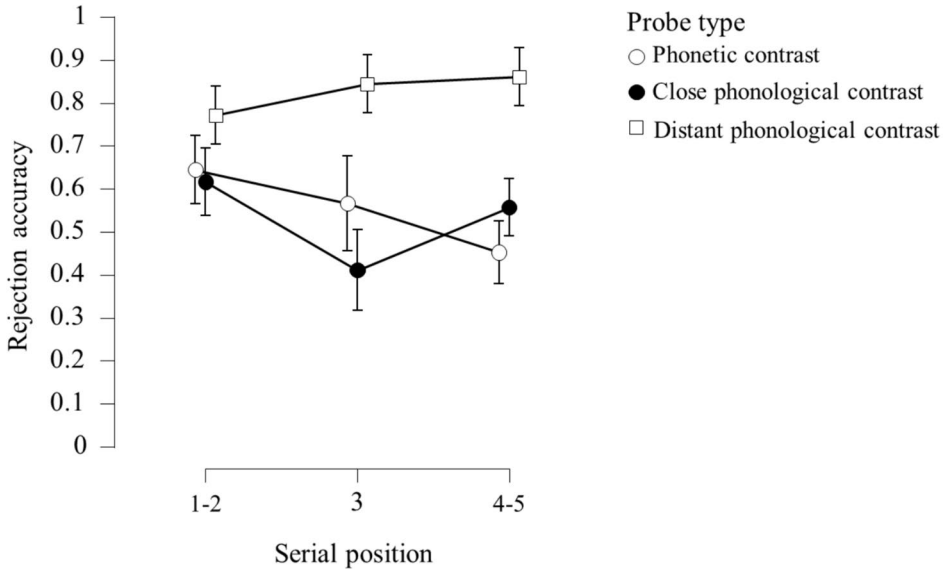

Fig. 2 Experiment 1 – rejection accuracy for negative probes as a function of probe type and target serial position. Error bars represent Bayesian credible intervals

Table 2 Bayesian factor values for rejection accuracy for negative probes as a function of probe type and target serial position

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Table 2a: **Experiment** 1					
Probe type + Probe type * Serial position	0.33	1.00	17993.49	$2.10\times10^{+12}$	1.77
Probe type	0.33	1.11×10^{-4}	2.22×10^{-4}	$2.34\times10^{+8}$	5.65
Table 2b:**Experiment** 2					
Probe type + Probe type * Serial position	0.17	0.74	13.88	$1.55\times10^{+19}$	1.69
Probe type + Hearing abilities + Probe type * Serial position	0.17	0.27	1.80	$5.59\times10^{+18}$	24.15
Probe type	0.17	2.09×10^{-9}	1.04×10^{-8}	$4.41\times10^{+10}$	2.11
Probe type + Hearing abilities	0.17	5.02×10^{-10}	2.51×10^{-9}	$1.06\times10^{+10}$	4.25
Hearing abilities	0.17	1.12×10^{-20}	5.62×10^{-20}	0.24	2.33
Table 2c: **Experiment** 3					
Probe type + Probe type * Serial position	0.33	1.000	$7.61\times10^{+15}$	$2.96\times10^{+25}$	2.47
Probe type	0.33	2.63×10^{-16}	5.26×10^{-16}	$7.79\times10^{+9}$	1.40

All models include subject, and random slopes for all repeated-measures factors. In all Experiments, the serial position factor is added to the null model

Phonetic contrast trials showing a reversed recency effect relative to the two other conditions. Overall, the results reveal very poor accuracy for rejecting non-words in Close phonological contrast and Phonetic contrast trials. This was further examined by running Bayesian one-sample t-tests comparing condition- and position-specific performance to a distribution of chance-level performance. For Phonetic contrast trials, accuracy was at or close to chance level at positions 3 and 4-5 ($BF_{01} = 2.82$ and $BF_{01} = 2.43$, respectively), but not for positions 1-2 ($BF_{10} = 7.22$). The same results were observed for Close phonological contrast trials, performance being close to chance level at positions 3 and 4-5 (Bayesian one sample t-test: $BF_{01} = 1.08$; 1.85, respectively), but not at positions 1-2 ($BF_{10} = 4.95$). Finally, on Distant phonological contrast trials, performance was well above chance level at all positions ($BF_{10} = 1.46 \times 10^{+7}$ for positions 1-2, $BF_{10} = 1.63 \times 10^{+7}$ for position 3, and $BF_{10} = 4.30 \times 10^{+9}$ for positions 4-5).

We also assessed recognition accuracy for positive trials as a function of the phonetically modified or natural character of the target and its serial position. We observed that the model associated with the strongest evidence included trial type (Phonetic contrast vs. natural) and serial position. This model was 8.07 times more likely than the model including also their interaction (see Table 3a). As shown in Fig. 3, Phonetic contrast trials were generally associated with lower accuracy. However, both trial types were associated with above-chance performance ($BF_{10} = 3.32 \times 10^{+12}$ for natural trials and $BF_{10} = 935.23$ for Phonetic contrast trials). On Phonetic contrast trials, performance reached chance level at positions 1-2 (Bayesian one-sample t-test: $BF_{01} = 3.03$), but approached or reached above-chance-level performance at positions 3 and 4-5 ($BF_{10} = 2.59$ and $BF_{10} = 455450.27$, respectively). For natural trials, performance was above chance level at all positions ($BF_{10} = 70871.72$ for positions 1-2, $BF_{10} = 2.50\times10^{+7}$ for position 3 and $BF_{10} = 2.25\times10^{+16}$ for positions 4-5). The lower recognition rates for Phonetic contrast trials suggest that participants maintained the phonetically modified target stimulus as a phonologically categorized stimulus, which, due to its ambiguous nature, led to inconsistent categorizations when hearing the stimulus at encoding and recognition stages.

Next, we analyzed hits and false alarms by combining them via d' scores, as a function of probe type using a Bayesian repeated measure ANOVA. Serial positions were not taken into account given the small number of trials for each position (Macmillan & Creelman, 2005). As shown in Table 4a, we observed decisive evidence in favor of an effect of probe type. Mean d' values were overall low, with the highest d' values observed for Distant phonological contrast trials (d' = 1.76) and d' values for the two other probe types approaching but not reaching 0 (0.80 for Close phonological contrast trials, and 0.85 for Phonetic contrast trials) (see Fig. 4). Furthermore, there was evidence in favor of an absence of difference between Close phonological contrast and Phonetic contrast trials, mirroring the preceding analyses for negative probe rejection accuracy (Bayesian paired-samples t-test: $BF_{01} = 4.87$).

Discrimination task

Finally, we analyzed accuracy in the discrimination task as a function of stimulus pair type (i.e., identical non-words, Distant phonological contrast, Close phonological contrast, Phonetic contrast). We found decisive evidence in favor of an effect of stimulus type ($BF_{10} = 1.42\times10^{+10}$); accuracy

Table 3 Bayesian factor values for recognition accuracy for positive probes as a function of trial type and target serial position

Models	P(M)	P(Mldata)	BF_M	BF_{10}	error %
Table 3a:**Experiment** 1					
Trial type + Serial position	0.20	0.89	31.63	$6.61\times10^{+6}$	2.74
Trial type + Serial position + Trial type * Serial position	0.20	0.11	0.50	818597.67	8.91
Serial position	0.20	0.00	0.01	16776.63	2.93
Trial type	0.20	4.72×10^{-5}	1.89×10^{-4}	351.50	2.76
Table 3b: **Experiment** 2					
Serial position + Serial position * Trial type	0.17	0.66	9.86	$2.93\times10^{+10}$	2.28
Serial position + Hearing abilities + Serial position * Trial type	0.17	0.31	2.21	$1.35\times10^{+10}$	3.50
Serial position	0.17	0.02	0.11	$9.50\times10^{+8}$	3.28
Serial position + Hearing abilities	0.17	0.01	0.04	$3.90\times10^{+8}$	2.89
Hearing abilities	0.17	9.32×10^{-12}	4.66×10^{-11}	0.41	2.10
Table 3c: **Experiment** 3					
Serial position	0.33	0.86	11.89	$5.26\times10^{+8}$	4.43
Serial position + Serial position * Trial type	0.33	0.14	0.34	$8.86\times10^{+7}$	3.47

All models include subject, and random slopes for all repeated measures factors. In Experiments 2 and 3, the trial type factor is added to the null model

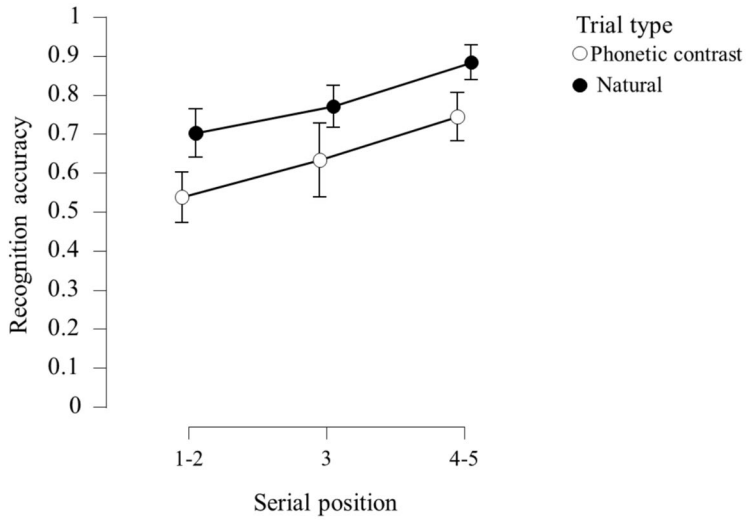

Fig. 3 Experiment 1 – Recognition accuracy for positive probes as a function of trial type and target serial position. Error bars represent Bayesian credible intervals

rates were 0.98 for identical stimulus pairs, 0.99 for Distant phonological contrast as well as for Close phonological contrast negative trials, and 0.87 for Phonetic contrast negative trials. Critically, although accuracy was lower for detecting deviations in the Phonetic contrast pairs as compared to the other negative pairs, participants detected the phonetic deviations of the Phonetic contrast pairs in a very reliable manner, in contrast to the WM task where detection performance was at or close to chance level.

In sum, the results of Experiment 1 indicate that probes differing minimally at the phonological or phonetic level from target items are not reliably rejected in a WM task, while perceptually, participants are able to detect both phonetic and minimal phonological deviations. Rejection accuracy for Phonetic contrast trials seems to vary according to serial position, with above chance performance only for initial items, and no overall evidence in favor of above-chance accuracy. Similar results were observed for Close phonological contrast trials, which showed close to chance performance for mid-list and end-of-list items. Also, in positive trials, phonetically modified targets led to reduced recognition accuracy, indicating that they were not systematically maintained

Table 4 Bayesian factor values for d' scores as a function of probe type

Models	P(M)	P(M|data)	BF_M	BF_{10}	error %
Table 4a: **Experiment** 1					
Probe type	0.50	1.000	$6.73\times10^{+6}$	$6.73\times10^{+6}$	1.02
Table 4b: **Experiment** 2					
Probe type	0.25	0.59	4.30	$4.29\times10^{+9}$	0.89
Probe type + Hearing abilities	0.25	0.41	2.10	$2.99\times10^{+9}$	1.61
Hearing abilities	0.25	8.07×10^{-11}	2.42×10^{-10}	0.59	1.77
Table 4c: **Experiment** 3					
Probe type	0.50	1.00	$4.63\times10^{+13}$	$4.63\times10^{+13}$	1.40

All models include subject and random slopes for all repeated-measures factors

in a veridical manner but in a phonologically recoded manner, given that, by design, the phonetic modification implied ambiguous features that render phonological categorization uncertain. At the same time, performance was overall above chance for Phonetic contrast trials, indicating that recognition might be easier than rejection and that the precision of representations might hence be sufficient to more reliably recognize positive probes (Dobbins et al., 2000; Roediger & McDermott, 1995). The results of Experiment 1 therefore do not provide evidence for a spontaneous use of phonetic levels of representation in verbal WM. One may, however, argue that this relatively low response accuracy could at least partly be due to hearing loss in some participants, which may have interacted with the perception and memorization of phonetically modified stimuli by favoring processing at a top-down, abstract phonological level. Although this possibility is not in line with the participants' good abilities for detecting phonetic deviations in the minimal pair speech perception task, it is nevertheless important to empirically verify this possibility. Indeed, previous studies have shown that even young healthy adult participants can show a relatively large range of auditory abilities, including increased hearing thresholds similar to those observed with age-related hearing decline (Verhaegen et al., 2014).

Experiment 2

Experiment 2 aimed at replicating the results of Experiment 1, by additionally including pure tone audiometric assessment of the participants in order to control for the possible impact of diminished hearing abilities on phonetic-level versus phonological-level retention in verbal WM. Furthermore, in order to allow for optimal statistical sensitivity regarding evidence in favor of the null effect, we recruited a larger sample (Brysbaert, 2019).

Method

Participants

Fifty (25 women) monolingual French speakers from the university community, with no history of medication or drug abuse, psychiatric or neurological disorders, or learning disabilities were recruited via ads posted on social media or via word of mouth. Note that these participants were part of a larger study also aiming at assessing word-level phonological

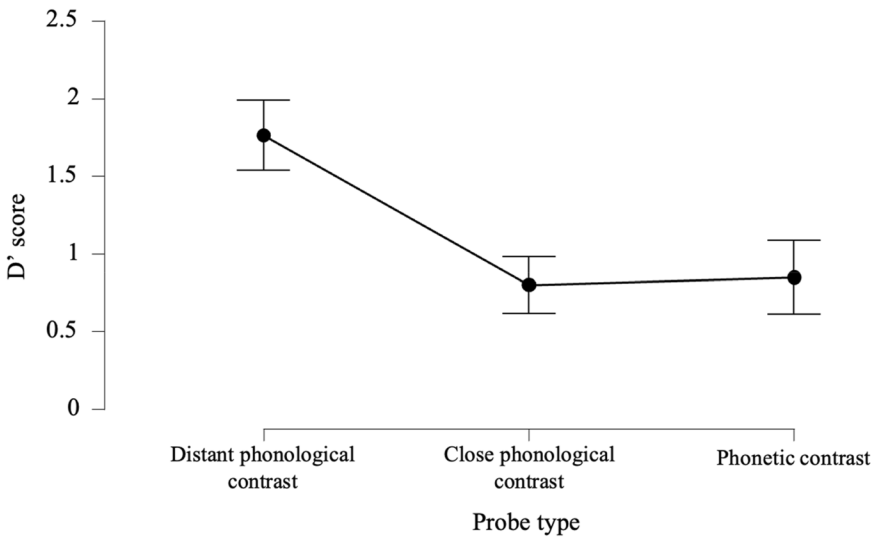

Fig. 4 Experiment 1 – d' scores as a function of probe type. Error bars represent Bayesian credible intervals

and semantic WM precision (Bouffier & Majerus, in preparation). Three tasks assessing phonetic-phonological, word-level phonological, and semantic levels of representation in WM were administered to the participants, in three separate experimental sessions. The present experiment thus focused on a WM task (phonetic-phonological) administered in isolation in a single session. This larger sample size allowed for assessing maximal sensitivity in favor of the null hypothesis. Bayesian sensitivity analysis (indicative effect size of Cohen's d = 0.6) showed that a sample size of 35 participants was needed to reach a minimal level of evidence in favor of the null in 90% of the cases. However, given that we might need to exclude some participants based on their hearing status, we recruited a larger sample.

In addition to the inclusion criteria mentioned in Experiment 1, hearing abilities were assessed using a screening pure tone audiometric test. We used the audiometric screening procedure implemented by the Madsen™ Xeta™ audiometer system. This screening test was always administered after the WM and minimal pair discrimination tasks. Frequencies of, respectively, 1,000 Hz, 2,000 Hz, 4,000 Hz, 8,000 Hz, 500 Hz, and 250 Hz, at 40 dB or 15 dB amplitude levels, were presented alternatively in the right and left ear. Participants were required to press a button when they heard the target frequency. If they failed to respond to a given stimulus, the next trial was automatically initiated. Although all participants accurately responded to all frequencies at an intensity of 40 dB, we excluded five participants who presented potential bilateral hearing loss at 15 dB for at least one frequency below 4,000 Hz (within the spectrum of frequencies that define speech sounds). One further participant was discarded for having already participated in a similar study. The final sample included 44 participants (22 women) (age: mean = 23.34 years, SD = 2.86 years; number of years of education: mean = 15.25, SD = 1.74). The study was approved by the ethics committee of the University of Liège (project number: 2016/358).

Materials and procedure

Except for audiometric screening, all other materials and task procedures were identical to Experiment 1. The tasks were presented in the following order: WM task, discrimination task, and audiometric screening.

Data analysis

As in Experiment 1, data were analyzed using a Bayesian framework implemented in JASP (JASP Team, 2019, Version 0.11.1). We conducted repeated-measures ANOVAs with hearing abilities as a covariate, measured as the number of correctly identified frequencies.

Results and discussion

Experimental task

Accuracy and d' values

We assessed response accuracy in terms of proportions of correct responses (i.e., hits in the positive condition and rejection accuracies in the negative conditions) as a function of probe type and target serial position (see Fig. 5), while entering hearing ability as covariate. Descriptive statistics can be found in the OSM.

When assessing negative trials as a function of probe type and serial position, the model associated with the strongest evidence included both the probe type factor and the interaction between probe type and serial position. This model was 2.77 times more likely than the model including also hearing abilities (see Table 2b). Since there was no evidence in favor of an effect of the serial position factor ($BF_{10} = 0.29$), this variable was added to the null model. As in Experiment 1, paired t-tests showed an advantage for Distant phonological contrast compared to Close phonological contrast and Phonetic contrast trials ($BF_{10} = 185209.77$ and $BF_{10} = 1.43\times10^{+7}$, respectively). There was, however, also a difference between the two latter trial types, performance being higher for Close phonological contrast trials compared to Phonetic contrast trials ($BF_{10} = 25.16$). Bayesian one-sample t-tests again revealed inconclusive evidence for above chance performance for Phonetic contrast trials ($BF_{10} = 1.60$), but above-chance level performance for Close phonological contrast trials ($BF_{10} = 23179.16$) and Distant phonological contrast trials ($BF_{10} = 2.93 \times 10^{+20)}$. The interaction between probe type and serial position was again explored running a series of pairwise Bayesian paired t-tests. The difference between the Close phonological contrast and Phonetic contrast trials was mainly due to a difference between positions 4-5 ($BF_{10} = 794.55$). As in Experiment 1, there was no conclusive evidence for such a difference at positions 1-2 ($BF_{10} = 0.84$), and position 3 ($BF_{10} = 1.83$), with again a reverse recency effect for the Phonetic contrast trials. Finally, when conducting Bayesian one-sample t-tests comparing performance to a distribution of chance-level performance, we observed that for Phonetic contrast trials, performance was at or close to chance level at positions 3 and 4-5 (Bayesian one-sample t-test: $BF_{01} = 1.27$ and $BF_{01} = 5.31$, respectively) and was above chance level only at positions 1-2 ($BF_{10} = 13.02$). For Close phonological contrast trials, accuracy was at chance level only at position 3 (Bayesian one sample t-test: $BF_{01} = 4.61$), but clearly not at positions 1-2 and 4-5 ($BF_{10} = 12362.99$ and $BF_{10} = 662980.44$, respectively). Finally, for Distant phonological contrast trials, performance was above chance level at all positions ($BF_{10} = 1.50 \times 10^{+8}$ for positions 1-2, $BF_{10} =$

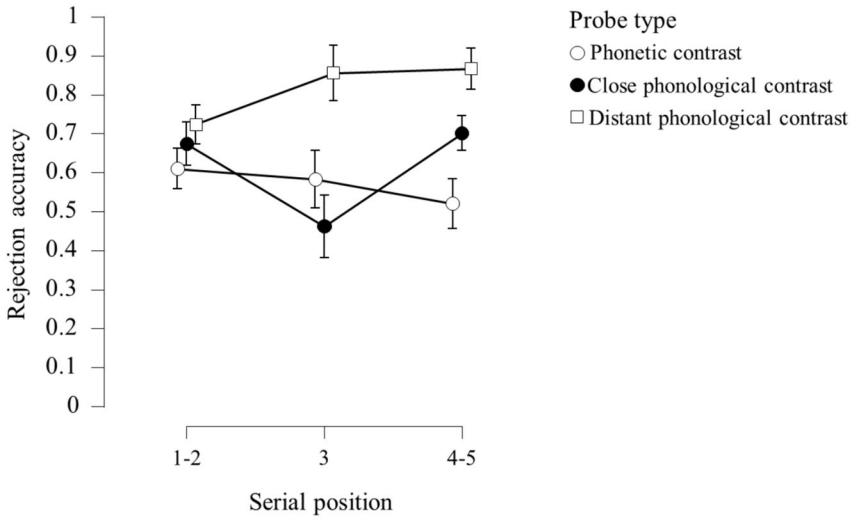

Fig. 5 Experiment 2 – rejection accuracy for negative probes as a function of probe type and target serial position. Error bars represent Bayesian credible intervals

$3.18 \times 10^{+12}$ for position 3, and $BF_{10} = 7.70 \times 10^{+18}$ for positions 4-5). In sum, the pattern of accuracy for negative probes was very similar to Experiment 1, with a strong decrease of performance for Close phonological contrast and Phonetic contrast trials relative to Distant phonological contrast trials. The only difference was a higher accuracy of Close phonological contrast trials for positions 4-5 relative to Phonetic contrast trials.

When assessing positive trials as a function of trial type and serial position, the model associated with the strongest evidence included serial position and the interaction between trial type and serial position. This model was 2.17 times more likely than the model also including hearing abilities and should be retained as being the most parsimonious model (see Table 3b). Since there was no evidence in favor of an effect of the trial type factor ($BF_{10} = 0.21$), this variable was added to the null model. A direct Bayesian paired-samples t-test on trial type showed that unlike Experiment 1, there was evidence in favor of an *absence* of a difference ($BF_{01} = 1.44$). As in Experiment 1, both trial types were above chance level ($BF_{10} = 1.49 \times 10^{+14}$ for natural trials and $BF_{10} = 4.04 \times 10^{+8}$ for Phonetic contrast trials). Exploration of the trial type by serial position interaction showed that there was an advantage for natural trials at positions 4-5 ($BF_{10} = 3.42$) as in Experiment 1, but not for the other positions ($BF_{01} = 1.82$ and $BF_{01} = 1.15$, for positions 1-2 and 3, respectively). Furthermore, as shown in Fig. 6, natural trials were associated with recency effects, while accuracy at positions 3 and 4-5 was comparable for Phonetic contrast trials. Finally, accuracy was above chance level for all positions not only for natural trials, but also for Phonetic contrast trials (natural trials: $BF_{10} = 1262.52$ for positions 1-2, $BF_{10} = 5.66 \times 10^{+6}$ for position 3, $BF_{10} = 5.08 \times 10^{+26}$ for positions 4-5; Phonetic contrast trials: $BF_{10} = 3.87$ for positions 1-2, $BF_{10} = 7.98 \times 10^{+8}$ for position 3, $BF_{10} = 9.18 \times 10^{+12}$ for positions 4-5). Hence, recognition accuracy for positive probes in Phonetic contrast trials was clearly higher than in Experiment 1.

We then conducted a Bayesian repeated-measures ANOVA to assess d' values as a function of probe type, while entering hearing abilities as a covariate. The model associated with the strongest evidence included the probe type factor only. This model was 1.41 times more likely than the less parsimonious model also including hearing abilities (see Table 4b). Mean d' values were 1.61 for Distant phonological contrast trials, 1.08 for Close phonological contrast trials, and 0.87 for Phonetic contrast trials (see Fig. 7). As in Experiment 1, discrimination accuracy was very low for Phonetic contrast trials. Unlike Experiment 1 however, discrimination performance was higher for Close phonological contrast trials than for Phonetic contrast trials ($BF_{10} = 10.61$). At the same time, discrimination accuracy trials remained decisively higher for Distant phonological contrast trials than for Close phonological contrast trials ($BF_{10} = 42503.63$).

Discrimination task

We conducted a repeated-measures ANOVA to assess response accuracy in the discrimination task as a function of pair type (i.e., identical non-words, Distant phonological contrast, Close phonological contrast, Phonetic contrast), while also entering hearing abilities as a covariate. The model associated with the strongest evidence included pair type alone ($BF_{10} = 2.34 \times 10^{12}$) and was 5.37 times more likely than the model including pair type and hearing abilities. Mean accuracy rate was 0.98 for identical trials, 0.98

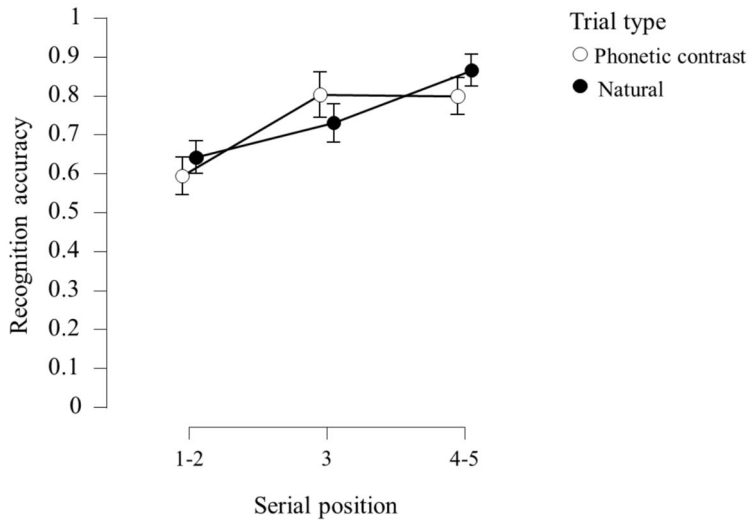

Fig. 6 Experiment 2 – recognition accuracy for positive probes as a function of trial type and target serial position. Error bars represent Bayesian credible intervals

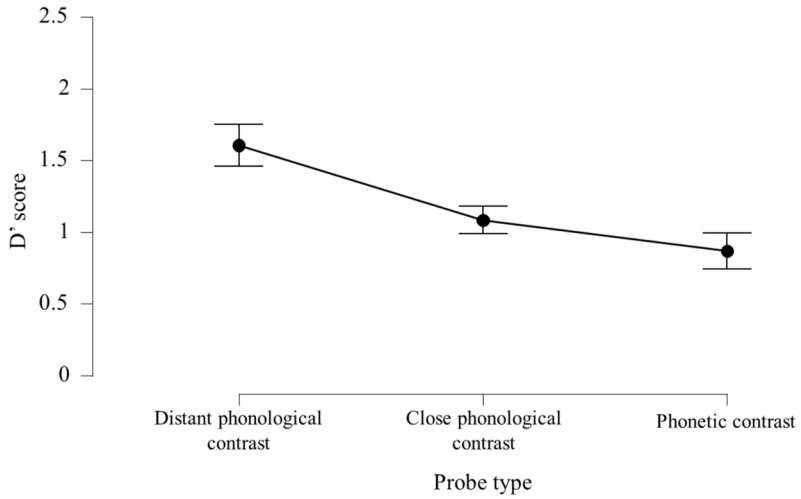

Fig. 7 Experiment 2 – d' scores as a function of probe type. Error bars represent Bayesian credible intervals

for Distant phonological contrast and Close phonological contrast trials, and 0.83 for Phonetic contrast trials. Although accuracy for phonetically modified non-words was reliably lower than accuracy for the other trials, it was well above chance level (Bayesian one-sample t-test: $BF_{10} = 8.73 \times 10^{13}$). These results are very similar to those observed in Experiment 1 and confirm that while WM accuracy is strongly impacted by phonetic/phonological proximity of target-probe pairs, participants are consistently able to perceive subtle phonetic differences in a discrimination task.

In sum, Experiment 2 yielded similar findings to Experiment 1, with overall similar probe type and/or serial position effects. Rejection accuracy was at chance-level on Phonetic contrast and Close phonological contrast negative trials for mid-list and/or end-of-list items. Similarly, d' values were impacted by phonological and phonetic proximity, with Phonetic contrast trials leading to the lowest results. However, we also noticed two differences between Experiments 1 and 2: the first difference is the higher rejection accuracy and d' scores for Close phonological contrast trials. Note that this difference was mainly due to higher performance for end-of-list trials. The second difference was the higher recognition accuracy in Experiment 2 for positive trials involving phonetically modified targets. These potential differences were explored with a between-experiment comparison aiming to assess their statistical robustness. When carrying out a mixed ANOVA

on rejection accuracy for negative trials, with Experiment as between-subject factor, we observed that the model with the strongest evidence did not include the main effect of Experiment or any interaction involving this factor: the best model included the probe type factor, the serial position factor and their interaction ($BF_{10} = 3.42 \times 10^{34}$), and this model was 3.62 times more likely than the next-best model also including the Experiment factor. Second, when conducting a between-experiment ANOVA on positive trials as a function of trial type and serial position, the model to-be-retained included the Experiment factor in addition to the trial type factor and the serial position factor, as well as the interaction between Trial type and Experiment ($BF_{10} = 2.13 \times 10^{17}$). We discuss potential reasons for this interaction in the General discussion section. It is moreover important to note that the number of stimuli was restricted in both experiments. Therefore, we conducted a third experiment with a new set of stimuli to examine the reproducibility and generalizability of the observed results (e.g., Neath et al., 2022). This additional experiment aimed to demonstrate that the results observed in Experiments 1 and 2 can be extended to a different stimulus set, as well as to clarify the discrepancies observed between Experiments 1 and 2 regarding recognition accuracy on positive trials. Moreover, Experiment 3 allowed us to provide a tighter control of the contrasts used for the Distant phonological condition.

Experiment 3

Experiment 3 aimed at reproducing the results of Experiments 1 and 2 with another set of stimuli. The type of contrasts manipulated for Phonetic contrast and Close phonological contrast trials focused on voicing parameters as in the preceding experiments, but on the basis of a different stimulus set. We additionally addressed the potential issue in Experiments 1 and 2 concerning the greater variability of Distant phonological contrast target-probe pairs, relative to the other target-probe pairs. In Experiment 3, we restricted negative Distant phonological contrast target-probe pairs to changes in voicing and place of articulation (e.g., contrasts /b/ - /k/ or /p/ - /g/) only. Stimuli involving nasal (i.e.,/m/,/n/) lateral (i.e,/l/), or vibrant (i.e.,/R/), consonant changes were not used anymore.

Method

Participants

Sixty-eight young adults from the university community (38 women, age: mean = 22.10 years, SD = 2.52 years; number of years of education: mean = 14.69, SD = 1.89) participated in the study. Sixty-four were recruited via ads posted on social media or via word of mouth. Four participants were furthermore tested via the participant pool of the University of Liège in exchange for course credit. Mixing the sources of recruitment has the advantage of reducing sampling bias. All participants were native French speakers, had no history of drug or medication abuse, and did not suffer from neurological or psychiatric disorders, or learning disabilities. They gave their informed consent prior to participating in the study.

With this larger sample size relative to Experiments 1 and 2, we aimed at obtaining higher sensitivity for evidence in favor of the null hypothesis, as our Bayesian sensitivity analysis showed that a sample size of at least 50 participants was needed to reach a minimal level of evidence in favor of the null in 95% of the cases.

As in Experiment 2, we assessed hearing abilities with a pure tone audiometric screening test implemented by the Madsen™ Xeta™ audiometer system. This screening test was again administered after the WM and minimal pair discrimination tasks. No participants were excluded based on this audiometric screening test. However, one participant was excluded for not being a monolingual French speaker (i.e., this participant spoke a second language fluently and on a daily basis). Six further participants were excluded due to experimenter error. The final sample included 61 participants (34 women) (age: mean = 22.26 years, SD = 2.57 years; number of years of education: mean = 14.72, SD = 1.87).

The study was approved by the ethics committee of the University of Liège (project number: 2223-051).

Materials and procedure

Experimental task

Stimuli In order to create a new stimulus set while keeping a similar task setup as for Experiments 1 and 2, we started by selecting the non-words in the previous experiments that had not been used as target-probe pairs (i.e., the filler non-words). We then created for these non-words new negative probe non-words, as a function of the three stimulus conditions (Distant phonological contrast, Close phonological contrast, Phonetic contrast).

All stimuli were composed of four phonemes and started with a consonant. As in the two previous experiments, we controlled for phonological neighborhood density using the Levenshtein distance. The number of phonological neighbors ranged from 0 to 12 (mean = 3.06, SD = 2.41), with only one stimulus with a neighborhood density above 10. Furthermore, Close phonological contrast and Distant phonological contrast trials were matched for their number of phonological neighbors (Bayesian independent t-test: $BF_{01} = 2.87$). For targets, mean number of phonological neighbors was 3.19 (SD = 3.41), while for probes, mean number of phonological neighbors was 3.50 (SD = 2.25). Table S10 in the OSM shows the new stimulus set and associated metrics.

More specifically for the creation of phonetically modified non-words, we edited the waveform of eight new stimuli using Audacity software (http://www.audacityteam.org). The phonetically modified stimuli were submitted to 23 participants not part of the experimental sample for a forced-choice phoneme identification task and a minimal pair discrimination task in order to confirm that the stimuli were perceived as ambiguous (i.e., leading to inconsistent identification) and distinct from the initial phoneme, thereby increasing the likelihood that the initial voiceless consonant was now perceived as its voiced counterpart. As in the previous experiments, this was an iterative process, and stimuli that did not immediately lead to the expected identification pattern were further edited. In the final stimuli sample, we cut out an average of 77% of the onset signal in voiceless consonants.

The list of these stimuli as well as the final amount of reduction of initial consonant signal is displayed in Table 5. Note that the two ambiguous stimuli used for practice trials used in the previous experiments were also kept for this experiment.

Table 5 Experiment 3 – natural and phonetically modified stimuli

Natural			
Voiced	Voiceless	Modified stimulus	Reduction of onset noise (in %)
jepre	**chepre**		
/ʒɛpʁ/	/ʃɛpʁ/	/ʃ*ɛpʁ/	79.84
glonque	**clonque**		
/glɔ̃k/	/klɔ̃k/	/k*lɔ̃k/	80.25
grour	**crour**		
/gʁuʁ/	/kʁuʁ/	/k*ʁuʁ/	65.05
vanvre	**fanvre**		
/vɑ̃vʁ/	/fɑ̃vʁ/	/f*ɑ̃vʁ/	87.14
bliffe	**pliffe**		
/blif/	/plif/	/p*lif/	71.95
guiande	**quiande**		
/gjɑ̃d/	/kjɑ̃d/	/k*jɑ̃d/	80.47
zufce	**sufce**		
/zyfs/	/syfs/	/s*yfs/	72.77
dilin	**tilin**		
/dilɛ̃/	/tilɛ̃/	/t*ilɛ̃/	82.00

The practice stimuli/s*isk/and/ʃ*ɛzg/used in Experiments 1 and 2 were reused in Experiment 3

Procedure Except for the modification of the stimuli, the procedure and order of administration of the tasks was identical to Experiment 2. As for Experiment 2, the first task to be administered was the WM task, followed by the discrimination task and lastly the audiometric screening. Although the certainty scale presented at the end of each trial was again not included in the analyses, it was retained in this experiment in order to allow for an exact replication.

Data analysis

As in the two previous Experiments, data were analyzed using a Bayesian framework implemented in JASP (JASP Team, 2024, Version 0.19.3). Since all participants were at ceiling in the audiometric screening, with 56 participants detecting all 12 frequencies at 15 dB and only five with 11 out of 12 successful detections, we did not include hearing abilities as a covariate.

Results and discussion

Experimental task

Accuracy and d' values

We assessed response accuracy in terms of proportions of correct responses (i.e., hits in the positive conditions and rejection accuracies in the negative conditions) as a function of probe type and target serial position (see Fig. 8). Descriptive statistics can be found in the OSM 1.

When assessing negative trials as a function of probe type and serial position, the model associated with the strongest evidence included both the probe type factor and the interaction between probe type and serial position. This model was 3.80×10^{15} times more likely than the model including only probe type (see Table 2c). Since there was again no evidence in favor of an effect of the serial position factor ($BF_{10} = 0.56$), this variable was added to the null model. Paired t-tests showed an advantage for Distant phonological contrast compared to Close phonological contrast and Phonetic contrast trials ($BF_{10} = 97582.49$ and $BF_{10} = 2.10 \times 10^{+11}$, respectively). As in Experiment 2, performance was again higher on Close phonological contrast trials compared to Phonetic contrast trials ($BF_{10} = 1602.37$). Bayesian one-sample t-tests revealed this time more conclusively that performance on Phonetic contrast trials did not differ from chance ($BF_{01} = 6.87$). However, Close phonological contrast and Distant phonological contrast trials reliably differed from chance ($BF_{10} = 46571.46$ and $BF_{10} = 1.11 \times 10^{+13}$, respectively). When exploring the interaction between probe type and serial position with pairwise Bayesian paired t-tests comparing Close phonological contrast and Phonetic contrast trials, we observed a difference between positions 4-5 as in Experiment 2 ($BF_{10} = 1.93 \times 10^{+10}$), but also a difference at the level of position 3 ($BF_{10} = 17.23$). For this latter comparison, performance was higher for Phonetic contrast trials than

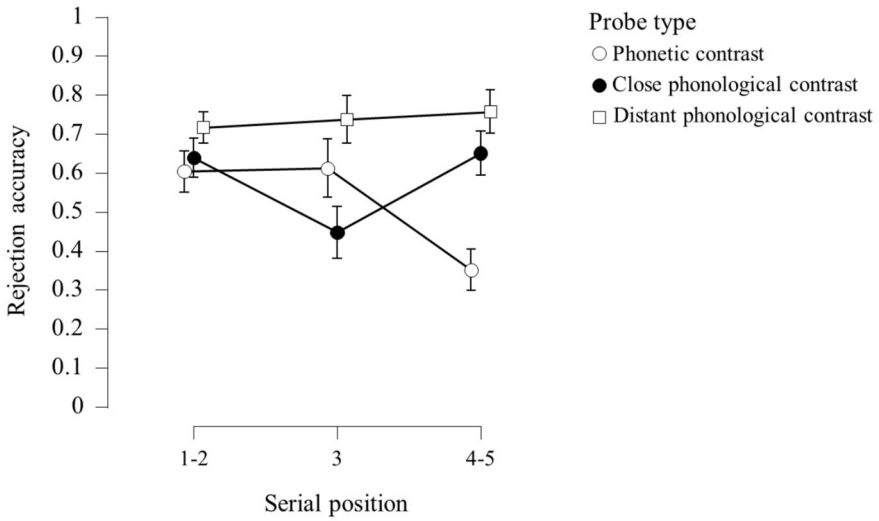

Fig. 8 Experiment 3 – rejection accuracy for negative probes as a function of probe type and target serial position. Error bars represent Bayesian credible intervals

for Close phonological contrast trials; a tendency towards this pattern of result had already been observed in the two previous experiments. As in the two previous experiments, there was no difference at positions 1-2 ($BF_{01} = 4.30$). Finally, when conducting Bayesian one-sample t-tests comparing performance to a chance-level distribution, we observed that for Phonetic contrast trials, performance was above chance level at positions 1-2 and 3 (Bayesian one-sample t-test: $BF_{10} = 115.89$ and $BF_{10} = 10.80$, respectively), but *below* chance level at positions 4-5 ($BF_{10} = 5026.68$), with a very marked inverted recency effect. For Close phonological contrast trials, accuracy was at chance level only at position 3 (Bayesian one-sample t-test: $BF_{01} = 1.60$), but clearly above chance at positions 1-2 and 4-5 ($BF_{10} = 12358.29$ and $BF_{10} = 1026.57$, respectively). For Distant phonological contrast trials, performance was again above chance level at all positions ($BF_{10} = 7.06 \times 10^{+9}$ for positions 1-2, $BF_{10} = 1.19 \times 10^{+6}$ for position 3, and $BF_{10} = 6.90 \times 10^{+8}$ for positions 4-5). In sum, the interaction between probe type and serial position observed in the two previous experiments was also observed in Experiment 3, but this time with a more marked difference between Phonetic contrast and Close phonological contrast trials. While performance on Phonetic contrast trials was above chance both at serial positions 1-2 and 3, performance drastically dropped for end-of-list items, revealing an even more pronounced inverted recency effect than in the two previous experiments.

When assessing positive trials as a function of trial type and serial position, the model associated with the strongest evidence included serial position (see Fig. 9). This model was 5.94 times more likely than the model also including the interaction between serial position and trial type (see Table 3c). As in Experiment 2, there was no evidence in favor of an effect of the trial type factor ($BF_{10} = 0.16$), and this variable was hence added to the null model. Also as in Experiment 2, there was evidence in favor of an *absence* of difference between the two trial types ($BF_{01} = 4.43$), with both trial types well above chance level ($BF_{10} = 2.07 \times 10^{+18}$ for natural trials and $BF_{10} = 6.30 \times 10^{+15}$ for Phonetic contrast trials); this was true for all serial positions (natural trials: $BF_{10} = 1.50 \times 10^{+6}$ for positions 1-2, $BF_{10} = 1.20 \times 10^{+7}$ for position 3, $BF_{10} = 4.65 \times 10^{+23}$ for positions 4-5; Phonetic contrast trials: $BF_{10} = 41.14$ for positions 1-2, $BF_{10} = 869193.72$ for position 3, $BF_{10} = 3.41 \times 10^{+17}$ for positions 4-5). Hence, recognition accuracy confirmed the similarity between the two trial types observed in Experiment 2, with even less difference between the two trial types, as observed by the absence of interaction.

We then conducted a Bayesian repeated-measures ANOVA to assess d' values as a function of probe type. As shown in Table 4c, we observed decisive evidence in favor of an effect of trial type ($BF_{10} = 4.54 \times 10^{+13}$). Mean d' values were 1.36 for Distant phonological contrast trials, 0.90 for Close phonological contrast trials, and 0.59 for Phonetic contrast trials (see Fig. 10). As in the two previous experiments, discrimination accuracy remained very low for Phonetic contrast trials. Bayesian paired-samples t-tests revealed that discrimination performance was higher for Distant phonological contrast trials than for Close phonological contrast trials ($BF_{10} = 198890.34$), and also as in Experiment 2, for Close phonological contrast compared to Phonetic contrast trials ($BF_{10} = 1362.41$)

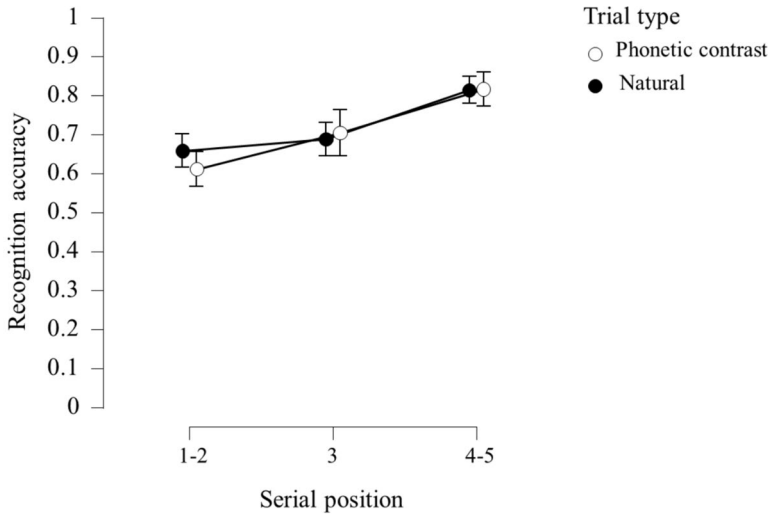

Fig. 9 Experiment 3 – recognition accuracy for positive probes as a function of trial type and target serial position. Error bars represent Bayesian credible intervals

Discrimination task

In a final step, we conducted a repeated-measures ANOVA to assess response accuracy in the discrimination task as a function of pair type (i.e., identical non-words, Distant phonological contrast, Close phonological contrast, Phonetic contrast). There was decisive evidence in favor of pair type. Mean accuracy rate was 0.98 for identical trials, 0.99 for Distant phonological contrast, 0.96 for Close phonological contrast trials, and 0.72 for Phonetic contrast trials. As in the two previous experiments, discrimination accuracy was lower for these latter trials than for the others, but still well above chance level (Bayesian one-sample t-test: $BF_{10} = 5.97 \times 10^{+10}$). These results are in line with those observed in Experiment 1 and 2 and consolidate the assumption that while discrimination accuracy is reliably high for Phonetic contrast trials and even at ceiling for Close phonological contrast trials when presented in minimal pairs, discrimination of the same pairs dramatically drops in a WM context, when there is a delay between the two items of the pair.

In sum, Experiment 3 replicates the main findings of the two previous experiments, showing impaired performance with increasing phonological/phonetic similarity. Due to the larger sample size and increased statistical sensitivity of Experiment 3, some effects that were associated with tendential evidence were now characterized by substantial levels of evidence. While participants are capable of above chance-level discrimination on some serial positions even for Phonetic contrast trials, memory rejection accuracy remains overall unreliable for these trials. These findings and their implications are discussed below.

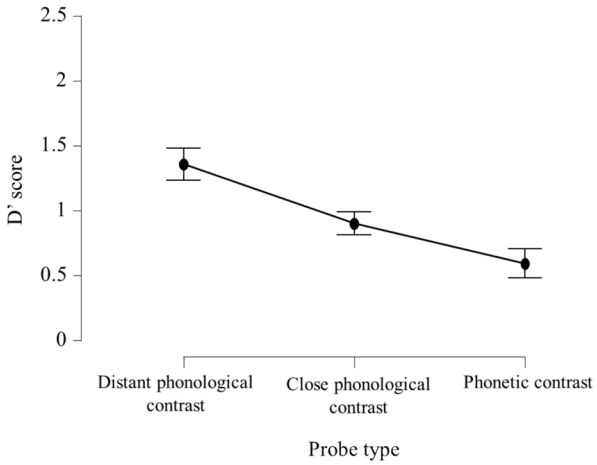

Fig. 10 Experiment 3 – d' scores as a function of probe type. Error bars represent Bayesian credible intervals. Error bars represent Bayesian credible intervals

General discussion

The present study examined the level of precision at which information is processed in verbal WM by manipulating the degree of phonetic and phonological proximity between target and probe non-words in a short-term recognition task. In all experiments, rejection accuracy was limited for phonetically and phonologically close negative probe stimuli, with chance level performance for mid-list and/or end-of-list items. While recognition accuracy for phonetically modified targets in positive trials was also low in Experiment 1, this was not the case in Experiments 2 and 3, which in addition

controlled for the participants' hearing abilities. In Experiments 2 and 3, participants also showed higher rejection accuracy for phonologically close probe stimuli (i.e., Close phonological contrast trials) as compared to phonetically close stimuli (i.e., Phonetic contrast trials), although this advantage was only robust at final list positions in both experiments.

The present study aimed to examine the precision with which verbal information can be maintained in WM. This question has been subject to competing claims. On the one hand, some studies had shown that subtle phonetic features can be maintained and reproduced in a WM context using a task directly and explicitly focusing on phonetic levels of representation (Hepner & Nozari, 2019; Joseph et al., 2015). On the other hand, other studies found no evidence for the use of phonetic-level representations in memory when attention was not explicitly directed towards this level of information (e.g., Rhodes et al., 2019). It should, however, be noted that these latter studies did not directly focus on explicit, verbal WM tasks but rather involved implicit memory paradigms. By directly examining the sensitivity to phonetic and phonological probes in a standard probe recognition verbal WM task, the present study tends to support a position assuming limited phonetic-level representation of memoranda in WM. In all three experiments, we observed very poor discrimination of phonetic deviations between probe and target stimuli for negative probes, with performance being at or close to chance for mid-list and/or end-of-list positions. The finding of a reversed recency effect on phonetically modified negative trials might seem surprising, as studies by Joseph and colleagues (2015) and Hepner and Nozari (2019) on auditory-verbal WM precision had shown recency effects in continuous reproduction paradigms. This difference might stem from the fact that these studies used single CV and VC syllables and that memory load was lower than in our study (up to four syllables per list vs. five non-words and six syllables per list in our study). It might be the case that start-of-list items could be refreshed more often and thus led to more robust and precise memory representations allowing to distinguish stimuli at a subphonemic level, while end-of-list items could not benefit from this longer rehearsal time and hence tended to be more strongly assimilated to the probe non-word. A recent study has indeed observed a reversed recency effect for stimuli associated with less spacing between encoding and recall phases (Sheaffer & Levy, 2022).

Evidence for phonetic-level recognition was observed in positive probes, with above-chance recognition accuracy for Experiment 1 for mid-list and end-of-list items. In Experiments 2 and 3, recognition accuracy for phonetically modified targets reached the accuracy observed for natural targets at start-of-list and mid-list positions in Experiment 2, and at all positions in Experiment 3. While this result may be interpreted as reflecting phonetic-level recognition accuracy, it could also reflect a general tendency to answer "yes" for phonetically modified stimuli, given that the same tendency was observed for phonetically modified negative probes. At the same time, note that the slower response times for the positive phonetically modified stimuli suggests hesitation for these stimuli, and hence at least some sensitivity to the phonetic alterations.

Even at the phonological level, precision remained limited as rejection accuracy for phonologically close probe non-words (differing by a phoneme taken from a very close phoneme category) in all three experiments was also close to chance level for mid-list positions (and also for end-of-list positions in Experiment 1). Only when the differing phonemes of targets and negative probes stemmed from more distant phoneme categories was rejection accuracy reliable and systematically above chance. These results suggest that WM precision is highly limited not only for phonetic information, but also for phoneme-level information. We do not want to imply here that subtle phonological or phonetic levels of representation can never be achieved in WM, but that this level of representation is rather unstable and task context dependent (as a function of serial position or as a function of recognition vs. rejection memory process). Similarly, in the case of a serial recall task, one might expect participants to be able to recall the memoranda, but we also expect repetition errors to occur, involving to a significant extent single phoneme transformations (Acheson & McDonald, 2009; Ellis, 1980; Gupta et al., 2005). For example, Savill et al. (2017) reported 22% of phoneme intrusions (i.e., phonemes other than correct target phonemes) in a serial recall task of non-words. Finally, extralist and intralist intrusions during serial recall were modelled by Guitard et al. (2025) in the Embedded Computational Framework of Memory (eCFM), capturing human-like tendency to recall stimuli not presented at encoding, yet phonologically close to the memoranda. Our results are in line with a recent study which also showed that WM precision is limited at the single-phoneme level (Bouffier et al., 2022). Using a similar probe recognition paradigm to that used here but with words instead of non-words, the authors showed a gradual decrease of rejection accuracy as a function of target-probe phoneme overlap. At the same time, performance remained above chance level even when targets and probes differed at the single-phoneme level. A first reason is the nature of the single phoneme differences, as the phonologically close stimuli used by Bouffier et al. (2022) were closer to Distant phonological contrast trials than to Close phonological contrast trials, with only three out of 25 target-probe pairs differing by their voicing parameter. A second reason is that word stimuli were used by Bouffier et al. (2022), allowing for additional semantic disambiguation and representation of phonologically close stimuli (Hulme et al., 1997; Saint-Aubin & Poirier, 2005). Semantic coding was not

possible in the present study as non-words were used in order to focus directly and exclusively on phonetic and phonological levels of representation in auditory-verbal WM.

The results observed in this study furthermore parallel the findings of the speech perception literature, by revealing context-dependent processing of phonetic levels of speech stimuli. In many language-processing paradigms, speech stimuli are processed at a phonological rather than at a phonetic level. This is shown by the categorical speech perception effect according to which participants are able to discriminate between two sounds if they represent distinct phonemes but not if they represent two acoustical variants close to the prototypical acoustic representation of the same phoneme (Liberman et al., 1957, 1961; see also Iverson & Kuhl, 1995; Kuhl, 1991; Kuhl et al., 1992). However, this does not mean that speech stimuli cannot be processed at all at the acoustic level in language judgment tasks: when participants are invited to rate the acoustic quality of speech stimuli rather than to do a same-different judgment task, then they are capable of acoustic-level judgments (Kuhl, 1991; Massaro & Cohen, 1983; Miller, 1994). These results suggest that our primary processing style involves the phonological level of processing of speech stimuli, and that acoustic levels of processing are referred to only when we are implicitly or explicitly directed towards this level of processing. The present results of limited precision in WM at both phonological and phonetic levels are also in line with other findings from the language-processing literature that indicate limited accuracy for processing speech stimuli at the single-phoneme level. For example, immediate repetition of (short) single non-words is typically not associated with maximal accuracy, with most errors involving single phoneme substitutions (e.g., Vitevitch & Luce, 1999). At the same time, as noted in the *Introduction*, phonetically graded and hence phonological ambiguous information can be maintained (Brown-Schmidt & Toscano, 2017; McMurray et al., 2009; Toscano et al., 2010), and this over up to 35 syllables (Falandays et al., 2020). Consequently, phonetically graded information, or at least the detection that there is such an information that cannot be directly categorized at the phonological level, might not be discarded until sufficient semantic information is gathered from the remaining sentence context to allow for semantic, top-down determination of the nature of the target word. These results are not contradictory to the present findings, as our study focused directly on the ability to maintain phonetic representations and not on delayed phonological categorization processes depending on upcoming semantic disambiguation processes. At a more general level, the present study adds to the broad literature supporting language-based WM models which suggest that verbal WM is strongly determined by the structure and representations of the language system (Acheson & MacDonald, 2009; Baddeley et al., 1998; Burgess & Hitch, 1999, 2006; Cowan, 1995; Gupta, 2009; Majerus, 2013, 2019; Martin et al., 1999; Martin & Saffran, 1992; Martin et al., 1996). These models consider that verbal WM reflects, directly or indirectly, the temporary activation of phonological, lexical and semantic representations that define the language system. Although these models often do not include phonetic levels of processing, they could theoretically be extended to include these levels. At the same time, the present study also suggests that this extension may not be the most critical one as phonetic levels of information only appear to have a limited impact on the precision of representations in WM precision.

The results of this study can also be discussed relative to the more widely studied phonological similarity effect, which is the finding that WM items are recalled more poorly when they are phonologically similar (Conrad & Hull, 1964). While it has traditionally been assumed that it is mostly order recall that is impacted by phonological similarity (Gupta et al., 2005), more recent studies suggest that item recall (i.e., poorer recall of phonologically similar items irrespective of the order in which they are recalled) can also be impacted by phonological similarity if the latter cannot be used to cue item recall (Roodenrys, Guitard, et al., 2022a, Roodenrys, Miller, & Josifovski, 2022b). From a broad perspective, these results may be considered as supporting those of the present study, by suggesting limited representational precision at the item level in verbal WM tasks. However, note that the negative phonological similarity effect observed in the studies by Roodenrys, Miller, and Josifovski was interpreted as reflecting inter-item suppression effects for phonologically overlapping items at the moment of recall. In order to occur, these suppression effects would in fact require that the shared phonological elements are maintained with a sufficient level of precision. Also note that these studies did not manipulate subtle phonetic overlap as we did in the present study.

One interesting aspect of this study that requires further discussion is the observation of similar levels of recognition accuracy for positive natural and phonetically modified stimuli in Experiments 2 and 3, while the latter stimulus category led to substantially reduced recognition accuracy in Experiment 1, which we had considered as reflecting inherently error-prone phonological re-categorization of the ambiguous targets. The only difference between the experiments is the fact that the participants in Experiments 2 and 3 were subjected to audiometric screening. While auditory abilities had little impact per se when introduced as a covariate in the analyses in Experiment 2, implicit task expectations may have altered the participants' encoding strategies in the WM task with a stronger focus on acoustic and phonetic aspects. This interpretation is supported by the similarity of findings between Experiments 2 and 3 in this regard, both experiments using the same procedure of audiometric screening. This kind of implicit expectations have been shown in other contexts to influence task behavior. For example, Dumitru and Pasqualotto (2018) showed that wearing helmets versus

baseball caps influenced participants' visual perception, with participants wearing a helmet providing more accurate estimations of angle and depth for visual targets. The same explanation could also account for the slightly enhanced performance for Close phonological contrast trials in Experiments 2 and 3. It could be argued that the fact that the participants of Experiment 2 also participated in other auditory-verbal tasks (Bouffier & Majerus, in preparation) might have sensitized them to search for subtle target-probe differences. However, the absence of a difference between the two positive conditions was replicated in Experiment 3, in which participants did not carry out any additional WM task. A more likely explanatory factor is the smaller sample size in Experiment 1, which may not have allowed for a full characterization of the effects under investigation. The main effects reported in this study became increasingly robust and clear-cut with progressive increase of the sample size, from Experiment 1 to Experiment 3. It is furthermore worth mentioning that the discrimination task yielded similar results in Experiments 1 and 2, and even slightly poorer discrimination performance for Phonetic contrast pairs in Experiment 3. It could be argued that this invalidates our hypothesis of increased sensitivity to subtle acoustic differences induced by the announcement of audiometric screening. However, the discrimination task was administered each time after the WM task, and thus participants were, at this time point, equally familiarized to the type of stimuli and contrasts that were being probed. Furthermore, the discrimination task explicitly focused on auditory discrimination abilities (i.e., hearing differences between stimuli), greatly reducing the importance of the impact implicit acoustic sensitization processes may have at this level.

Although the present study has focused on the voicing feature in the phonetic and close phonological contrasts, we can hypothesize that similar results could be observed for other types of contrasts. For example, early work by Liberman et al. (1957) on the /b-d-g/-continuum, which involves changes at the level of formant transitions, shows a similar tendency to categorize stimuli into supra-ordinate categories as for the voicing manipulations used in the present study. These variations will also impact verbal WM performance as shown by a study by Schweppe et al. (2011), who observed impaired recall for sequences involving rapid changes in format transitions such as /pa-ta-ka/compared to sequences composed of more dissimilar phoneme categories such as /fa-na-ga/(Schweppe et al. 2011).

Similarly, one limit of this study is that phonetic and phonological deviations only included consonants. The reason was that most consonants are very brief speech stimuli, in comparison to vowels, and hence need particularly precise representations to be accurately encoded and maintained. Early work furthermore seems to suggest that vowels include a smaller number of dimensions than consonants (Wickelgren, 1965, 1966). Moreover, in school-aged children, repetition accuracy of consonants, but not of vowels, distinguishes good readers from poorer readers (Brady et al., 1987). In the present study, different results may, however, be obtained when using vowels. In the speech perception literature, it has indeed been shown that vowels may be perceived both at phonetic and phonological levels in different task contexts (Fry et al., 1962; Pisoni, 1973; Schouten & van Hessen, 1992). Vowels also carry information about rhythm and intonation, as well as about a speaker's emotion or dialect, and hence may require more extended acoustic-level processing (Fry et al., 1962). Other studies have, however, shown that these acoustic-phonetic features may not be maintained for an extended time, as vowel discrimination performance rapidly decreases when a temporal delay is inserted between two vowels to be discriminated (Pisoni, 1973). Joseph et al. (2015) assessed WM precision for vowels and nevertheless showed that at least at low WM loads and for a task setup explicitly directing attention towards acoustic levels of processing, phonetic variations around a given vowel can be maintained and reproduced for a delay of up to 25 s. However, we do not know to what extent phonetic levels of information characterizing vowels are maintained in WM tasks when using more naturalistic stimuli and task setups.

Another caveat concerns the use of a 3-s retention interval. The use of this delay allowed to discard purely sensory processes. However, increased rejection accuracy for phonetically/phonemically close target-probe negative pairs may be observed if this delay is reduced or removed, as also suggested by the high performance in the discrimination tasks, which did not include any delay.

A further limit is the absence of manipulation of WM load (i.e., the number of stimuli to be maintained). Evidence shows that a higher WM load leads to reliance on more lexically mediated processes and a limited ability to process fine-grained phonetic information (Mattys & Wiget, 2011). This is in line with the present study, which showed very high accuracy in the discrimination task (i.e., a very low WM load task condition) but impaired recognition accuracy for the same stimuli in the high WM load task condition we used (five non-words per list). Joseph et al. (2015) and Hepner and Nozari (2019), in addition of using paradigms explicitly focusing the participants' attention on phonetic level of representation, used low WM load task conditions and observed evidence for representation at the phonetic level in WM. Future studies should use a parametric modulation of WM load in order to determine whether sensitivity to phonetic level of detail in WM is task-dependent or WM load-dependent.

To conclude, the present study shows very limited evidence for a spontaneous representation of memoranda at a

phonetic level in verbal WM, i.e., when participants' attention is not directly and explicitly directed towards phonetic levels of processing. Our findings mirror, in the verbal WM domain, the dominance of phonological levels of representation that is also observed in natural language-processing conditions. The precision of verbal WM representations appears to depend on the nature and resolution of the representations in the language system itself.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.3758/s13421-025-01768-z.

Funding This work was supported by grant F.R.S.-FNRS N°1.A.238.17F (Fund for Scientific Research FNRS, Belgium). We would like to thank Dounia Benchegra, Alexia Campanella, Lucie Cheval, Simon Delvenne, Noémie Depierreux, Ilyas Gutierrez Suarez, Loïc Harsin, Elise Mestré, Vanessa Mias and Rajana Salamova and for their help in data acquisition.

Data availability The data and materials for all Experiments are available via the Open Science Framework (OSF) repository: https://osf.io/e8gah/?view_only=d0d24de505d14ecebd947b4ad865fe80

None of the experiments was preregistered.

Code availability The code used to carry out sensitivity analyses can be found at https://github.com/nicebread/BFDA (Schönbrodt & Stefan, 2018). All other analyses have been carried out using JASP (JASP team, 2024, Version 0.19.3).

Declarations

Conflicts of interest The authors declare that they have no conflicts of interest.

Ethics approval Experiments were performed in line with the principles of the Declaration of Helsinki. Approval was granted by the Ethics Committee of the University of Liège (project number: 2016/358 for Experiments 1 and 2 and 2223-051 for Experiment 3).

Consent to participate Informed consent was obtained from all individual participants included in the study.

Consent for publication Participants signed informed consent regarding publishing their data on repositories.

References

Acheson, D. J., & MacDonald, M. C. (2009). Verbal working memory and language production: Common approaches to the serial ordering of verbal information. *Psychological Bulletin,135*(1), 50–68. https://doi.org/10.1037/a0014411

Atkins, A. S., & Reuter-Lorenz, P. A. (2008). False working memories? Semantic distortion in a mere 4 seconds. *Memory & Cognition,36*(1), 74–81. https://doi.org/10.3758/MC.36.1.74

Baddeley, A., Gathercole, S., & Papagno, C. (1998). The phonological loop as a language learning device. *Psychological Review, 105*(1), 158–173. https://doi.org/10.1037/0033-295x.105.1.158

Bays, P. M., Catalao, R. F. G., & Husain, M. (2009). The precision of visual working memory is set by allocation of a shared resource. *Journal of Vision, 9*(10), 7–7. https://doi.org/10.1167/9.10.7

Beckman, J. N. (1998). *Positional faithfulness: An optimality theoretic treatment of phonological asymmetries*. Routledge.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International, 5*, 341–345.

Bouffier, M., & Majerus, S. (2025). *Working memory precision for phonological and semantic information [Manuscript in preparation]*. University of Liège.

Bouffier, M., Poncelet, M., & Majerus, S. (2022). The linguistic constraints of precision of verbal working memory. *Memory & Cognition, 50*(7), 1464–1485. https://doi.org/10.3758/s13421-022-01283-5

Brady, S., Mann, V., & Schmidt, R. (1987). Errors in short-term memory for good and poor readers. *Memory & Cognition,15*(5), 444–453. https://doi.org/10.3758/bf03197734

Brown-Schmidt, S., & Toscano, J. C. (2017). Gradient acoustic information induces long-lasting referential uncertainty in short discourses. *Language, Cognition and Neuroscience, 32*(10), 1211–1228. https://doi.org/10.1080/23273798.2017.1325508

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition, 2*(1), 16. https://doi.org/10.5334/joc.72

Burgess, N., & Hitch, G. J. (1999). Memory for serial order: A network model of the phonological loop and its timing. *Psychological Review,106*(3), 551–581. https://doi.org/10.1037/0033-295X.106.3.551

Burgess, N., & Hitch, G. J. (2006). A revised model of short-term memory and long-term learning of verbal sequences. *Journal of Memory and Language, 55*(4), 627–652. https://doi.org/10.1016/j.jml.2006.08.005

Clark, K. M., Hardman, K. O., Schachtman, T. R., Saults, J. S., Glass, B. A., & Cowan, N. (2018). Tone series and the nature of working memory capacity development. *Developmental Psychology, 54*(4), 663–676. https://doi.org/10.1037/dev0000466

Connine, C. M., Blasko, D. G., & Hall, M. (1991). Effects of subsequent sentence context in auditory word recognition: Temporal and linguistic constraints. *Journal of Memory and Language,30*(2), 234–250. https://doi.org/10.1016/0749-596X(91)90005-5

Conrad, R., & Hull, A. J. (1964). Information, acoustic confusion and memory span. *British Journal of Psychology,55*(4), 429–432. https://doi.org/10.1111/j.2044-8295.1964.tb00928.x

Cowan, N. (1995). *Attention and memory: An integrated framework*. Clarendon Press.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences,24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922

Cowan, N. (2010). The magical mystery four: How is working memory capacity limited, and why? *Current Directions in Psychological Science,19*(1), 51–57. https://doi.org/10.1177/0963721409359277

Crowder, R. G., & Morton, J. (1969). Precategorical acoustic storage (PAS). *Perception & Psychophysics,5*(6), 365–373. https://doi.org/10.3758/BF03210660

Dobbins, I. G., Khoe, W., Yonelinas, A. P., & Kroll, N. E. (2000). Predicting individual false alarm rates and signal detection theory: A role for remembering. *Memory & Cognition, 28*(8), 1347–1356. https://doi.org/10.3758/bf03211835

Dumitru, M. L., & Pasqualotto, A. (2018). Helmets improve estimations of depth and visual angle to safe targets. *Attention, Perception, & Psychophysics,80*(8), 1879–1884. https://doi.org/10.3758/s13414-018-1605-9

Ellis, A. W. (1980). Errors in speech and short-term memory: The effects of phonemic similarity and syllable position. *Journal of Verbal Learning and Verbal Behavior,19*(5), 624–634. https://doi.org/10.1016/S0022-5371(80)90672-6

Falandays, J. B., Brown-Schmidt, S., & Toscano, J. C. (2020). Long-lasting gradient activation of referents during spoken language processing. *Journal of Memory and Language, 112*, 104088. https://doi.org/10.1016/j.jml.2020.104088

Fry, D. B., Abramson, A. S., Eimas, P. D., & Liberman, A. M. (1962). The identification and discrimination of synthetic vowels. *Language and Speech, 5*(4), 171–189. https://doi.org/10.1177/002383096200500401

Garnier, M., Dohen, M., Buttiaux, L., Gerber, S. (2018). Clarification et correction d'indices segmentaux: Une étude pilote sur les consonnes occlusives du français. In: *Proceedings. XXXIIe Journées d'Études sur la Parole*, 478-486. https://doi.org/10.21437/JEP.2018-55

Gorgoraptis, N., Catalao, R. F. G., Bays, P. M., & Husain, M. (2011). Dynamic updating of working memory resources for visual objects. *Journal of Neuroscience, 31*(23), 8502–8511. https://doi.org/10.1523/JNEUROSCI.0208-11.2011

Guitard, D., Saint-Aubin, J., Reid, J. N., & Jamieson, R. K. (2025). An embedded computational framework of memory: The critical role of representations in veridical and false recall predictions. *Psychonomic Bulletin & Review*, https://doi.org/10.3758/s13423-025-02669-7. Advance online publication.

Gupta, P. (2009). A computational model of non-word repetition, immediate serial recall, and non-word learning. In A. S. C. Thorn & M. P. A. Page (Eds.), *Interactions between short-term and long-term memory in the verbal domain* (pp. 108–135). Psychology Press.

Gupta, P., Lipinski, J., & Aktunc, E. (2005). Reexamining the phonological similarity effect in immediate serial recall: The roles of type of similarity, category cuing, and item recall. *Memory & Cognition,33*(6), 1001–1016. https://doi.org/10.3758/bf03193208

Hamilton, A. C., & Martin, R. C. (2007). Proactive interference in a semantic short-term memory deficit: Role of semantic and phonological relatedness. *Cortex,43*(1), 112–123. https://doi.org/10.1016/S0010-9452(08)70449-0

Hepner, C. R., & Nozari, N. (2019). Resource allocation in phonological working memory: Same or different principles from vision? *Journal of Memory and Language,106*, 172–188. https://doi.org/10.1016/j.jml.2019.03.003

Hulme, C., Roodenrys, S., Schweickert, R., Brown, G. D. A., Martin, S., & Stuart, G. (1997). Word-frequency effects on short-term memory tasks: Evidence for a redintegration process in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(5), 1217–1232. https://doi.org/10.1037/0278-7393.23.5.1217

Iverson, P., & Kuhl, P. K. (1995). Mapping the perceptual magnet effect for speech using signal detection theory and multidimensional scaling. *The Journal of the Acoustical Society of America, 97*(1), 553–562. https://doi.org/10.1121/1.412280

JASP Team (2019). JASP (Version 0.11.1) [Computer software].

JASP Team (2024). JASP (Version 0.19.3)[Computer software]. https://jaspstats.org/faq/how-do-i-cite-jasp/

Jeffreys, H. (1998). *Theory of probability* (3rd ed). Clarendon Press , Oxford University Press.

Joseph, S., Iverson, P., Manohar, S., Fox, Z., Scott, S. K., & Husain, M. (2015). Precision of working memory for speech sounds. *Quarterly Journal of Experimental Psychology, 68*(10), 2022–2040. https://doi.org/10.1080/17470218.2014.1002799

Kruschke, J. K. (2011). Bayesian assessment of null values via parameter estimation and model comparison. *Perspectives on Psychological Science, 6*(3), 299–312. https://doi.org/10.1177/1745691611406925

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & Psychophysics,50*(2), 93–107. https://doi.org/10.3758/BF03212211

Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science,255*(5044), 606–608.

Lee, M. D., & Wagenmakers, E.-J. (2013). *Bayesian Cognitive Modeling: A Practical Course*. Cambridge University Press. https://doi.org/10.1017/CBO9781139087759

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

Liberman, A. M., Harris, K. S., Hoffman, H. S., & Griffith, B. C. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology, 54*(5), 358–368. https://doi.org/10.1037/h0044417

Liberman, A. M., Harris, K. S., Kinney, J. A., & Lane, H. (1961). The discrimination of relative onset-time of the components of certain speech and nonspeech patterns. *Journal of Experimental Psychology, 61*(5), 379–388. https://doi.org/10.1037/h0049038

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature Neuroscience, 17*(3), 347–356. https://doi.org/10.1038/nn.3655

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Lawrence Erlbaum Associates.

Majerus, S. (2013). Language repetition and short-term memory: An integrative framework. *Frontiers in Human Neuroscience,7*, 357. https://doi.org/10.3389/fnhum.2013.00357

Majerus, S. (2019). Verbal working memory and the phonological buffer: The question of serial order. *Cortex,112*, 122–133. https://doi.org/10.1016/j.cortex.2018.04.016

Majerus, S., & Lorent, J. (2009). Is phonological short-term memory related to phonological analysis stages in auditory sentence processing? *European Journal of Cognitive Psychology, 21*(8), 1200–1225. https://doi.org/10.1080/09541440902733216

Martin, N., & Saffran, E. M. (1992). A computational account of deep dysphasia: Evidence from a single case study. *Brain and Language,43*(2), 240–274. https://doi.org/10.1016/0093-934X(92)90130-7

Martin, N., Saffran, E. M., & Dell, G. S. (1996). Recovery in deep dysphasia: Evidence for a relation between auditory–verbal STM capacity and lexical errors in repetition. *Brain and Language, 52*(1), 83–113. https://doi.org/10.1006/brln.1996.0005

Martin, R. C., Lesch, M. F., & Bartha, M. C. (1999). Independence of input and output phonology in word processing and short-term memory. *Journal of Memory and Language, 41*(1), 3–29. https://doi.org/10.1006/jmla.1999.2637

Massaro, D. W., & Cohen, M. M. (1983). Categorical or continuous speech perception: A new test. *Speech Communication,2*(1), 15–35. https://doi.org/10.1016/0167-6393(83)90061-4

Mathôt, S., Schreij, D., & Theeuwes, J. (2012). OpenSesame: An open-source, graphical experiment builder for the social sciences. *Behavior Research Methods,44*(2), 314–324. https://doi.org/10.3758/s13428-011-0168-7

Mattys, S. L., & Wiget, L. (2011). Effects of cognitive load on speech recognition. *Journal of Memory and Language, 65*(2), 145–160. https://doi.org/10.1016/j.jml.2011.04.004

McMurray, B., Tanenhaus, M. K., & Aslin, R. N. (2009). Within-category VOT affects recovery from "lexical" garden paths: Evidence against phoneme-level inhibition. *Journal of Memory and Language,60*(1), 65–91. https://doi.org/10.1016/j.jml.2008.07.002

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review,63*(2), 81–97. https://doi.org/10.1037/h0043158

Miller, J. L. (1994). On the internal structure of phonetic categories: A progress report. *Cognition,50*(1–3), 271–285. https://doi.org/10.1016/0010-0277(94)90031-0

Mızrak, E., & Oberauer, K. (2021). What is time good for in working memory? *Psychological Science, 32*(8), 1325–1337. https://doi.org/10.1177/0956797621996659

Monsell, S. (1978). Recency, immediate recognition memory, and reaction time. *Cognitive Psychology, 10*(4), 465–501. https://doi.org/10.1016/0010-0285(78)90008-7

Moore, D. R., Rosenberg, J. F., & Coleman, J. S. (2005). Discrimination training of phonemic contrasts enhances phonological processing in mainstream school children. *Brain and Language,94*(1), 72–85. https://doi.org/10.1016/j.bandl.2004.11.009

Neath, I., Hockley, W. E., & Ensor, T. M. (2022). Stimulus-based mirror effects revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 48*(12), 1833–1849. https://doi.org/10.1037/xlm0000901

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods,36*(3), 516–524. https://doi.org/10.3758/BF03195598

Oberauer, K., & Lin, H.-Y. (2017). An interference model of visual working memory. *Psychological Review, 124*(1), 21–59. https://doi.org/10.1037/rev0000044

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics,44*(4), 369–378. https://doi.org/10.3758/BF03210419

Pisoni, D. B. (1973). Auditory and phonetic memory codes in the discrimination of consonants and vowels. *Perception & Psychophysics,13*(2), 253–260. https://doi.org/10.3758/BF03214136

Rhodes, R., Han, C., & Hestvik, A. (2019). Phonological memory traces do not contain phonetic information. *Attention, Perception, & Psychophysics,81*(4), 897–911. https://doi.org/10.3758/s13414-019-01728-1

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

Roodenrys, S., Guitard, D., Miller, L. M., Saint-Aubin, J., & Barron, J. M. (2022a). Phonological similarity in the serial recall task hinders item recall, not just order. *British Journal of Psychology,113*(4), 1100–1120. https://doi.org/10.1111/bjop.12575

Roodenrys, S., Miller, L. M., & Josifovski, N. (2022b). Phonemic interference in short-term memory contributes to forgetting but is not due to overwriting. *Journal of Memory and Language, 122*, 104301. https://doi.org/10.1016/j.jml.2021.104301

Saint-Aubin, J., & Poirier, M. (2005). Word frequency effects in immediate serial recall: Item familiarity and item co-occurrence have the same effect. *Memory, 13*(3-4), 325–332. https://doi.org/10.1080/09658210344000369

Savill, N., Ellis, A. W., & Jefferies, E. (2017). Newly-acquired words are more phonologically robust in verbal short-term memory when they have associated semantic representations. *Neuropsychologia, 98*, 85–97. https://doi.org/10.1016/j.neuropsychologia.2016.03.006

Schönbrodt, F. D., & Stefan, A. M. (2018). BFDA: An R package for Bayes factor design analysis (version 0.4.0). Retrieved from https://github.com/nicebread/BFDA

Schouten, M. E. H., & van Hessen, A. J. (1992). Modeling phoneme perception. I : Categorical perception. *The Journal of the Acoustical Society of America, 92*(4), 1841–1855. https://doi.org/10.1121/1.403841

Schweppe, J., Grice, M., & Rummer, R. (2011). What models of verbal working memory can learn from phonological theory: Decomposing the phonological similarity effect. *Journal of Memory and Language, 64*(3), 256–269.

Sheaffer, R., & Levy, D. A. (2022). Negative recency effects in delayed recognition: Spacing, consolidation, and retrieval strategy processes. *Memory & Cognition,50*(8), 1683–1693. https://doi.org/10.3758/s13421-022-01293-3

Toscano, J. C., McMurray, B., Dennhardt, J., & Luck, S. J. (2010). Continuous perception and graded categorization: Electrophysiological evidence for a linear relationship between the acoustic signal and perceptual encoding of speech. *Psychological Science,21*(10), 1532–1540. https://doi.org/10.1177/0956797610384142

Verhaegen, C., Collette, F., & Majerus, S. (2014). The impact of aging and hearing status on verbal short-term memory. *Aging, Neuropsychology, and Cognition, 21*(4), 464–482. https://doi.org/10.1080/13825585.2013.832725

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language, 40*(3), 374–408. https://doi.org/10.1006/jmla.1998.2618

Wagenmakers, E. J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., & Morey, R. D. (2018). Bayesian inference for psychology part. I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review,25*(1), 35–57. https://doi.org/10.3758/s13423-017-1343-3

Wickelgren, W. A. (1965). Distinctive features and errors in short-term memory for English vowels. *The Journal of the Acoustical Society of America, 38*(4), 583–588. https://doi.org/10.1121/1.1909750

Wickelgren, W. A. (1966). Distinctive features and errors in short-term memory for English consonants. *Journal of the Acoustical Society of America, 39*(2), 388–398.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision, 4*(12), 11. https://doi.org/10.1167/4.12.11

Zokaei, N., Gorgoraptis, N., Bahrami, B., Bays, P. M., & Husain, M. (2011). Precision of working memory for visual motion sequences and transparent motion surfaces. *Journal of Vision, 11*(14), 2–2. https://doi.org/10.1167/11.14.2