A multilingual preregistered replication of the semantic mismatch effect on serial recall

Jan Philipp Röer [a], Raoul Bell [b], Axel Buchner [b], Jean Saint-Aubin [c], René-Pierre Sonier [c], John E. Marsh [d, e], Stuart B. Moore [f], Matthew B. A. Kershaw [d], Robert Ljung [g], and Sebastian Arnström [g]

[a] Department of Psychology and Psychotherapy, Witten/Herdecke University, Witten, Germany

[b] Department of Experimental Psychology, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

[c] École de Psychologie, Université de Moncton, New Brunswick, Canada

[d] School of Psychology, University of Central Lancashire, Preston, Lancashire, United Kingdom

[e] Luleå University of Technology, Department of Business Administration, Technology and Social Sciences, Humans and Technology, Sweden

[f] School of Psychology, Keele University, Staffordshire, United Kingdom

[g] Faculty of Engineering and Sustainable Development, University of Gävle, Gävle, Sweden

Correspondence concerning this article should be addressed to

Jan Philipp Röer

Department of Psychology and Psychotherapy

Witten/Herdecke University

Alfred-Herrhausen-Straße 44

58455 Witten, Germany

Telephone number: +49 2302 926 845

E-mail address: jan.roeer@uni-wh.de

Abstract

Visual-verbal serial recall is disrupted when task-irrelevant background speech has to be ignored. Contrary to previous suggestion, it has recently been shown that the magnitude of disruption may be accentuated by the semantic properties of the irrelevant speech. Sentences ending with unexpected words that did not match the preceding semantic context were more disruptive than sentences ending with expected words. This particular instantiation of a deviation effect has been termed the semantic mismatch effect. To establish a new phenomenon, it is necessary to show that the effect can be independently replicated and does not depend on specific boundary conditions such as the language of the stimulus material. Here we report a preregistered replication of the semantic mismatch effect in which we examined the effect of unexpected words in four different languages (English, French, German, and Swedish) across four different laboratories. Participants performed a serial recall task while ignoring sentences with expected or unexpected words that were recorded using text-to-speech software. Independent of language, sentences ending with unexpected words were more disruptive than sentences ending with expected words. In line with previous results, there was no evidence of habituation of the semantic mismatch effect in the form of a decrease in disruption with repeated exposure to the occurrence of unexpected words. The successful replication and extension of the effect to different languages indicates the expression of a general and robust mechanism that reacts to violations of expectancies based on the semantic content of the irrelevant speech.

Keywords: auditory distraction, irrelevant speech, selective attention, working memory

Running Head: A multilingual replication of the semantic mismatch effect

A multilingual preregistered replication of the semantic mismatch effect on serial recall

It is a well-established and widely known phenomenon that cognitive performance declines when task-irrelevant background speech has to be ignored. The standard paradigm to study the effect of irrelevant speech in the laboratory involves the serial recall task (Colle & Welsh, 1976; Ellermeier & Zimmer, 1997; Jones & Macken, 1993). In this task, a list of to-be-remembered items—usually a random permutation of the digits 1 to 9—is sequentially presented on a computer screen. Immediately after the presentation of the last item or after a retention interval of a few seconds, recall is prompted, and participants are asked to recall the items in the order in which they had been presented. When task-irrelevant speech is played during the presentation and/or retention of the items, memory for the correct order of the items is poorer than when no task-irrelevant speech is played (Beaman & Jones, 1997; Bell, Röer, Lang, & Buchner, 2019a, 2019b; Elliott, 2002; Jones, Saint-Aubin, & Tremblay, 1999; Parmentier & Beaman, 2015).

For a long time, it was assumed that the content of irrelevant speech was not an important ingredient of the disruption it produces to serial recall. A finding that has received a lot of attention, for example, is that it makes little, if any, difference whether irrelevant speech is played forward or backward (LeCompte, Neely, & Wilson, 1997; Surprenant, Neath, & Bireta, 2007; see also Jones, Miles & Page, 1990). This finding has been replicated several times by different research groups (Marsh, Hughes, & Jones, 2009; Röer, Bell, & Buchner, 2014; but see Ueda, Nakajima, Ellermeier, & Kattner, 2017). Similarly, irrelevant speech in a familiar language is found to be as disruptive as irrelevant speech in an unfamiliar language (Ellermeier, Kattner, Ueda, Duomoto, & Nakajima, 2015; Jones et al., 1990; Marsh et al., 2009). The semantic similarity between the to-be-remembered items and the to-be-ignored distractors also has often surprisingly little effect on serial recall (Buchner et al., 1996; Marsh et al., 2008; but see Neely & LeCompte, 1999). These findings have been used to discard semantic properties as relevant to the disruption of serial recall. For example, Ellermeier and Zimmer (2014) summarized the literature as suggesting that “[t]he semantics of the

irrelevant speech do not seem to be important" (p. 11), and, in consequence, argued for a psychoacoustic perspective of the irrelevant speech effect.

Based on findings showing that the processing of the irrelevant stimuli does interfere with tasks that require semantic processing of the relevant stimuli such as reading comprehension, proof reading, or free recall of semantically related material (Bell et al., 2008; Jones et al., 1990; Marsh et al., 2008, 2009), an *interference-by-process* account (Marsh & Jones, 2010; Marsh et al., 2008, 2009) has been proposed according to which interference is a function of the similarity between the processing required by the primary task and the pre-attentive processing of the auditory distractors. For instance, the automatic seriation of the to-be-ignored sound is assumed to yield order cues that interfere with the order cues during articulatory-based rehearsal of the items during a serial recall task. One reason why the semantic properties of the auditory distractors do not seem to interfere with serial recall performance may be that the serial recall task makes virtually no demands on the processing of meaning: Usually, digits drawn from the same small and extremely well-known set are to be recalled in every trial so that the primary demand of the task is to recall the order of those digits whereas remembering their identities is a trivial task component.

There are two possible explanations for the apparent lack of semantic interference in the serial recall paradigm (Marsh et al., 2014). Given that the auditory modality is completely irrelevant and can be entirely ignored in the paradigm, it seems possible to speculate that the analysis of to-be-ignored speech does not reach a semantic level and that one is able to effectively ignore the content of task-irrelevant background speech altogether. According to this first hypothesis, the analysis of the irrelevant speech is blocked before its meaning is properly analyzed—a view taken by early proponents of structural models of selective attention according to which an inherent processing limitation necessitates selection attention, thereby preventing the categorization or identification of task-irrelevant information (Broadbent, 1957, 1958; Eriksen & St. James, 1986; Treisman, 1960). On these approaches, a filter at an early stage within a set of discrete processing stages is required to prevent the massive inflow of perceptual information from overloading the cognitive system. This filter allows information pertaining to the pre-categorical physical properties of senso-

ry information (e.g., pitch, timbre, intensity, spatial location) to pass through to capacity-limited processing stages for categorization and identification, but prevents (Broadbent, 1958, 1971) or attenuates (Treisman, 1964, 1969) the entry of post-categorical semantic properties (but see, Deutsch & Deutsch, 1963; Duncan, 1980). However, a second hypothesis is equally suited to account for the apparent lack of semantic interference in the serial recall paradigm: The meaning of the auditory distractors is always processed, but the processing usually occurs automatically, so that interference only manifests when the semantic processing of the irrelevant information comes into direct conflict with the semantic processing of the relevant information required by the primary task.

A growing body of empirical work provides evidence in favor of the second hypothesis according to which the meaning of the distractors is always processed but does not always come into conflict with the serial recall task (Marsh et al., 2014; Röer et al., 2017b). For instance, in a recent study (Röer, Körner, Buchner, & Bell, 2017b), auditory distractor words from different semantic categories had to be ignored during a standard serial recall task. In line with previous findings (e.g., Jones et al., 1990), these words were equally disruptive when they were played forward or backward. In what participants were led to believe to be an unrelated norming study, they were asked to spontaneously produce exemplars of categories from which the distractor words had been drawn. Previously presented distractor words were produced with a higher probability than words from a matched set, suggesting that the content of irrelevant speech, while not always having an observable effect on ongoing performance, is still processed to the degree that it may influence behavior in a subsequent task. These findings are clearly at odds with the view that the semantic properties of task-irrelevant speech are filtered at an early stage of processing (Broadbent, 1958, 1975; Treisman, 1964; 1969).

The evidence mentioned so far is in line with the interference-by-process principle (Marsh et al., 2008, 2009) according to which understanding auditory distraction requires a careful analysis of the processes involved in the serial recall task and in the automatic processing of the auditory distractors. However, the principle can be applied less well to explain results demonstrating that the content of irrelevant speech has a disruptive effect on serial recall performance when distrac-

tors are self-relevant or emotional (Buchner, Mehl, Rothermund, & Wentura, 2006; Buchner, Rothermund, Wentura, & Mehl, 2004; Marsh et al., 2018; Röer, Bell, & Buchner, 2013; Röer, Körner, Buchner, & Bell, 2017a). For example, sentences containing one's own name were found to be more disruptive than sentences containing a yoked-control partner's name (Röer et al., 2013) and taboo words were more disruptive than neutral words (Röer et al., 2017a). These findings suggest that the emotional or self-relevant meaning of the distractor words may capture attention, and thus require theories to specify a role of attention in the disruption of serial recall by auditory distractors (e.g., Bell et a., 2019; Hughes, 2014). However, acoustic properties of the to-be-ignored words cannot be completely ruled out as eliciting factors because one's own name and taboo words are highly overlearned stimuli that may capture attention based on their acoustic properties (Röer et al., 2019).

It is thus of high theoretical interest that, recently, semantic effects have been reported that go beyond that of the individual word meaning. Specifically, words without any inherent capacity to capture attention gain disruptive power when presented in a mismatching semantic context. Vachon et al. (2020) found that distractor sequences that contain a single deviant item from a different category disrupt serial recall more than sequences without such a deviation. This categorical deviant effect was demonstrated with a single letter in a sequence of digits, with a digit in a sequence of letters, and with a word from a different semantic category (e.g., a tool in a list of fruit). Furthermore, sentences ending with unexpected words that did not match the preceding semantic context were found to be more disruptive than sentences ending with expected words (Röer, Bell, Körner, & Buchner, 2019). This semantic mismatch effect occurred regardless of whether the sentences were familiar proverbs or novel sentences with no specific long-term memory representations (see also Röer, Buchner, & Bell, 2020). The size of the semantic mismatch effect is comparable to that of the auditory deviant effect which has had a large impact on theory development over the last decade (Bell, Mieth, Röer, Troche, & Buchner, 2019; Bell, Röer, Marsh, Storch, & Buchner, 2017; Hughes, Hurlstone, Marsh, Vachon, & Jones, 2013; Hughes, Vachon, & Jones, 2005; Hughes, Vachon, & Jones, 2007; Körner, Röer, Buchner, & Bell, 2017, 2019; Röer, Bell, Marsh, &

Buchner, 2015; Vachon, Labonté, & Marsh, 2017) and was recently included as one of the benchmark findings that theories of working memory should be able to account for (Oberauer et al., 2018). Within the few experiments that are available, the semantic mismatch effect has also been found to be quite stable in the sense that sentence endings with semantic mismatches were found to maintain their disruptive power even after several repetitions. The effect therefore seems to be more persistent than other forms of auditory distraction such as, for example, the disruptive effect of one's own name which substantially decreases after a few repetitions (Röer et al., 2013).

The semantic mismatch effect is thus a newly discovered phenomenon with high theoretical leverage for theories on auditory distraction and working memory as it allows one to draw conclusions about the fate of to-be-ignored auditory information. However, before theories on attention and working memory are adapted to account for a novel phenomenon, it is essential to establish that this phenomenon is indeed reproducible (Simons, 2014). While it seems promising that the semantic mismatch effect has already been successfully replicated in several experiments, an important caveat at this point is that these replications have all been reported in the same language from a single laboratory. Obviously, trust in a newly discovered effect substantially increases if it can be replicated in different languages and laboratories. For example, as a prerequisite for including a phenomenon in the list of benchmark findings, an important criterion is reproducibility, and it was explicitly mentioned that replications from *different* laboratories are to be preferred over replications in one laboratory (Oberauer et al., 2018). The present study provides such a multiple-laboratory (and multiple-language) replication of the semantic mismatch effect. As a further method to increase trust in the newly discovered finding, the replications presented here have been preregistered to provide an unbiased estimate of its replicability (Nosek, Ebersole, DeHaven, & Mellor, 2018). We preregistered our method, materials, and planned analyses prior to the start of the data collection.

Replicating a theoretically interesting, newly discovered effect across multiple laboratories is desirable, not least in order to demonstrate that the effect does not depend on highly specific boundary conditions that will make it difficult, or even impossible, to replicate the effect in other la-

boratories. A famous example for this is the fourth experiment of Baddeley, Thomson, and Buchanan (1975) who equated short and long words on all dimensions except pronunciation time and found an advantage of short words over long words. With the exact same words, the effect has been replicated many times (e.g., Cowan et al., 1992; Longoni, Richardson, & Aiello, 1993; Lovatt, Avons, & Masterson, 2000; Nairne, Neath, & Serra, 1997). However, all attempts so far to create another set of short and long words that differ only in pronunciation time found no difference in the recall of short and long words (Caplan, Rochon, & Waters, 1992; Lovatt et al., 2000; Neath, Bireta, & Surprenant, 2003; Service, 1998).

Replications can hardly ever be “exact” (Hüffmeier, Mazei, & Schultze, 2016)—changes are necessarily introduced by replicating a finding in another laboratory as the sample is drawn from a different population and, in the present instance, the stimulus material has to be adapted to the language of the participants. In this sense, multiple-laboratory replications always represent important tests of the generalizability of an effect. When the semantic mismatch effect is used to draw general conclusions about the semantic processing of to-be-ignored speech (Röer et al., 2019), the implicit assumption is that the effect reflects a general aspect of cognitive processing. Such generalizations are just a normal part of interpreting effects, but they can be dangerous simplifications. To date, the effect has only been demonstrated with German stimulus material. Although we have no reason to believe that the effect is limited to the German language, there are effects that seem to be limited to certain languages, such as the contextual diversity effect on serial recall performance which is reliably found with Spanish (Parmentier, Comesãna & Soares, 2017), but not with English words (Guitard, Miller, Neath & Roodenrys, 2019). With reference to the semantic mismatch effect, there is one language-specific peculiarity that may or may not be of relevance in this context. Unlike, for example, in English it is common in German for the semantically crucial verb component to come at the very end of the sentence, so it cannot be completely ruled out that German-speaking participants may be particularly well-practiced at retaining words for a while in order to integrate their meaning with incoming semantic information. In a worst-case scenario, this peculiarity of the German language may seriously compromise any efforts to replicate the semantic

mismatch effect with non-German speaking participants. Thus, the present multiple-laboratory preregistered replication also serves as a test of the generalizability of the effect. To have a good empirical basis for comparison, we examined the effect of semantically unexpected words in four different languages (English, French, German, Swedish). The replication study is based on Experiment 2b of Röer et al. (2019). With the exception of the wordings of the proverbs used as stimulus material, our aim was to keep all other aspects of the experimental design as consistent as possible across languages and laboratories.

Method

Ethics Statement

Research was performed in accordance with the Declaration of Helsinki and received the approval of the ethics committee of the Faculty of Mathematics and Natural Sciences at Heinrich Heine University Düsseldorf. Written informed consent was obtained from all participants prior to participation.

Preregistration Statement

A time-stamped preregistration document was published prior to the start of data collection outlining in detail the method and planned analysis using the format provided by https://aspredicted.org. The preregistration document is available at the project page on OSF under https://osf.io/4r5up/.

There is one minor deviation of the actual study from the preregistered plan. We originally planned to collect at least 60 participants in each language. During the scheduled testing period, demand on the laboratories in England and Sweden was intense. Therefore, we were only able to collect data from 59 participants in English and Swedish laboratories. In all other aspects, we followed the preregistered plan.

Participants

The total sample consisted of 252 participants (180 women; *M* age = 24, *SD* = 7). Participants received course credit or a monetary compensation for participating. All participants reported normal hearing and normal or corrected-to-normal vision. The English language sample was recruited at the University of Central Lancashire in the United Kingdom and consisted of 59 participants (43 women; *M* age = 24, *SD* = 7). The French language sample was recruited at Université de Moncton in Canada and consisted of 60 participants (45 women; *M* age = 20, *SD* = 3). The German language sample was recruited at Heinrich Heine University Düsseldorf in Germany and consisted of 74 participants (53 women; *M* age = 22, *SD* = 4). The Swedish language sample was recruited at the University of Gävle in Sweden and consisted of 59 participants (39 women; *M* age = 29, *SD* = 10). All data were collected in person in 2018 and 2019 (i.e., before the COVID-19 pandemic).

Materials

A standard serial recall task was used with eight to-be-remembered digits that were sampled randomly without replacement from the set {1, 2, ... 9}. Digits were presented at a rate of 1 Hz (800 ms on, 200 ms off) in black font on a white background in the center of the computer screen. From a viewing distance of 50 cm, the to-be-remembered digits subtended a vertical visual angle of 1.34° and a horizontal angle of 0.83°.

Auditory distractors were 24 proverbs. In the training block, 8 proverbs were presented in the expected version. In the subsequent experimental block, 8 proverbs were presented in the expected version and 8 proverbs were presented in the unexpected version. In the expected version, the proverbs ended with the correct final word, for example “A poor workman always blames his tools” and “It's no use crying over spilt milk”. In the unexpected version, the proverbs ended with a final word from a different proverb resulting in a violation of semantic expectations, for example “A poor workman always blames his *milk*” and “It's no use crying over spilt *tools*”. The average number of

syllables and the average number of words was equivalent in the expected and unexpected versions in each language.

Auditory distractors were recorded digitally at 44.1 kHz with 16-bit encoding. To avoid acoustic differences in pronunciation between words as a function of whether they were semantically expected or unexpected, all auditory distractor sentences were recorded using text-to-speech software. In the English sample, the text-to-speech software of Mac OS X 10.14 was used to generate the stimuli. The sentences were spoken by the female voice “Kate” in the English sample. In the French sample, the Google Text-to-Speech software was used to generate the stimuli. The sentences were spoken by the voice “fr-CA-Wavenet-D”. This is a male voice with a French-Canadian accent. In the German sample, the text-to-speech software of Mac OS X 10.11 was used to generate the stimuli. The sentences were spoken by the female voice “Anna”. In the Swedish sample, the text-to-speech software of Mac OS X 10.9.3 was used to generate the stimuli. The sentences were spoken by the female voice “Alva”. The recordings and a list of distractor sentences used in each language are available at the project page on OSF under https://osf.io/4r5up/. During the experiment the distractor sentences were played binaurally at a normal conversational speech level using headphones with high-insulation hearing protection covers.

Procedure

Before starting the experiment, the participants were instructed that all sounds would be task-irrelevant and should be ignored. The written instructions are available at the project page on OSF under https://osf.io/4r5up/.

Participants first completed a training block that consisted of 8 quiet trials and 8 expected trials presented in a random order. The subsequent experimental block consisted of 8 quiet trials, 8 expected trials, and 8 unexpected trials that each participant completed in a different random order. Proverbs were randomly drawn without replacement from the total set of 16 proverbs. Each proverb was only presented once (i.e., either in the expected or in the unexpected condition).

After the presentation of the last to-be-remembered digit, eight question marks appeared on the screen. Participants used the number pad of the keyboard attached to the computer that controlled the experiment to replace the question marks with the digits they still remembered. It was not possible to correct a response, but it was possible to skip over a serial position by pressing a "don't know" button on the keyboard. The experiment took about 16 minutes.

Design

A 4×3×8 repeated measures design was used with language (English, French, German, Swedish) as the between-subjects independent variable, auditory condition (quiet, expected, unexpected) and serial position (1 to 8) as the within-subject independent variables and serial recall performance as the dependent variable.

The semantic mismatch effect refers to the finding that sentences with unexpected endings that do not match the preceding semantic context are more disruptive than sentences ending with expected words that match the preceding semantic context. Thus, the critical test is the comparison between serial recall performance in the expected and unexpected condition. Given a total sample size of $N = 252$, $\alpha = \beta = .05$, and the assumption that the average population correlation between the expected and the unexpected condition is $\rho = .5$ (estimated based on the results of Experiment 2b of Röer et al., 2019), a semantic mismatch effect as small as $f = 0.11$ (i.e., a "small" effect in terms of the conventions suggested by Cohen, 1988) could be detected. In terms of partial eta squared, this corresponds to an effect size of $\eta_p^2 = .05$. Power calculations were conducted using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007).

Results

The data on which the analyses are based are available at the project page on OSF under https://osf.io/4r5up/. Digits recalled in the serial position they were presented in were scored as correct. A multivariate approach was used for all within-subject comparisons. In the present appli-

cation, all multivariate test criteria correspond to the same exact *F* statistic which is reported. The level of α was set to .05. Partial eta squared (η_p^2) is reported as a measure of relative effect size, that is, the variance explained relative to the variance not explained by any of the other experimental variables.

All reported analyses were preregistered prior to the start of the data collection. No data were excluded before or after the analyses. We will report the combined analysis across all languages. Given that in the main analysis the semantic mismatch effect did not differ as a function of language, it was not necessary to perform separate follow-up analyses of the semantic mismatch effect in each language. In all other aspects, the following analyses adhere to the preregistered analysis plan.

Preregistered Confirmatory Analyses

Figure 1 shows recall performance as a function of auditory condition at each serial position. A 4×3×8 repeated-measures MANOVA with language (English, French, German, Swedish) as between-subject variable and auditory condition (quiet, expected, unexpected) as well as serial position (1 to 8) as within-subject variables yielded significant main effects of language, $F(3,248) = 8.19$, $p < .001$, $\eta_p^2 = .09$, auditory condition, $F(2,247) = 161.54$, $p < .001$, $\eta_p^2 = .57$, and serial position, $F(7,242) = 181.51$, $p < .001$, $\eta_p^2 = .84$. There were significant interactions of language and serial position, $F(21,732) = 2.25$, $p = .001$, $\eta_p^2 = .06$, and of auditory condition and serial position, $F(14,235) = 10.13$, $p < .001$, $\eta_p^2 = .38$. Importantly, there was no interaction of language and auditory condition, $F(6,496) = 1.12$, $p = .347$, $\eta_p^2 = .01$, demonstrating that the effect of auditory distraction did not differ among languages. There was also no three-way-interaction, $F(42,711) = 1.08$, $p = .334$, $\eta_p^2 = .06$.

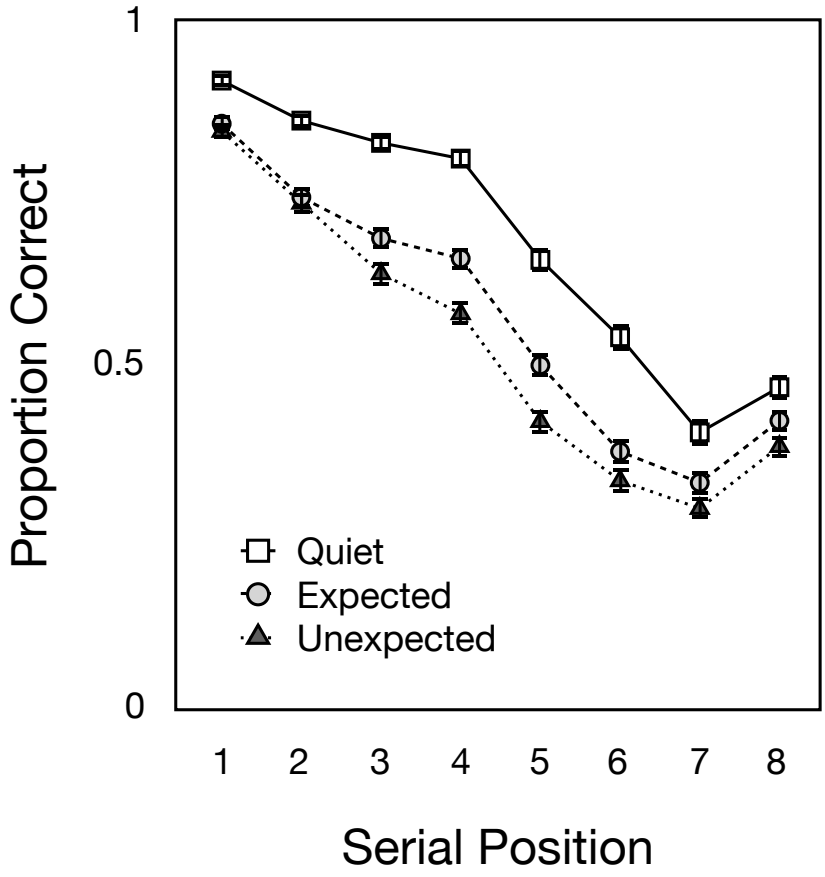

Figure 1. Overall recall performance as a function of auditory condition (quiet, expected, unexpected) for each serial position. The error bars represent the standard errors of the means.

Orthogonal contrasts revealed an irrelevant speech effect in form of a significant reduction of recall performance in the distractor conditions relative to the quiet control condition, $F(1,248) = 300.36$, $p < .001$, $\eta_p^2 = .55$, and a semantic mismatch effect in form of a significant reduction of recall performance in the unexpected condition relative to the expected condition, $F(1,248) = 31.73$, $p < .001$, $\eta_p^2 = .11$. The between-subject variable language did not interact with the irrelevant speech effect, $F(3,248) = 1.58$, $p = .195$, $\eta_p^2 = .02$, and, most importantly, language also did not interact with the semantic mismatch effect, $F(3,248) = 0.78$, $p = .503$, $\eta_p^2 = .01$ (see Figure 2).

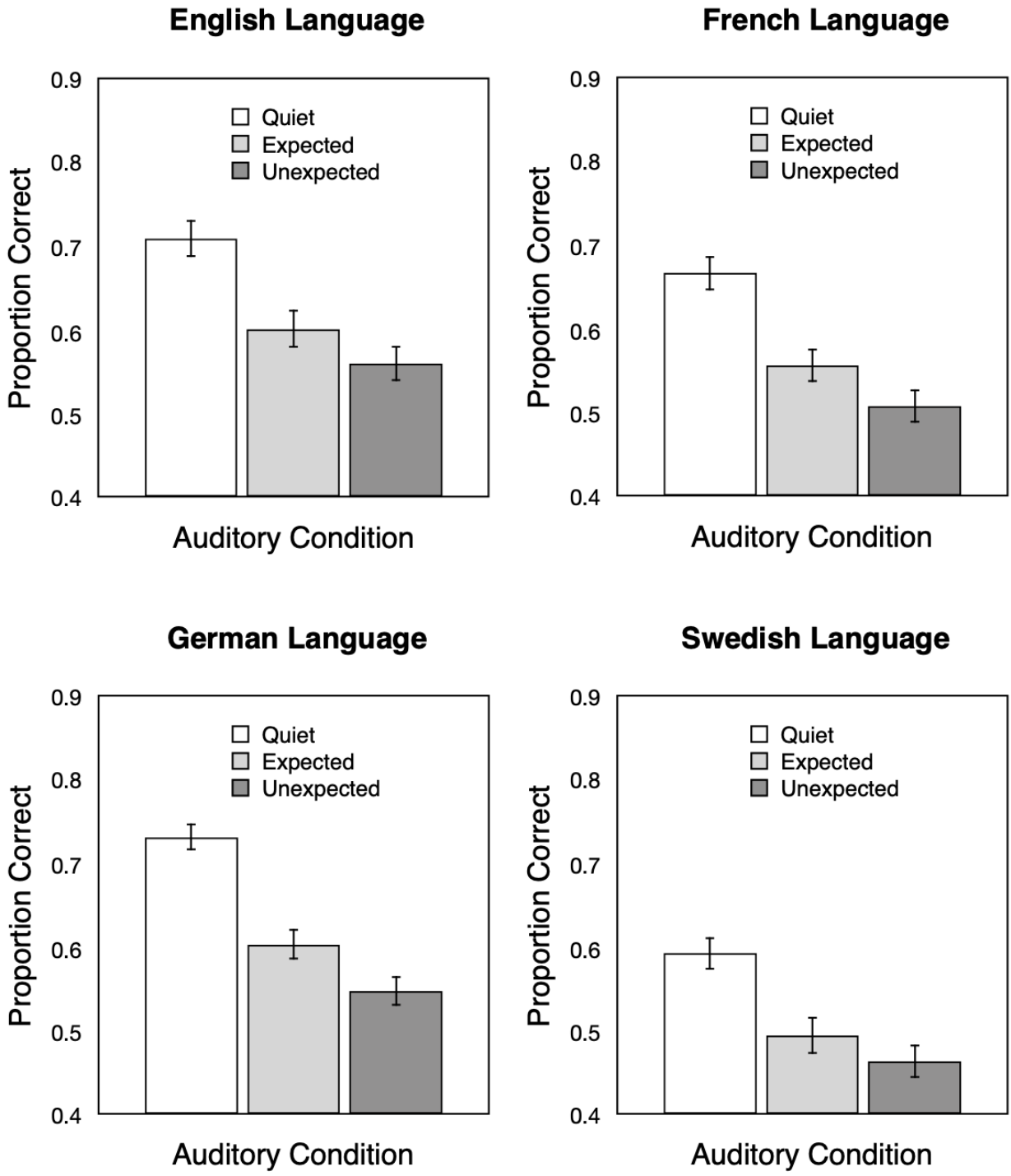

Figure 2. Recall performance as a function of auditory condition (quiet, expected, unexpected) and language (English, French, German, Swedish) collapsed across serial positions. The error bars represent the standard errors of the means.

As specified in the preregistration, we were also interested in whether or not the semantic mismatch effect habituates. Figure 3 shows recall performance as a function of auditory condition at each ordinal trial position. If the disruptive potential of unexpected words is reduced with repeated exposure, then the semantic mismatch should gradually decrease over the course of the experiment. However, a 4×2×8 repeated measures MANOVA with language (English, French, German, Swedish) as between-subject variable and auditory condition (expected, unexpected) as well as ordinal trial position (1 to 8) as within-subject variables showed no evidence for a decrease of the semantic mismatch effect across trials. Specifically, there was no interaction between the linear contrast component of the ordinal trial position variable and the variable contrasting the expected and unexpected condition, $F(1,248) = 2.33$, $p = .128$, $\eta_p^2 = .01$. There was also no three-way interaction with language, $F(3,248) = 0.18$, $p = .907$, $\eta_p^2 < .01$.

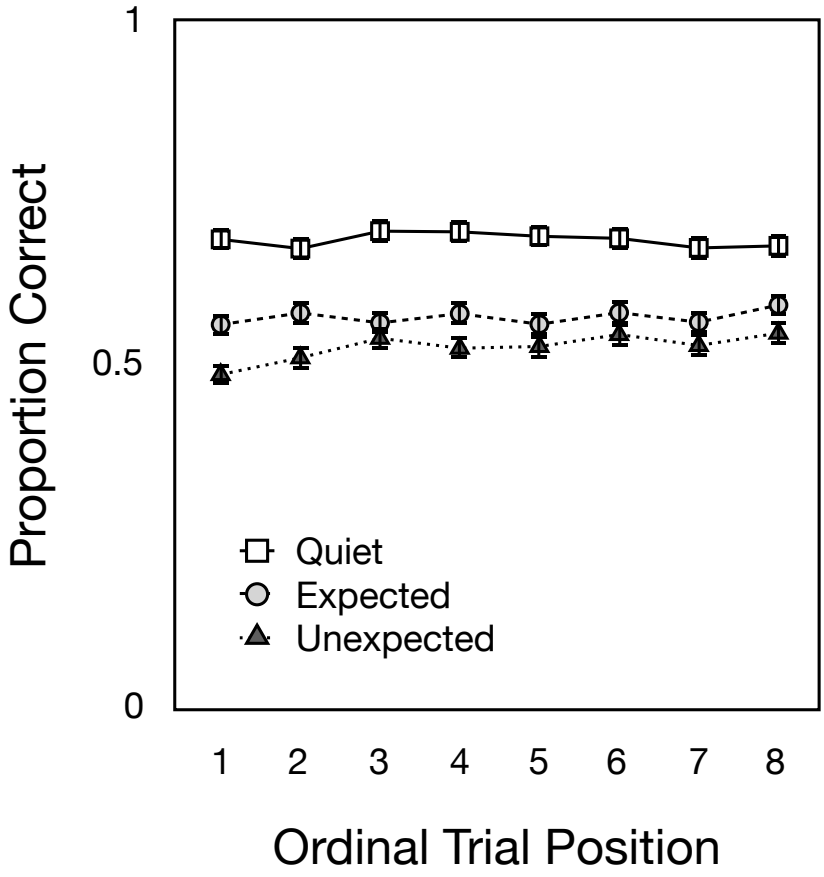

Figure 3. Recall performance as a function of auditory condition (quiet, expected, unexpected) and ordinal trial position with 1 corresponding to the first trial in each condition, 2 corresponding to the second trial, and so on. The error bars represent the standard errors of the means.

Discussion

Across all laboratories, sentences with unexpected endings that did not match the preceding semantic context were more disruptive than sentences with expected endings. The sample effect size of the semantic mismatch effect observed in the present study ($\eta_p^2 = .11$) is comparable with the sample effect size of the semantic mismatch effect ($\eta_p^2 = .13$) in the experiment that the replication was based on (Experiment 2b of Röer et al., 2019). Importantly, there was no interaction with the language variable, demonstrating that the disruptive effect of semantic mismatches did not differ among languages. The multilingual replication of the semantic mismatch effect can therefore be considered successful.

Although there was no interaction of language and auditory condition, there was a significant main effect of language indicating differences in the absolute level of serial recall performance between samples. It can be seen from Figure 2 that the Swedish participants performed more than ten percent poorer on average than participants from the other laboratories. We did not expect this difference and can only speculate about its cause. One aspect that could be relevant in this context is the pronunciation time of the to-be-remembered items which had an effect on memory—at least in some studies. The word length effect refers to the finding that lists of words with a long pronunciation time are more poorly remembered than lists of words with a short pronunciation time (e.g., Baddeley et al., 1975; Cowan et al., 1992; Nairne et al., 1997; but see Caplan et al., 1992; Guitard et al., 2018; Neath et al., 2003). In English and German, there is only one numeral from 1 to 9 that has two syllables (i.e., “seven” and “sieben”, respectively) while all other numerals have only one syllable. In French, depending on the pronunciation of "quatre", there is also one or no numeral with two syllables (New, Pallier, Brysbaert & Ferrand, 2004). In Swedish, by contrast, there are three numerals from 1 to 9 that have two syllables (i.e., “fyra”, “åtta”, and “nio”). Thus, the word length effect may serve as a post-hoc explanation of why overall performance was poorer in the Swedish sample than in the other samples. However, in the present study we were mainly interested in the effects of auditory distraction and, despite the absolute differences in serial recall, the relative differences among the distractor conditions did not differ as a function of language. The main conclusion of the present study therefore is that the semantic mismatch effect on serial recall can be consistently reproduced in different languages.

A further interesting observation is that, in the present replication, the size of the semantic mismatch effect did not decline over the course of the experiment. Note that, as in the original study, a different distractor sentence was presented in each trial. The finding that performance did not benefit from the repeated exposure to different semantic mismatches is consistent with previous studies (Röer et al., 2019; Röer et al., 2020) in which there was no evidence for unspecific habituation of the semantic mismatch effect. It also fits well with what has been observed with the categorical deviant effect (Vachon et al., 2020). Sequences that contained a single deviant item from a

different category still disrupted serial recall more than sequences without such a deviation when unspecific foreknowledge was given (i.e., participants were informed whether the upcoming distractor sequence will contain a deviant item, or not) and even specific foreknowledge (i.e., participants were informed about the identity of the deviant item) did not reduce the categorical deviant effect. This may indicate that effects that require the integration of incoming information into a larger context of meaning such as the semantic mismatch effect and the categorical deviant effect (i.e., inter-lexical effects) cannot be overcome as easily as effects that concern only the individual word meaning (i.e., lexical effects), for example one's own name (Röer et al., 2013). From a practical standpoint, this feature of the semantic mismatch effect seems attractive as it allows to examine the effect across a large number of trials.

What is obvious from the semantic mismatch effect is that the meaning of the task-irrelevant sentence preceding the mismatch is processed and that the presence (and processing of) the mismatch produces disruption. This finding is inconsistent with the notion that the semantic properties of sound are filtered out at early processing stages (Broadbent, 1958, 1975; Treisman, 1964; 1969) or that they only affect tasks that require the processing of meaning such as text comprehension. The finding that words without any inherent attention-grabbing properties disrupt serial recall when they do not match the preceding semantic context sheds light on the fate of to-be-ignored speech and can help to refine theories of auditory distraction. Theories that specify a role for attention (Bell et al., 2019; Cowan, 1999; Hughes et al., 2013) appear to be suitable to explain why semantic mismatches disrupt serial recall. A theoretical account of the semantic mismatch effect has to start from the assumption that irrelevant speech is processed semantically to some degree, otherwise semantic mismatches or categorical deviants would not be detected (see also Röer et al., 2019; Vachon et al., 2020). The effect suggests that the model of the auditory environment does not only include acoustic but also semantic features and that—in case of a mismatch between the predicted and the incoming stimulation—the processing of a semantically unexpected word disrupts the maintenance of information in short-term memory. Now that the phenomenon has been established across different laboratories and languages, more fine-grained theories

about the underlying processes can be developed, and future studies may establish the similarities and differences between the newly discovered phenomenon and more established phenomena such as the auditory deviant effect. The multilingual stimulus set that has been developed here (and is made available to other researchers) will hopefully prove useful to achieve these goals.

In sum, the present results confirm the semantic mismatch effect in a preregistered multiple-language, multiple-laboratory replication. Reproducibility across laboratories is seen as an important step in establishing trust in a newly discovered phenomenon (e.g., Hüffmeier et al., 2016; Simons, 2014). The fact that the semantic mismatch effect can be obtained in different languages and independently of the specific circumstances prevailing in individual laboratories shows that the effect does not depend on highly specific boundary conditions and thus suggests that the effect may reflect a general property of the processing of to-be-ignored auditory information. Now that the effect has been established in different languages, theories on attention and distraction should be adapted to incorporate the effect. The existence of the effect suggests that there must be a process that analyzes the content of irrelevant speech in the to-be-ignored channel even at an inter-lexical level, derives predictions about its continuation, and reacts with more intense processing when unexpected content is encountered, the latter of which leads to a reduction in the efficacy of other cognitive processes and thus to a decline in ongoing cognitive performance. Any explanation of the semantic mismatch effect should also account for the persistence of the effect. The disruptive potential of unexpected words is not reduced with repeated exposure, suggesting that incoming information is routinely integrated into a meaningful context, in the course of which mismatches inevitably lead to increased processing.

References

Baddeley, A. D., Thomson, N., & Buchanan, M. (1975). Word length and the structure of short-term memory. *Journal of Verbal Learning & Verbal Behavior*, *14*, 575-589.

Beaman, C. P., & Jones, D. M. (1997). Role of serial order in the irrelevant speech effect: Tests of the changing-state hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 459-471.

Bell, R., Buchner, A., & Mund, I. (2008). Age-related differences in irrelevant-speech effects. *Psychology and Aging*, *23*, 377-391.

Bell, R., Mieth, L., Röer, J. P., Troche, S. J., & Buchner, A. (2019). Preregistered replication of the auditory deviant effect: A robust benchmark finding. *Journal of Cognition*, *2*, 13.

Bell, R., Mund, I., & Buchner, A. (2011). Disruption of short-term memory by distractor speech: Does content matter? *Quarterly Journal of Experimental Psychology, 64*, 146-168.

Bell, R., Röer, J. P., Lang, A. G., & Buchner, A. (2019a). Distraction by steady-state sounds: Evidence for a graded attentional model of auditory distraction. *Journal of Experimental Psychology: Human Perception and Performance*, *45*, 500-512.

Bell, R., Röer, J. P., Lang, A. G., & Buchner, A. (2019b). Reassessing the token set size effect on serial recall: Implications for theories of auditory distraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 1432-1440.

Bell, R., Röer, J. P., Marsh, J. E., Storch, D., & Buchner, A. (2017). The effect of cognitive control on different types of auditory distraction. *Experimental Psychology*, *64*, 359-368.

Broadbent, D. E. (1957). A mechanical model for human attention and immediate memory. *Psychological Review*, *64*, 205-215.

Broadbent D. E. (1958). *Perception and Communication.* New York: Pergamon Press.

Buchner, A., Irmen, L., & Erdfelder, E. (1996). On the irrelevance of semantic information for the "Irrelevant Speech" effect. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *49A*, 765-779.

Buchner, A., Mehl, B., Rothermund, K., & Wentura, D. (2006). Artificially induced valence of distractor words increases the effects of irrelevant speech on serial recall. *Memory & Cognition*, *34*, 1055-1062.

Buchner, A., Rothermund, K., Wentura, D., & Mehl, B. (2004). Valence of distractor words increases the effects of irrelevant speech on serial recall. *Memory & Cognition*, *32*, 722-731.

Caplan, D., Rochon, E., & Waters, G. S. (1992). Articulatory and phonological determinants of word length effects in span tasks. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *45*, 177–192.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (Rev. ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Colle, H. A., & Welsh, A. (1976). Acoustic masking in primary memory. *Journal of Verbal Learning & Verbal Behavior*, *15*, 17-31.

Cowan, N., Day, L., Saults, J. S., Keller, T. A., Johnson, T., & Flores, L. (1992). The role of verbal output time in the effects of word length on immediate memory. *Journal of Memory and Language*, *31*, 1–17.

Cowan, N. (1999). An Embedded-Processes Model of working memory. In A. Miyake & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 62-101). New York: Cambridge University Press.

Deutsch, J. A., & Deutsch, D. (1963). Attention: Some theoretical considerations. *Psychological Review*, *70*, 80-90.

Duncan, J. (1980). The locus of interference in the perception of simultaneous stimuli. *Psychological Review*, *87*, 272-300.

Ellermeier, W., Kattner, F., Ueda, K., Duomoto, K., & Nakajima, Y. (2015). Memory disruption by irrelevant noise-vocoded speech: Effects of native language and the number of frequency bands. *The Jounal of the Acoustical Society of America*, *138*, 1561-1569.

Ellermeier, W., & Zimmer, K. (1997). Individual differences in susceptibility to the "irrelevant speech effect". *The Journal of the Acoustical Society of America*, *102*, 2191-2199.

Ellermeier, W., & Zimmer, K. (2014). The psychoacoustics of the irrelevant sound effect. *Acoustical Science and Technology*, *35*, 10-16.

Elliott, E. M. (2002). The irrelevant-speech effect and children: theoretical implications of developmental change. *Memory & Cognition*, *30*, 478-87.

Eriksen, C. W., & St. James, J. D. (1986). Visual attention within and around the field of focal attention: A zoom lens model. *Perception & Psychophysics*, *40*, 225-240.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175-191.

Guitard, D., Gabel, A. J., Saint-Aubin, J., Surprenant, A. M., & Neath, I. (2018). Word length, set size, and lexical factors: Re-examining what causes the word length effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*, 1824–1844.

Guitard, D., Miller, L. M., Neath, N., & Roodenrys, S. (2019). Does contextual diversity affect serial recall? *Journal of Cognitive Psychology*, 31, 379-396.

Hughes, R. W. (2014). Auditory distraction: A duplex-mechanism account. *PsyCH Journal*, *3*, 30-41.

Hüffmeier, J., Mazei, J., & Schultze, T. (2016). Reconceptualizing replication as a sequence of different studies: A replication typology. *Journal of Experimental Social Psychology*, *66*, 81-92.

Hughes, R. W., Hurlstone, M. J., Marsh, J. E., Vachon, F., & Jones, D. M. (2013). Cognitive control of auditory distraction: Impact of task difficulty, foreknowledge, and working memory capacity supports duplex-mechanism account. *Journal of Experimental Psychology: Human Perception and Performance*, *39*, 539-553.

Hughes, R. W., Vachon, F., & Jones, D. M. (2005). Auditory attentional capture during serial recall: Violations at encoding of an algorithm-based neural model? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 736-749.

Hughes, R. W., Vachon, F., & Jones, D. M. (2007). Disruption of short-term memory by changing and deviant sounds: Support for a duplex-mechanism account of auditory distraction. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 1050-1061.

Jones, D. M., & Macken, W. J. (1993). Irrelevant tones produce an irrelevant speech effect: Implications for phonological coding in working memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 369-381.

Jones, D. M., Miles, C., & Page, J. (1990). Disruption of proofreading by irrelevant speech: Effects of attention, arousal or memory? *Applied Cognitive Psychology*, *4*, 89-108.

Jones, D. M., Saint-Aubin, J., & Tremblay, S. (1999). Modulation of the irrelevant sound effect by organizational factors: Further evidence from streaming by location. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *52A*, 545-554.

Körner, U., Röer, J. P., Buchner, A., & Bell, R. (2017). Working memory capacity is equally unrelated to auditory distraction by changing-state and deviant sounds. *Journal of Memory & Language*, *96*, 122-137.

Körner, U., Röer, J. P., Buchner, A., & Bell, R. (2019). Time of presentation affects auditory distraction: Changing-state and deviant sounds disrupt similar working memory processes. *Quarterly Journal of Experimental Psychology*, *72*, 457-471.

LeCompte, D. C., Neely, C. B., & Wilson, J. R. (1997). Irrelevant speech and irrelevant tones: The relative importance of speech to the irrelevant speech effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*, 472-483.

Longoni, A. M., Richardson, J. T., & Aiello, A. (1993). Articulatory rehearsal and phonological storage in working memory. *Memory & Cognition*, *21*, 11-22.

Lovatt, P., Avons, S. E., & Masterson, J. (2000). The word-length effect and disyllabic words. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *53*, 1–22.

Marsh, J. E., & Jones, D. M. (2010). Cross modal distraction by back- ground speech: What role for meaning? *Noise & Health, 12*, 210-216.

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2008). Auditory distraction in semantic memory: A process-based approach. *Journal of Memory and Language, 58*, 682–700.

Marsh, J. E., Hughes, R. W., & Jones, D. M. (2009). Interference by process, not content, determines semantic auditory distraction. *Cognition*, *110*, 23-38.

Marsh, J. E., Röer, J. P., Bell, R., & Buchner, A. (2014). Predictability and distraction: Does the neural model represent post-categorical features? *PsyCH Journal*, *3*, 58-71.

Marsh, J. E., Yang, J., Qualter, P., Richardson, C., Perham, N., Vachon, F., & Hughes, R. W. (2018). Postcategorical auditory distraction in short-term memory: Insights from increased task load and task type. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*, 882-897.

Nairne, J. S., Neath, I., & Serra, M. (1997). Proactive interference plays a role in the word-length effect. *Psychonomic Bulletin & Review*, *4*, 541-545.

Neath, I., Bireta, T. J., & Surprenant, A. M. (2003). The time-based word length effect and stimulus set specificity. *Psychonomic Bulletin & Review*, *10*, 430-434.

Neely, C. B., & LeCompte, D. C. (1999). The importance of semantic similarity to the irrelevant speech effect. *Memory & Cognition, 27*, 37–44.

New, B., Pallier, C., Brysbaert, M., & Ferrand, L. (2004). Lexique 2: A new French lexical database. *Behavior Research Methods, Instruments & Computers, 36,* 516–524.

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*, 2600-2606.

Oberauer, K., Lewandowsky, S., Awh, E., Brown, G. D. A., Conway, A. R. A., Cowan, N., Donkin, C., Farrell, S., Hitch, G. J., Hurlstone, M. J., Ma, W. J., Morey, C. C., Nee, D. E., Schweppe, J., Vergauwe, E., & Ward, G. (2018). Benchmarks for Models of Short Term and Working Memory. *Psychological Bulletin*, *144*, 885–958.

Parmentier, F. B. R., & Beaman, C. P. (2015). Contrasting effects of changing rhythm and content on auditory distraction in immediate memory. *Canadian Journal of Experimental Psychology*, *69*, 28-38.

Parmentier, F. B. R., Comesãna, M., & Soares, A. P. (2017). Disentangling the effects of word frequency and contextual diversity on serial recall performance, *Quarterly Journal of Experimental Psychology, 70*, 1–17.

Röer, J. P., Bell, R., & Buchner, A. (2013). Self-relevance increases the irrelevant speech effect: Attentional disruption by one's own name. *Journal of Cognitive Psychology*, *25*, 925-931.

Röer, J. P., Bell, R., & Buchner, A. (2014). Evidence for habituation of the irrelevant sound effect on serial recall. *Memory & Cognition*, *42*, 609-621.

Röer, J. P., Bell, R., Körner, U., & Buchner, A. (2019). A semantic mismatch effect on serial recall: Evidence for interlexical processing of irrelevant speech. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*, 515-525.

Röer, J. P., Bell, R., Marsh, J. E., & Buchner, A. (2015). Age equivalence in auditory distraction by changing and deviant speech sounds. *Psychology and Aging*, *30*, 849-855.

Röer, J. P., Buchner, A., & Bell, R. (2020). Auditory distraction in short-term memory: Stable effects of semantic mismatches on serial recall. *Auditory Perception & Cognition*, 143-162.

Röer, J. P., Körner, U., Buchner, A., & Bell, R. (2017a). Attentional capture by taboo words: A functional view of auditory distraction. *Emotion*, *17*, 740-750.

Röer, J. P., Körner, U., Buchner, A., & Bell, R. (2017b). Semantic priming by irrelevant speech. *Psychonomic Bulletin & Review*.

Service, E. (1998). The effect of word length on immediate serial recall depends on phonological complexity, not articulatory duration. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *51*, 283–304.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*, 76-80.

Surprenant, A. M., Neath, I., & Bireta, T. J. (2007). Changing state and the irrelevant sound effect. *Canadian Acoustics*, *35*, 86-87.

Treisman, A. M. (1960). Contextual cues in selective listening. *Quarterly Journal of Experimental Psychology*, *12*, 242-248.

Treisman, A. M. (1964). Verbal cues, language, and meaning in selective attention. *American Journal of Psychology*, *77*, 206-219.

Treisman, A. M. (1969). Strategies and models of selective attention. *Psychological Review*, *76*, 282-299.

Ueda, K., Nakajima, Y., Ellermeier, W., & Kattner, F. (2017). Intelligibility of locally time-reversed speech: A multilingual comparison. *Scientific Reports*, *7*, 1782.

Vachon, F., Labonté, K., & Marsh, J. E. (2017). Attentional capture by deviant sounds: A noncontingent form of auditory distraction? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 622-634.

Vachon, F., Marsh, J. E., & Labonté, K. (2020). The automaticity of semantic processing revisited: Auditory distraction by a categorical deviation. *Journal of Experimental Psychology: General*, *149*, 1360-1397.

Declarations

Funding

The research reported in this article was supported by a grant from the Deutsche Forschungsgemeinschaft (RO 4972/1-1) to JPR and a Discovery grant from the Natural Sciences and Engineering Research Council of Canada (RGPIN-2015-04416) to JS-A.

Conflicts of interest/Competing interests

All authors state that no conflicts of interest/competing interests exist.

Availability of data and materials

The preregistration document, stimuli, instructions, data, and a data dictionary can be found on OSF under https://osf.io/4r5up/.

Authors' contributions

Contributed to conception and design: JPR, RB, AB, JS-A, JEM

Contributed to acquisition of data: JPR, RB, AB, JS-A, R-PS, JEM, SBM, MBAK, RL, SA

Contributed to analysis and interpretation of the data: JPR, RB, AB, JS-A, R-PS, JEM, SBM, MBAK, RL, SA

Drafted and/or revised the article: JPR, RB, AB, JS-A, R-PS, JEM, SBM, MBAK, RL, SA

Approved the submitted version for publication: JPR, RB, AB, JS-A, R-PS, JEM, SBM, MBAK, RL, SA