Memory

Routledge
Taylor & Francis Group

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/pmem20

Characterizing production: the production effect is eliminated for unusual voices unless they are frequent at study

Rachelle M. Wakeham-Lewis, Jason Ozubko & Jonathan M. Fawcett

To cite this article: Rachelle M. Wakeham-Lewis, Jason Ozubko & Jonathan M. Fawcett (2022): Characterizing production: the production effect is eliminated for unusual voices unless they are frequent at study, Memory, DOI: 10.1080/09658211.2022.2115075

To link to this article: https://doi.org/10.1080/09658211.2022.2115075

View supplementary material

Published online: 15 Sep 2022.

Submit your article to this journal

View related articles

View Crossmark data

MEMORY
https://doi.org/10.1080/09658211.2022.2115075

Characterizing production: the production effect is eliminated for unusual voices unless they are frequent at study

Rachelle M. Wakeham-Lewis[a], Jason Ozubko [b] and Jonathan M. Fawcett[a]

[a]Department of Psychology, Memorial University of Newfoundland, St. John's, Newfoundland, Canada; [b]Department of Psychology, SUNY Geneseo, Geneseo, NY, USA

ABSTRACT

The production effect refers to the finding that items read aloud are better remembered than items read silently. This is often explained with reference to distinctiveness, arguing that aloud items become associated with distinctive sensorimotor features that facilitate retrieval at test. Based on this framework, more distinctive forms of production should result in larger production effects. The present study tested this theory by having participants study items silently or aloud in either their own voice or as a popular character. Participants were then tested for those items using recognition memory. Relative to silent items, aloud items read in the participants' own voice demonstrated a typical production effect; however, contrary to any predictions, no production effect was observed for the character voices. We next manipulated how frequently the character voice was used relative to the participants' own voice. This revealed a production effect for character voices only when those voices were more common than the participant's own voice. This pattern could not be attributed to cognitive demands or performance anxiety but was predicted by a novel computational account based on the Retrieving Effectively from Memory (REM) model. Our results show that the relation between distinctiveness and memory is not necessarily linear.

ARTICLE HISTORY
Received 27 June 2021
Accepted 14 August 2022

KEYWORDS
Production effect; memory; distinctiveness; recognition memory

It has long been understood that reading something aloud makes it more memorable than reading it silently. Hopkins and Edwards (1972) first reported this phenomenon after discovering that participants recalled more words in a standard list-learning paradigm if they had pronounced those words aloud at study rather than reading them silently. Conway and Gathercole (1987) reported similar findings and further found that even just mouthing a word made it more memorable than silent reading – although words read aloud were still more memorable than those mouthed. In more recent years, interest in this finding has been reinvigorated with its rebranding as the production effect by MacLeod et al. (2010) and its extension to other forms of production, such as writing (Forrin et al., 2012) or singing (Quinlan & Taylor, 2013).

Many modern theorists have attributed the advantage associated with production to distinctive encoding processes at study that are then leveraged to facilitate recognition or recall at test: This has been referred to as the *distinctiveness account* (MacLeod et al., 2010; Ozubko & MacLeod, 2010; Forrin et al., 2012). According to this account, producing something enhances its distinctiveness at encoding by appending a unique record of the productive act to the representation in memory not available for non-produced items. This production record (sometimes called the production trace; Fawcett, 2013) can then be used diagnostically at test to discriminate between studied items and unstudied items based on whether participants recall having produced them, thereby facilitating retrieval (for example, participants may think to themselves, "I remember saying this word out loud therefore this word must be old"; MacLeod et al., 2010). This strategic use of the production record is sometimes referred to as the *distinctiveness heuristic* (Dodson & Schacter, 2001; Taikh & Bodner, 2016).

In support of this general framework, Ozubko and MacLeod (2010) found that removing the utility of the production record as a discriminative indicator eliminated the production effect. In their study, participants memorised two separate wordlists – a critical mixed list in which half the words were to be read aloud, and half read silently, and a distracting pure list in which all words were either read aloud or silently. At test, participants were re-presented with each word and asked to identify its originating list. Their results demonstrated a production effect for the critical mixed list only when the words on the distracting

CONTACT Jonathan M. Fawcett jfawcett@mun.ca Department of Psychology, Memorial University of Newfoundland, St. John's, Newfoundland A1B 3X9, Canada

Supplemental data for this article can be accessed online at https://doi.org/10.1080/09658211.2022.2115075.

pure list were read silently. Thus, when all words on the distracting list were read aloud, participants were no longer able to use the production record to discriminate list membership (but see Bodner & Taikh, 2012). They concluded that the production effect is supported by the relative distinctiveness conferred by the production record.

This argument for distinctiveness is supported further by Forrin et al. (2012), who examined a broader variety of production modalities. Their results showed that the magnitude of words remembered increased according to the relative distinctiveness of the encoding modality – words read aloud were better remembered than words either mouthed, written, or whispered, which were subsequently better remembered than words read silently. Thus, their results suggest that more distinctive production modalities confer greater mnemonic benefits (see also, Fawcett et al., 2012). Subsequent work by Quinlan and Taylor (2013) extended this finding to singing, which was likewise found superior to either reading silently or aloud. In line with earlier theoretical explanations, this was attributed to the fact that the production record for items that were sung now included further elements such as intensity, pitch, and/or timbre (but see, Hassall et al., 2016).

Although it appears that more distinctive forms of production offer additional memory benefits, few studies have explored variation within a specific production modality (e.g., saying a word, but in a different voice). Thus, the goal of this study was to determine whether enhancing distinctiveness by changing how items are voiced would offer similar benefits to more distinctive forms of production, such as singing. To examine this, we used a paradigm in which, in addition to a silent condition, there were two possible production conditions: saying a word out loud in one's own voice and saying a word out loud in an unusual voice. Our initial experiment used three randomly interspersed voices reflecting popular characters or figures (Elvis, Dracula, or Kermit the Frog); our second experiment used only a single voice (Elvis). According to the distinctiveness account, we predicted that saying a word using an unusual voice should incorporate unique elements into the production record that would then be useful at test. Supporting this general idea, voice impersonations have been shown to activate unique brain regions not typically involved in speech (specifically the left anterior insula and inferior frontal gyrus), thus making this distinct from using one's own voice (McGettigan et al., 2013). For that reason, we expected a greater production effect for our voice conditions than for our aloud condition.

Experiment 1

For our initial investigation, we used a modified production task wherein participants were instructed via a visual pre-cue to read the following word either silently, aloud in their own voice or aloud in one of three possible character voices (Elvis, Dracula, or Kermit the Frog). Once all items had been studied, participants were then tested for those items using a recognition task.

Method

Participants

Our minimum target sample size was twenty-four (based on typical sample sizes in this area), although our stopping rule allowed for the possibility of gathering more participants if the term permitted. In total, thirty-one participants (17 female; 30 right-handed; mean age = 23.9, SD = 7.6) were recruited using advertisements placed around the Memorial University of Newfoundland campus or online via the departmental participant pool. Those registered in an eligible undergraduate psychology course were given partial credit towards their grade in exchange for their participation whereas others were provided a small honorarium ($10 per hour). Of those participants, one was excluded because they did not feel comfortable reading aloud (they instead studied *all* words silently) and five further were excluded on the basis that they did not feel comfortable reading aloud in a character's voice (they were instead permitted to study voice words as though they were aloud words). We were required to run these participants (despite their explicit and stated non-compliance) due to a local requirement ensuring students be offered the educational experience of taking part in research regardless of whether they wish to contribute data or take part fully in the task; their results were never analyzed, and they are mentioned here only for the sake of transparency as they took part only as observers. As a result, we had a total usable sample of twenty-five participants (13 female; 23 right-handed; mean age = 24.5, SD = 8.5).

Stimuli and apparatus

PsychoPy2 v1.84.2 (Peirce et al., 2019) was used to run the experiment loaded on a Mac mini running MacOS Sierra, version 10.12.3 with a 23-inch (1080p resolution) Dell P2317H display. Responses were recorded via a standard USB keyboard. All text was in white Arial font set to a normalised unit of 0.1 within the stimulus presentation programme and depicted against a uniform grey (RGB: 160, 160, 160) background. Two instruction images were presented during the practice and study phases: one portrayed the black outline of an eye against a white circle and black background and the other portrayed a mouth matching the same background.

A hired voice actor completed an audio recording demonstrating each of the character voices (Elvis, Dracula, and Kermit the Frog), which participants were required to impersonate throughout the experiment.[1] The audio recording consisted of the voice actor introducing himself as the character followed by a read through of a small word list to demonstrate to participants what that character's voice should sound like. The recording was a wave file and lasted approximately 12 s. These files were recorded using Reaper Digital Audio Workstation software and a Shure SM57 mic with the signal being transmitted through an XLR cable into a Focusrite Scarlett 18i20 USB interface. Participants listened to these recordings through a set of generic headphones connected via the headphone jack.

A pool of 120 monosyllabic nouns were selected from those provided by Tillotson et al. (2008). The mean body–object-interaction for the words in the Tillotson et al. (2008) norms was 3.75 (*SD* = 1.59). The mean imageability for the words in the Cortese and Fugett (2004) norms was 5.19 (*SD* = 1.30). The mean frequency for the words in the SUBTLEX$_{\text{US}}$ corpus (Brysbaert & New, 2009) was 94.48 (*SD* = 168.21). A custom Python script was used to randomly assign 30 words to each of the silent, aloud and voice conditions – with 10 words assigned to each unique voice. The remaining 30 words were used as foils. Words were assigned randomly on a subject-by-subject basis.

It is worth noting that this assignment protocol results in 66% of the words (60 of 90) being read aloud at study, either in one's own voice or in a character voice. Although the relative proportion of aloud and silent items at test has been found to influence the magnitude of the production effect (Icht & Algom, 2014), we were unconcerned as this proportion is typical of three-condition production experiments (e.g., Ozubko et al., 2020; Quinlan & Taylor, 2013), which have observed production effects (as well as differences between production modalities) without issue. Likewise, we opted to use relatively fewer foils at test (25% rather than 50% of all test phase trials) to reduce the duration of our task such that it fit within the testing period available to us, inclusive of all task procedures. Importantly, studies using fewer foils than target items at test have consistently observed a production effect in the past (e.g., Ozubko et al., 2012; Forrin et al., 2012; Ozubko et al., 2014), and there is no reason to expect this feature to interact with our voice manipulation.

Procedure

Participants were instructed that they would be presented with a list of words, one at a time, in the centre of the screen. Preceding each word there was an image instructing them to either read the word silently (an eye) or aloud (a mouth). For aloud trials, the voice they were to use was provided beneath the image (either "Elvis", "Dracula", "Kermit the Frog" or "Yourself"). The silent instruction was always accompanied by "Yourself". Participants were also told they should try to remember the words because they would be tested on them in the second half of the experiment. Prior to the start of the study phase, participants were told they would first listen to an audio demonstration of the voices and then complete a practice task, so they could get used to the instructions. Each phase is detailed below. As an exploratory manipulation aimed at evaluating performance anxiety, we varied whether the research assistant remained in the room with the participant during the experiment or sat in a separate room out of sight (although, unbeknownst to the participant, still within earshot). In total, 12 participants were run with the researcher in the room and 12 with the researcher outside the room (assigned at random), with the position of the researcher having not been recorded for the remaining participant.

Audio demonstration

Before beginning the study phase, participants listened to an audio demonstration of each voice to familiarise them with the impersonation component of the task. The demonstration was approximately 12 s and consisted of each name displayed in the centre of the screen as the recording played.

Practice phase

Once the audio demonstration finished, participants completed a series of practice trials. These trials were identical to the study phase (described below) with the exception that different words were used, which did not appear in the final test. Both production instructions (aloud, voice) were presented 9 times each (3 for each voice), while the read silently instruction was presented 3 times in a randomised loop, totalling 21 practice trials. Participants were told they would not be tested for these items but should treat the practice trials like the actual task. If participants had difficulty with the voices, or appeared to lack confidence in their impersonations, they were run through the audio demonstration and practice phase again (no record was retained as to how often this occurred).

Study phase

Once participants completed the second practice phase, they were instructed that they were now going to begin the actual experiment; they were informed that this would be the same as what they had been doing during the practice task, but that they should now remember the words as they would be tested for them later. Each trial began with a fixation ("+") in the centre of the screen for 500 ms followed by the instruction image (silent, aloud, voice) for 1000 ms, then a blank screen for 500 ms, and finally the word for 1000 ms. The length of each study trial summated to 3000 ms from the start of the presentation of the fixation to the offset of the word presented. In total there were 90 trials consisting of 30 silent items, 30 aloud items and 30 voice items.

Test phase

Once they had finished the study phase, participants completed a recognition task. Participants were told that they would be presented with a series of words consisting of "old" words from the list they had just studied as well as "new" words that were not on that list. Participants had to determine whether each word presented to them was "old" or "new" using a rating scale ranging from 1 to 6. The scale was used as a confidence rating such that 1 meant they were very sure the item was "new", 2 meant they were somewhat sure the item was "new", 3 meant they were unsure the item was "new", 4 meant they were unsure the item was "old", 5 meant they were somewhat sure the item was "old" and 6 meant they were very sure the item was "old". Participants were asked to allocate their responses such that they used each response value throughout the test phase. Each trial consisted of a fixation for 500 ms followed by one of the words (drawn randomly from the list of 30 foil and 90 studied words) presented in the centre of the screen and the confidence rating presented just below it. The confidence rating was depicted as a line with 6 notches. Underneath the first notch were the words "Very Sure New" and underneath the sixth notch were the words "Very Sure Old". The word and confidence rating remained on the screen until the participant submitted their response (using the number keys) which was then followed by the next trial.

Post-experimental questionnaire

Following completion of the computerised portion of the task, participants completed a demographic questionnaire and were asked to rate how self-conscious they felt while producing the voices on a scale from 1 (not self-conscious at all) to 10 (very self-conscious). This value was recorded along with a secret rating (ranging from 1 to 6) as to how well the researcher felt the participant did in differentiating the character voices from their own voice. Included in the demographic questionnaire was also the Generalized Anxiety Disorder 7-item (GAD-7) Scale (Spitzer et al., 2006) to measure general anxiety levels.[2]

Results and discussion

Table 1 depicts the empirical mean percentage hits and false alarms corresponding to each condition (based on a dichotomisation of the confidence data). However, our primary analyses instead used multi-level Bayesian probit regression to estimate *d′* (for a detailed explanation of the methodological and philosophical motivation behind our decision to analyze the data in this manner, please refer to Fawcett et al., 2016 or Fawcett & Ozubko, 2016). We fit our models using the *brms* (Bürkner, 2017, 2018) package within *R 3.6.1* (R Core Team, 2016) following the general fitting and model checking procedure described by Fawcett et al. (2016). Using this approach, we estimated *d′* for each condition following the procedure described by Fawcett and Ozubko (2016). Due to the shared false alarm rate, it was not meaningful to calculate separate estimates of response bias.

Table 1. Percentage "old" responses as a function of item type (silent, aloud, voice, foil) for each experiment and experimental group; standard errors provided in parentheses.

	n	Silent	Aloud	Voice	Foil
Experiment 1	25	53.5 (3.5)	69.5 (3.0)	53.5 (2.8)	17.5 (2.0)
Experiment 2					
High self	29	61.8 (2.8)	68.7 (2.7)	60.9 (3.4)	22.0 (1.9)
High voice	31	60.1 (2.7)	71.6 (3.0)	68.4 (2.1)	20.0 (2.2)

As depicted in the left panel of Figure 1, we observed a typical production effect with greater sensitivity to aloud items than silent items, difference = 0.45, $CI_{95\%}$ [0.21, 0.68]. However, counter to our expectations, performance in the voice condition not only failed to demonstrate a superior production effect – performance in that condition was numerically identical to performance in the silent condition, difference = 0.01, $CI_{95\%}$ [−0.20, 0.22], difference in production effects = 0.44, $CI_{95\%}$ [0.26, 0.62]. Given the robust nature of the production effect, this was surprising. In short, producing a word in an unusual voice appears to undermine – rather than augment – the production effect. This finding is unexpected based on all extant theoretical accounts of the production effect, none of which would anticipate *a priori* that reading aloud in an unusual voice would eliminate the effect. There are several possible explanations for the absence of a production effect within the voice condition, including performance anxiety, the cognitive demands of assuming a character, task-switching costs, or contextual effects at test.

With respect to performance anxiety, we anticipated this possibility and incorporated features into our design intended to mitigate or quantify its influence on our task. These measures included (a) self-consciousness ratings gathered from the subjects and experimenters; and (b) manipulation of the research assistant's location during the task (i.e., inside or outside the room). We were unable to analyze the data as a function of the self-consciousness ratings because these ratings were lost for the first half of the participants; analyses conducted instead using performance ratings (made by the researcher) failed to reveal any influence whatsoever (data available on request), and although generalised anxiety (measured using the GAD-7) predicted a general decline in memory performance in anxious participants, there was no evidence of an interaction. For example, using a median split, the production effect for a low-anxiety participant measured using either the aloud, PE = 0.43, $CI_{95\%}$ [0.09, 0.77], or voice condition, PE = −0.02, $CI_{95\%}$ [−0.33, 0.28], was not credibly different from what would be expected for a high-anxiety participant measured using either the aloud, PE = 0.47, $CI_{95\%}$ [0.11, 0.83], difference = 0.04, $CI_{95\%}$ [−0.45, 0.53], or voice condition, PE = 0.02, $CI_{95\%}$ [−0.30, 0.33], difference = 0.04, $CI_{95\%}$ [−0.40, 0.47].

Researcher location also failed to demonstrate an interaction capable of explaining the absence of a production effect for the voice condition. In this respect, there was a trend toward overall performance being greater across

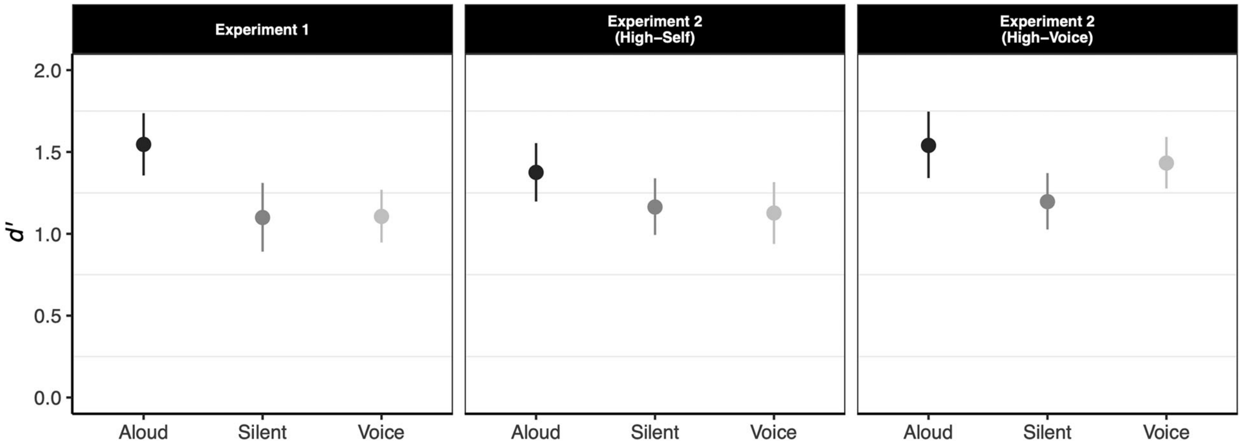

Figure 1. Sensitivity (*d′*) as a function of production condition (silent, aloud, voice), experiment and group (high-self, high-voice); error bars here reflect 95% CIs.

all conditions when the researcher remained in the room. More importantly, as depicted in Figure 2, the magnitude of the production effect tended to be – if anything – numerically larger when the researcher remained in the room, for both the aloud, difference in PE = 0.11, $CI_{95\%}$ [−0.37, 0.60], and the voice, difference in PE = 0.08, $CI_{95\%}$ [−0.35, 0.51], conditions, although these differences were small and highly uncertain. As a result, performance anxiety is unable to explain the absence of a production effect in the voice condition.

With respect to the cognitive demands associated with reading aloud as a character, it is possible that participants had difficulty balancing the performative elements of the task with the mnemonic requirements. We had tried to minimise the risk of this occurrence by including a large number of practice trials, but given our decision to use multiple character voices it is true that practice was spread across several voices. Nonetheless, this account might predict that any difference in the magnitude of the production effect between the aloud and voice conditions might dissipate across the study phase. To evaluate this possibility, our models were re-fit including serial position as a predictor (to allow *d′* calculations foil items were randomly assigned study-phase serial positions; two alternate random assignments produced similar results). As depicted in Figure 3, there was a trend favouring a larger production effect for items studied toward the end of the study phase, but this trend was of similar magnitude between the aloud and voice conditions: Because there is no reason to expect the cognitive demands associated with production in the voice condition to have a comparable effect on the aloud condition, this speaks against the cognitive demand account.[3] Further, this relation is only a weak trend (the slope for the production effect is 0.05, $CI_{95\%}$ [−0.09, 0.20], and 0.01, $CI_{95\%}$ [−0.12, 0.15], for the aloud and voice conditions, respectively, where serial position had been standardised prior to analysis) and there remains no production effect within the voice condition even for items presented at the end of the study phase (the production effect predicted for the final study-phase trial within the voice condition is 0.03, $CI_{95\%}$ [−0.30, 0.36]). Finally, it is worth noting that other forms of production that are arguably more anxiety provoking and difficult to enact (e.g., singing isolated words; Quinlan & Taylor, 2013) have shown production effects.

The remaining accounts – task-switching and contextual effects based on reinstatement – are more difficult to eliminate. Task-switching costs may occur as participants shift out of the old and into the new production modality at the outset of each trial; however, whereas there were an equal number of silent and aloud trials, the voice trials are further split across three unusual voices. Thus, when preparing to produce within the voice condition, there is a greater chance that participants will have just recently enacted some other condition. As a result, task-switching stands to incur a greater cost during those trials. Contextual effects instead refer to the idea that – during test – participants might reinstate production (e.g., imagine saying the item aloud or otherwise activate its sensorimotor representation) as a means of facilitating retrieval. In doing so, they are liable to imagine saying the word in their own voice, which may place words read in someone else's voice at a relative disadvantage. Because there were few iterations of any individual character voice (10 items each), it is unlikely that the participant would think to use this information as a retrieval cue. Our next experiment will speak to this possibility with a modification of our present paradigm.

Experiment 2

Our second experiment was designed to address concerns delineated following Experiment 1 and to provide a preliminary test of the contextual account. To address performance anxiety and the cognitive demands of producing in someone else's voice, we doubled the amount of practice provided to our participants and provided additional, detailed instruction. To further aid in

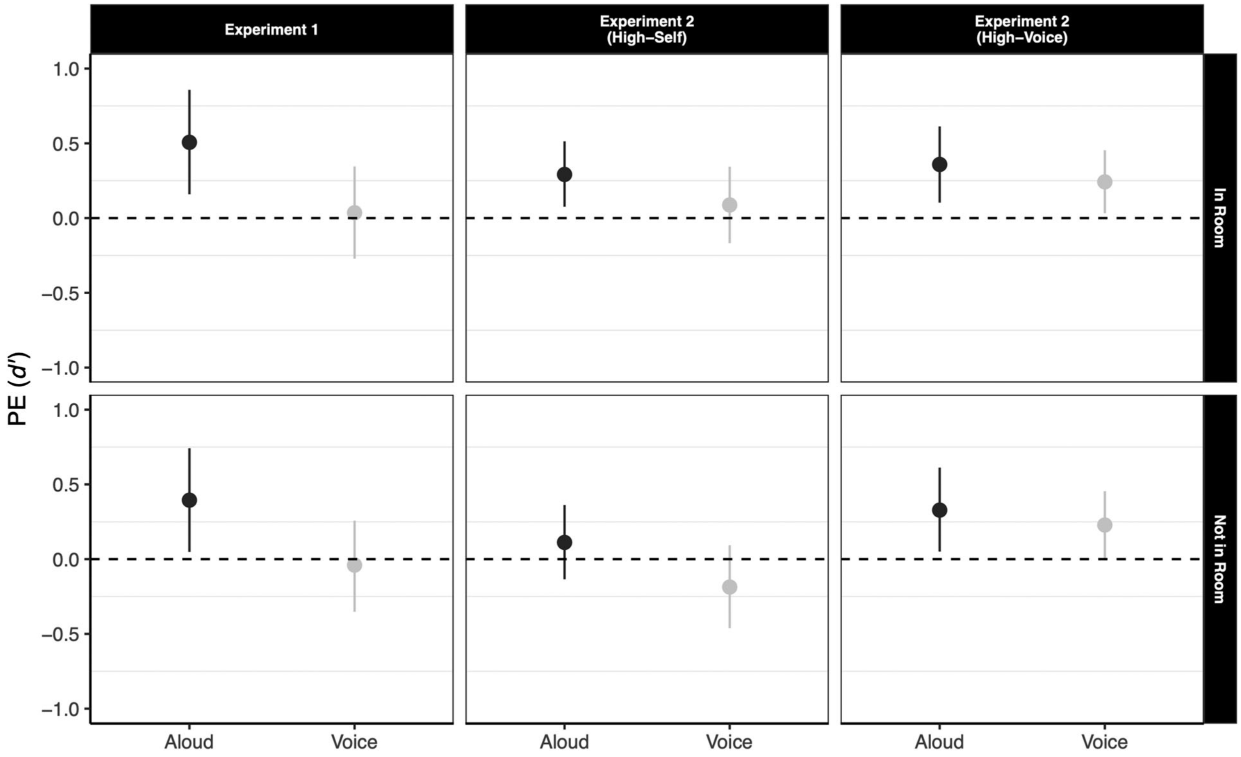

Figure 2. The production effect (*d′*) for Experiments 1 and 2, calculated using either the aloud (dark grey) or voice (light grey) condition, as a function of whether the researcher was in the room (in room, not in room) during test and group (high-self, high-voice); error bars here reflect 95% CIs. Dotted line reflects 0 (no production effect).

this respect – and to address concerns pertaining to differential task-switching costs across conditions – we also eliminated two of the voices (Dracula and Kermit the Frog), having participants only ever produce words in their own voice or as Elvis (which was the voice participants reported feeling most comfortable producing). Finally, to evaluate the possibility of contextual effects operating at retrieval, we also manipulated the relative number of aloud and voice trials at study to produce high-self (45 aloud trials) and high-voice (45 voice trials) groups. We reasoned that if participants were reinstating production as a means of augmenting retrieval, that they would be unlikely to do so with Elvis in mind if they had only received a small portion of voice trials (akin to Experiment 1). Therefore, we predicted no production effect within the voice condition for the high-self group but predicted it would be more likely to emerge in the high-voice group.

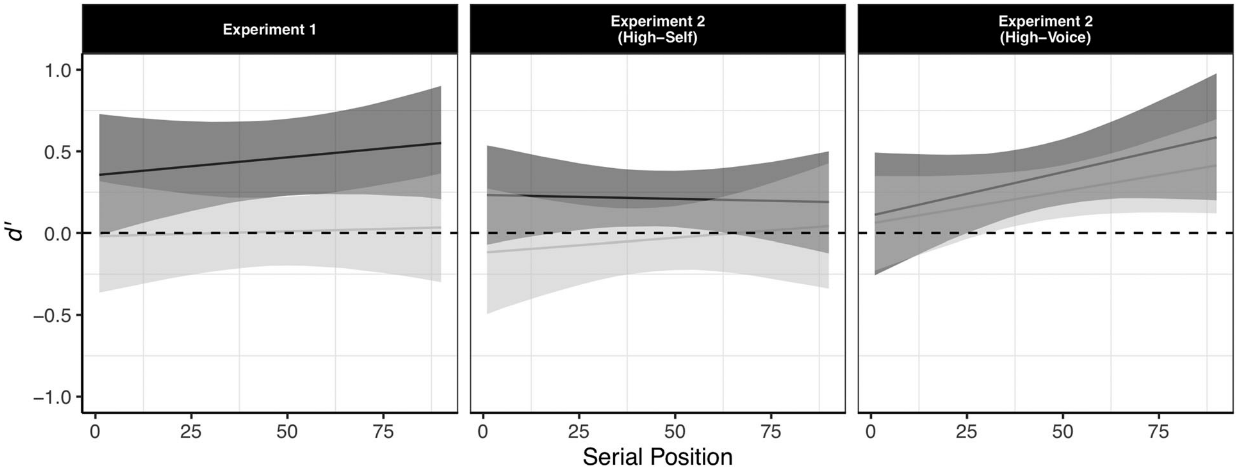

Figure 3. The production effect (*d′*) for Experiments 1 and 2, calculated using either the aloud (dark lines and ribbon) or voice (light line and ribbon) condition, as a function of study-phase serial position (1-90) and group (high-self, high-voice); error bars here reflect 95% CIs. Dotted line reflects 0 (no production effect). Note that there was evidence *against* an effect of serial position in either experiment (see text for further details).

Method

Participants

Our minimum target sample size was again twenty-four per group, although our stopping rule allowed for the possibility of gathering more participants if the term permitted. A total of sixty-six undergraduate participants (42 female; 63 right-handed; mean age = 23.7, SD = 9.2) were recruited for this study via the psychology research experience pool (PREP) as well as through poster advertisements placed around the Memorial University of Newfoundland St. John's campus. Those registered in an eligible psychology course participated in exchange for partial credit toward their overall grade while others were provided an honorarium in exchange for their time. Of these participants, four were excluded because they did not read the words as Elvis (as in the preceding experiment, they participated as observers), one demonstrated incredibly low recognition performance (~30% hit rate in each condition and a 30% false alarm rate; this participant was excluded *after* viewing their data) and one reported a neurological condition affecting their memory. Thus, this left us with a total sample size of sixty. One further participant reported consistently inverting the direction of the confidence rating, but rather than losing the participant these data were simply inverted during pre-processing (exclusion of the participant makes no difference and inspection of their data supported their self-reported inversion).

Prior to beginning the experiment, participants were randomly assigned to one of two groups: In total, 31 participants (17 with the researcher in the room and 14 with the researcher outside the room) were assigned to the high-voice condition in which 70% of the production instructions were "Elvis" and 30% were "Yourself"; the remaining 29 participants (16 with the researcher in the room and 13 with the researcher outside the room) were assigned to the high-self condition in which 70% of the production instructions were "Yourself" and 30% were "Elvis".

Stimuli and apparatus

The stimuli and apparatus were identical to Experiment 1, with the exception that Elvis was the only voice used.

Procedure

The procedure was identical to Experiment 1, with the following exceptions: (a) Participants received either relatively more aloud items (45 aloud/15 voice) or relatively more voice items (15 aloud/45 voice) during the study and test phases, depending on their group assignment (all participants received 30 silent items and training was the same between groups); (b) the demonstration and training phases were expanded and each repeated twice, with feedback provided in between – participants now completed 42 practice trials, 18 of which involved production using the Elvis voice. Such a large number of practice trials was included to lessen the cognitive demands associated with having to use an unusual voice during the primary task; and (c) we included anxiety measures more targeted toward self-consciousness: The Liebowitz Social Anxiety Scale (LSAS; Liebowitz, 1987) and the Self-Consciousness Scale-Revised (SCS-R; Scheier & Carver, 1985).

Results and discussion

Table 1 again depicts the mean percentage hits and false alarms corresponding to each condition. Our models were the same as in Experiment 1, except that they now included both item type (silent, aloud, voice, foil) and group (high-self, high-voice) as predictors. As depicted in the middle and right panel of Figure 1, our predictions were supported: A standard production effect was observed for aloud items in both the high-self, PE = 0.21, $CI_{95\%}$ [0.04, 0.38], and high-voice, PE = 0.34, $CI_{95\%}$ [0.15, 0.54], groups that did not differ credibly from one another, difference in PE magnitude = 0.13, $CI_{95\%}$ [−0.38, 0.11]; however, a production effect was observed for voice items only in the high-voice group, PE = 0.23, $CI_{95\%}$ [0.08, 0.39], with a small tendency toward a reverse production effect in the high-self group, PE = −0.04, $CI_{95\%}$ [−0.22, 0.15], difference in PE magnitude = 0.27, $CI_{95\%}$ [0.04, 0.51]. Put differently, within the high-self group, memorability of the voice condition was comparable to the silent condition, difference = −0.04, $CI_{95\%}$ [−0.22, 0.15], whereas in the high-voice group memorability of the voice condition was only slightly (and not credibly) smaller than the aloud condition, difference = −0.11, $CI_{95\%}$ [−0.29, 0.07].

This finding would appear at odds with the performance anxiety explanation discussed in relation to the findings of Experiment 1. If participants were nervous about reading items aloud as Elvis, there would be no reason to expect that a greater number of Elvis items would make any difference – and in fact, a decrease might have been expected in the magnitude of the standard aloud production effect (owing to distraction): However, the magnitude of the standard aloud production effect was, if anything, of numerically greater magnitude in the high-voice condition, despite comparable performance in the silent condition between these groups, difference in silent conditions = 0.03, $CI_{95\%}$ [−0.21, 0.27]. As depicted in Figure 4, few measures related to performance anxiety were credibly predictive of either the aloud or voice production effect, and none offered a compelling explanation as to why the production effect re-appeared for the voice condition in the high-voice group. In fact, only two of the 20 possible high/low comparisons plotted in that figure excluded 0 as a credible value: the high/low contrast for the aloud condition within the analysis of self-consciousness and the high/low contrast for the

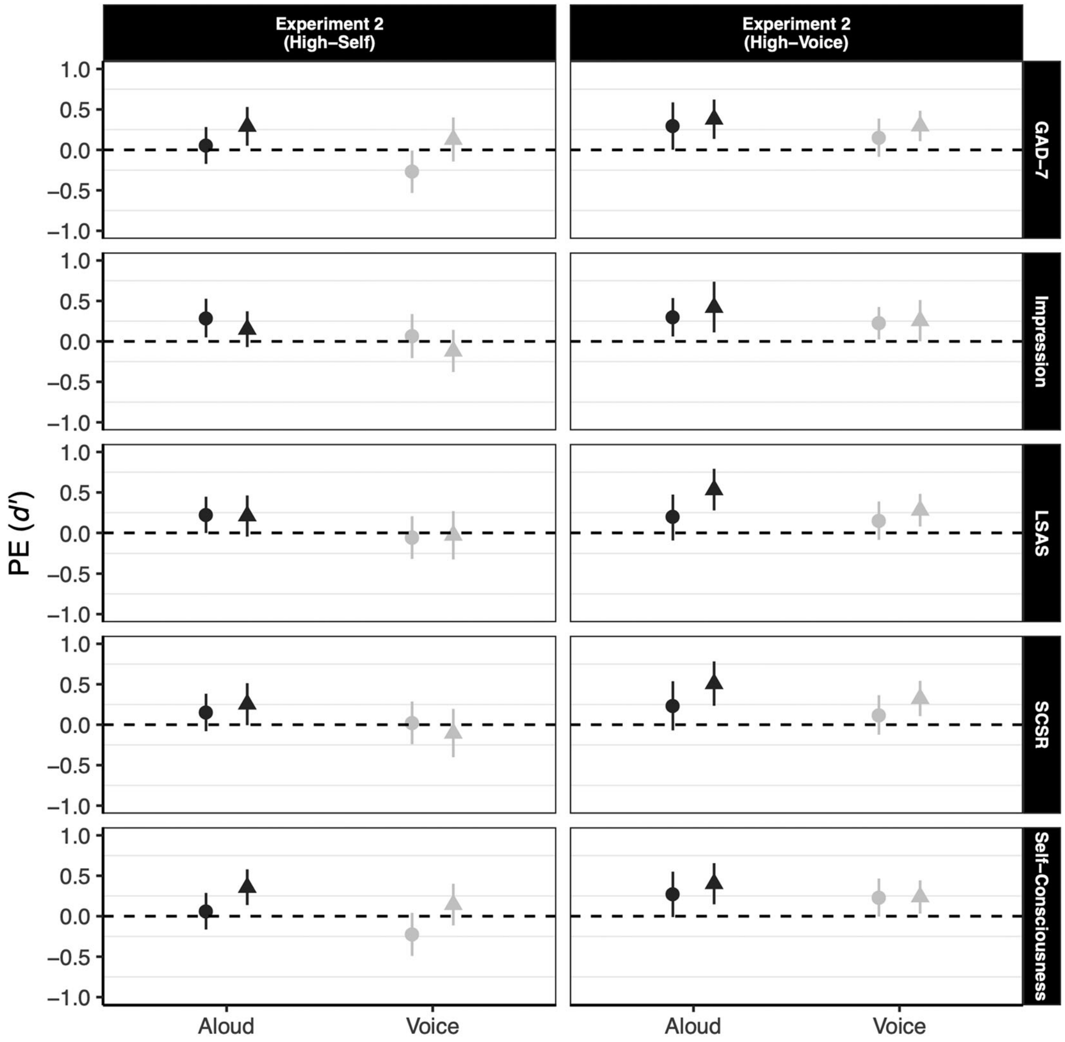

Figure 4. The production effect (*d′*) for Experiment 2 as a function of production condition (silent, aloud, voice) and a median split on the GAD-7, experimenter impression, LSAS, SCS-R or self-consciousness scores; circles reflect low scorers and triangles reflect high scorers for each metric with error bars presenting 95% CIs.

voice condition within the analysis of the GAD-7, both within the high-self condition. Also, as depicted in Figure 2, the data trended in the direction of larger aloud or voice production effects when the researcher was in the room, as compared to when they were not, but this was true of both the high-self and high-voice conditions.

The cognitive demands account also received limited support in the current data. Although one interpretation of our findings would be that participants in the high-voice group had greater opportunity to practice speaking as Elvis, giving rise to a voice production effect in later trials, there are at least two reasons to doubt this account. First, participants received over twice as much practice producing words as Elvis in this experiment compared to our previous experiment with minimal benefit to the high-self condition. If it were merely a practice effect, with diminishing cognitive demands revealing the voice production effect in the high-voice condition, a great deal of practice must be required. However, the second point is that a serial position analysis akin to the one presented in Experiment 1 indicates that the voice production effect was present numerically from early in the study phase within the high-voice group and was absent even late in the study phase within the high-self group, as depicted in Figure 2. If the voice production effect were masked by the cognitive demands of adopting a persona, in the high-self condition the voice production effect should be absent at early serial positions within the high-voice condition and emerge at late serial positions within the high-self condition. This was not the case. Despite an apparent trend favouring a slight increase in the magnitude of the production effect calculated using either the aloud or voice conditions within the high-voice group, neither analyses using serial position nor a comparable analysis separating the data into thirds based on the number of preceding repetitions within each condition found credible evidence of practice effects, and in either case a model excluding the serial position/repetition predictor was favoured by leave-one-out cross-validation over the model including the relevant predictor (*ΔLOO-IC* = 13.0 and 11.8 *against* serial position or repetition, respectively).

With respect to task-switching costs, it is still true that participants are more likely to switch tasks during voice trials for the high-self group and during aloud trials for the high-voice group. This may well contribute to the present pattern but cannot explain the absence of a voice production effect in the former or the presence of a voice production effect in the latter: Specifically, such an account would predict a reduction in the magnitude of the voice production effect for the high-self group but would also predict a reduction in the magnitude of the aloud production effect in the high-voice group. This is not observed. Even so, to further evaluate this possibility, we categorised each test phase item for both Experiment 1 and 2 based on whether participants had to "switch" to that voice at study (i.e., whether the preceding item matched that condition).[4] As depicted in Figure 5, the data trended in the direction expected by a task-switching account, with a numerically comparable production effect for non-switch self and voice trials. However, these findings must be interpreted with great caution as they are based on an average of ∼2 trials per participant in the rare conditions. As a result, it is perhaps unsurprising that none of the apparent differences in the magnitude of the production effect deriving from switching in that figure excluded 0 (i.e., no difference) as a credible value, and further, if a difference did emerge it could be attributed to the repeated (i.e., non-switch) trial "standing out" due to its rarity. Even so, this reflects a plausible partial explanation for our findings, the resolution of which would require additional data capable of maintaining the rarity of our conditions whilst ensuring a suitable number of trials within each cell of the analysis.

The remaining account from our initial experiment is context. Specifically, participants may read each item to themselves during the recognition task. In doing so, it is possible that they also reinstate the original encoding context, thinking about having said it aloud. Presuming they would naturally – or even automatically – do so in their own voice, this would stand to advantage aloud items under typical circumstances or when voice items are uncommon. However, when voice items are common, participants would be more likely to think about having said the test items in that voice, allowing the production effect to re-emerge.

General discussion

The present experiments tested the distinctiveness account of the production effect by determining whether items read in an unusual voice would be remembered particularly well. This rationale built on earlier findings that the magnitude of the production effect scales with the distinctive nature of the production modality (e.g., singing; Quinlan & Taylor, 2013). We wanted to take this a step further and specifically look at whether variation within a specific modality of production (in this case, variation in how words are spoken) would also provide additional memory benefits because of enhanced distinctiveness. Contrary to our expectations and to the predictions of the distinctiveness model, we did not find that saying words in a more distinctive manner enhanced the production effect. In fact, if anything, the production effect for items produced in an unusual voice was *less* robust, a finding not predicted by any major theoretical account.

In Experiment 1, we obtained a standard aloud production effect as measured by sensitivity in a signal detection model; however, counter to our expectations, performance in the silent and voice conditions were roughly equivalent. We delineated several theoretical perspectives that might account for these findings, including (a) performance anxiety, (b) the cognitive demands of assuming a character, (c) task-switching costs or (d) contextual effects at test. In either experiment, we found minimal evidence to support the role of performance anxiety or the cognitive demands of assuming a character in either the aloud or voice production effect. However, while some of our analyses (or manipulations, such as having the researcher in a different room) suggested that anxiety or self-consciousness might influence the magnitude of the production effect (see also, Forrin et al., 2019), none could explain the absence of the effect for the high-self condition or the emergence of the effect for the high-voice condition. With respect to cognitive demands, the magnitude of the voice production effect failed to credibly increase for items encountered later in the study phase, after the participant had more experience reading words as a character (although there was a numerical trend, as depicted in Figure 2).[5] Experiment 2 further observed the emergence of a voice production effect when most words read aloud were read in that voice: Importantly, the voice production effect emerged early in the study phase and was again not predicted by anxiety or the presence of the researcher. As for task switching, this remains a plausible explanation, although our present analyses failed to produce credible evidence that the reduction in the magnitude of the voice production effect was driven specifically by no-switch trials; as depicted in Figure 5, the voice production effect does numerically re-emerge for the no-switch trials within the high-self group, but the effect was measured imprecisely due to the small number of non-switch voice trials, allowing for effects close to (or less) than 0 (and notably credible values inclusive of the corresponding switch trial estimate). Switching had even less of an impact in Experiment 1 or in the high-voice group of Experiment 2.

Our results thus stand as a new challenge in understanding the production effect. How is it possible that making a voice *more* distinctive could *decrease* its

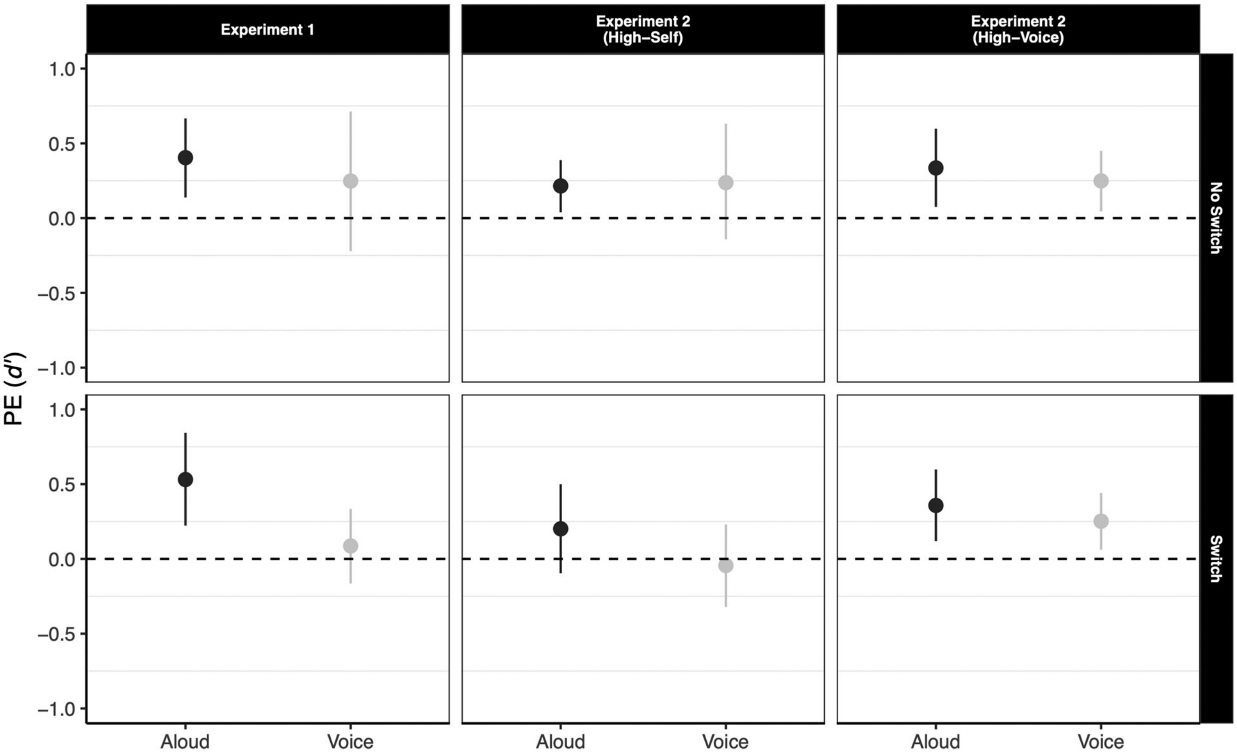

Figure 5. The production effect (*d′*) for Experiments 1 and 2, calculated using either the aloud (dark grey) or voice (light grey) condition, as a function of whether the condition of the preceding study trial matched the current trial (no switch) or mismatched the current trial (switch); error bars here reflect 95% CIs. Dotted line reflects 0 (no production effect). *Note*: Some cells in this analysis contain as few as 1–2 trials per participant and ought to be interpreted with caution (see in-text for details).

memorability? This pattern seemingly contradicts the distinctiveness-based account, which supposes that the more distinctive an item is at study the more memorable it should be at test. However, we believe this data can be explained. Although not definitive, our findings remain compatible with the notion that the identity in which the items had been read at study moderates the utility of the production record. With respect to the mechanisms involved, one possibility is that participants reinstate the productive act during commission of the recognition task itself. This could be in the form of reading the item silently to themselves (presumably in their own voice) or even imagining having produced the word as an intrinsic retrieval cue. This might be viewed as an inversion of the typical distinctiveness account where – rather than using memory of having produced an item to discriminate "old" from "new" items – participants reinstate having produced the item as a means of enriching the available retrieval cues. Importantly, according to this *reinstatement account*, the nature of the associated imagery and its relation to study phase conditions should predict ultimate performance. In Experiment 1 or in the high-self condition of Experiment 2, participants rarely read items in a different voice during study, which means they would be relatively less likely to incorporate that voice into their processing of items at test. As well, reading is an automatic skill (Augustinova & Ferrand, 2014) and participants may be inclined to not just automatically pronounce words but do so internally in their own voice. As a result, the production record for those items would be less useful as a means of re-activating the item (due to mismatch between the production record stored with the episode and the retrieval cue generated at test). Within the high-voice condition of Experiment 2, the relative ubiquity of voice items may have instead encouraged participants to incorporate the voice into their processing of items at test, possibly by imagining having read the item as Elvis in addition to having read the item in their own voice. This would make better use of the production record for voice items, owing to the better match with the generated retrieval cues.

To evaluate this possibility further, we implemented a computational model of the production effect and our unusual voice findings (described in the Online Supplement) based on the REM model introduced by Shiffrin and Steyvers (1997). We chose to use REM rather than MINERVA2 (which had previously been used to model the production effect by Jamieson et al., 2016) because the REM model provides a natural means of representing the distinctive nature of the unusual voices and of handling unique voice representations. With respect to the former, REM represents individual traces within memory as a series of numerical features. In our case, each word was represented by 30 features, reflecting visual, semantic, etc. components incorporated into the encoding episode for all conditions. As has been done by past computational models (e.g., Jamieson et al., 2016), the production effect was implemented by appending an additional 10 features reflecting sensorimotor components associated with the act of production itself (these features were replaced with 0's for the silent condition, as they were not encoded). Importantly, whereas features can take on any positive integer, higher values are considered more distinctive in memory. In this manner, the act of adopting an unusual voice during production may be emulated by increasing the propensity for productive features to take on "rarer" values within the voice condition (as compared to the aloud condition) whilst maintaining the same number of sensorimotor features as the aloud items. At test, the features associated with the target item for each trial are re-presented as a probe, and activation summed based on the content of memory: Here, we emulated participants reading the words at test in their own voice or the unusual voice (Elvis) by modifying the sensorimotor features accordingly.

Further details pertaining to the REM model can be found in the Online Supplement (and code is available on request), but to summarise, we failed to replicate any production effect when the production record was excluded entirely from the test probe, approximated the pattern observed in Experiment 1 and the high-self condition of Experiment 2 when the production record always reflected the participant's own voice, and approximated the pattern observed in the high-voice condition of Experiment 2 when the production record reflected a mixture of one's own voice and the unusual voice. The first implication of these simulations is how easily the REM model can account for the production effect by simply adding additional sensorimotor features. However, no production effect was observed unless participants covertly cued themselves with those sensorimotor features at test. In addition to being one of the first applications of an REM model to the production effect, this finding also lends credence to the notion that participants may use covert reinstatement of the production record to aid retrieval.[6] Importantly, when the production record used at test mismatches the production record encoded at study (as may have been the case for voice items in Experiment 1), the production effect largely disappears. However, if participants are cued to make use of the voice production record, even inconsistently, the production effect re-emerges in this condition. This reflects a quantitative demonstration of the reinstatement idea, including the inconsistent production effect for voice items across our experiments.

However, many questions remain including whether reinstatement is a conscious or unconscious process, why participants would not always use the character voice (or why they would *ever* use the character voice) at test and why other forms of especially distinctive production (such as singing) are not similarly impacted. Concerning the former, reinstatement could either be used as a conscious strategy (reflecting the mirror image of the distinctiveness heuristic, covertly emulating the production record to aid retrieval rather than using access to it as a guide to retrieval) or it could derive naturally from reading the word at test. Further research is needed to resolve either possibility, but participants do self-report explicitly using the production record at test in both within- and between-subject designs (Fawcett & Ozubko, 2016) and a recent study using functional magnetic resonance imaging during a production task found evidence of sensorimotor activation consistent with convert reinstatement (or retrieval) of the production record at test (Bailey et al., 2021).

With respect to why participants might vary in their application of their own voice or the character voice during reinstatement, we can only speculate. It is likely that using one's own voice would come more naturally when either reading the test items or reinstating production of the test items. For that reason, perhaps participants would not think to use any other voice unless induced (e.g., by having them read most items in that voice at study). In contrast, singing is a production strategy that has been reported to offer additional benefits even above reading aloud (Quinlan & Taylor, 2013), even though one might imagine participants equally unlikely to reinstate singing at test. Here, the difference is that singing is still fundamentally one's own voice, albeit with additional features (e.g., timbre; as argued by Quinlan & Taylor, 2013). Further, all currently supportive examples of the singing superiority effect use a foil matching procedure based on Fawcett et al. (2012) whereby participants are told which condition test items have been drawn from (which could prompt reinstatement of earlier singing), and not all attempts to replicate the additive benefits of singing have been successful (for a meta-analysis, see Whitridge, Huff, Ozubko, Lahey & Fawcett, 2022).

It is also worth noting the relation between this account and an either contextual or self-referential explanation. With respect to the former, it is possible that adopting a persona during the production of voice items might cause a shift in mental context, effectively segregating the voice items from their aloud counterparts (and resulting in a mismatch in context at test). Such a shift need not occur (or need not be as drastic) simply because one has altered the intonation of their voice whilst producing the word (as is liable to happen when singing a single word). In addition to altering mental context, it is also possible that adopting a persona discourages self-referential processing, and that self-referential processing contributes to the production effect. According to research on the self-reference effect (for a classic meta-analysis, see Symons & Johnson, 1997), processing information in relation to the "self" results in a major improvement to memory by encouraging deeper encoding and connecting the memory to one's self-representation. Although production tends to bear greater similarity to the enactment effect (e.g., Roberts et al., 2022), it is possible that the act of production encourages a similar connection; if so, making that connection to a self-enacted character – rather than the self – would reduce any encoding advantage unless the character were also enacted at test. Either the contextual or self-referential account might interpret the re-emergence of the voice production effect in Experiment 2 as reinstatement of mental context or re-activation of the schema corresponding to the character.

The present studies represent the first production experiments conducted using character voices, but are not the only ones to explore how saying something in an unusual way influences memory. Cho and Feldman (2013) had participants listen to words spoken in a familiar or unfamiliar accent and then asked them to either listen quietly, repeat the word in their own voice (Experiment 1a), or to mimic the accent of the speaker (Experiment 1b). Unlike the present results, they observed no difference in the magnitude of the production effect – as measured using recall or recognition – as a function of whether participants repeated the word in their own voice or using the provided accent.[7] However, also unlike the present study, each word was presented twice, participants did not know they would be tested, voice (in this case speaking in one's own voice or imitating someone else) was manipulated between-subjects, and – most importantly – participants did not need to *generate* the accent, they only needed to mimic an audio recording. With respect to repeating the words twice, it is possible that the second repetition was less effortful, affording greater benefit and undoing any deleterious effect of the accent; similarly, not knowing they would be tested may have diminished performance for the listen condition (which required the least effort) and advantaged the unfamiliar accent and/or imitation conditions (where participants would need to listen harder or engage more in the productive act). Concerning the manipulation itself, it is unclear whether the voice effects observed in the present study would persist if manipulated between-subjects. However, we believe the most reasonable explanation for differences in our findings is that imitation does not require the same degree of effort or contextual change as is associated with generating persona appropriate accented speech. Specifically, to produce a word as Elvis the participant must activate a schematic representation of that character's speech patterns and then apply that knowledge to generate a word in real-time; to imitate a word spoken as Elvis, the participant must only attempt to re-produce an acoustic pattern, without much consideration to how or why the word was presented in the perceived manner. As such, it is our interpretation that imitation is liable to require less effort and to result in less immersion with the character being imitated. These claims are purely speculative, and further research is needed to resolve differences between these experiments.

Although Cho and Feldman (2013) is the only *published* study of which we are aware manipulating the manner in which one produces a word (beyond work conducted using singing), we also became aware during revision of an unpublished thesis on this topic exploring associative memory between names and faces.[8] Patel (2020) had participants study name-face pairs by reading the name either in their own voice or using an unspecified bizarre voice. Although their predictions had been the same as our own, they also observed marginally worse (associative) recall for the bizarre voice than one's own voice and a similar non-significant trend in recognition memory. Critically, they did not include a silent condition, and therefore were unable to evaluate whether a production effect was observed, and how its magnitude varied according to voice, but given that production has not been found to influence associative memory for name-face pairs (e.g., Hourihan & Smith, 2016), it is reasonable to assume performance for the bizarre voice condition was perhaps even lower than what the silent condition might have been if it had been included. The author in that case attributed their findings to the cognitive effort associated with implementing the voice, akin to the account explored in the present work (see also, Wakeham-Lewis, 2019), and like us they observed the voice < aloud pattern even for items studied in later blocks, after participants would have received substantial practice. Patel's (2020) partial replication of our finding might also shed light on the role of task-switching, as they observed a similar pattern to our own in a paradigm wherein the self and voice conditions were more balanced with respect to frequency (average of ~42%/~58% voice/self trials and no silent trials).

Whatever the mechanism responsible for the present findings, it is worth noting that our current data are – at the least – incompatible with the typical distinctiveness account of the production effect. As described at the outset of this article, we had ourselves predicted that reading items as a character voice should result in a particularly distinctive memory trace.[9] This was not the case in either experiment. Whereas we failed to observe any voice production effect for Experiment 1 or the high-self group of Experiment 2, we also failed to observe any particular benefit of using a character voice over using one's own voice in the high-voice group of Experiment 2. This provides an important boundary condition to the claim that the magnitude of the production effect scales with the distinctive nature of the productive modality (e.g., Fawcett et al., 2012; Forrin et al., 2012; Quinlan & Taylor, 2013). For example, reading a word aloud as Elvis certainly ought to incorporate additional distinctive elements into the production record, as evidenced by the unique pattern of brain activation observed for impersonations not otherwise observed during typical speech (McGettigan et al., 2013). Nonetheless, these distinctive features are only useful under certain circumstances. Whether this is due to the utility or application of the production record (see Fawcett, 2013) remains to seen. In short, whereas the present findings would appear more complementary than adversarial to a distinctiveness account, they suggest that distinctiveness is not everything, and that an appropriately calibrated retrieval strategy is also liable to be important. Luckily, applications of the production effect in the wild (except perhaps in theatre) are unlikely to involve alterations in character between study and retrieval.

In summary, the present experiments provide an unexpected challenge to classic (and future) theoretical perspectives within the production literature. Whereas production remains a useful retention strategy, we have identified an important boundary condition on its usefulness (and, indeed, once again failed to uncover a production modality superior to simply reading aloud; Ozubko et al., 2020). Future research is necessary to fully characterise the nature of this boundary condition and the mechanisms through which it emerges, though for the time being we can conclude that while distinctiveness can sometimes improve memory, distinctiveness and memorability do not share a one-to-one relation and enhancing distinctiveness may even – at times – hurt memory.

Notes

1. Prior to conducting our study, we generated a large number of prospective character voices, including both male and female characters, which we then had a sample of ~12 graduate and undergraduate laboratory members attempt to impersonate and rate for similarity to their own voice using a multidimensional scaling task. The final voices were selected on the basis that they were well-known to our target demographic, easily imitated and as differentiated from their typical speaking voice as possible. Regrettably, all female characters were removed because they were either judged to be insufficiently distinctive (e.g., Lois Griffin) or too similar to a more common male character (e.g., Minnie as compared to Mickey Mouse). Importantly, analyses of our data as a function of participant sex demonstrate the same pattern across all measures for male and female participants. Also, without access to acoustic information pertaining to the voices used, we freely admit we are unable to isolate what it is about the individual voices that drives any observed effects. Importantly, the voices used in Experiment 1 (each possessing distinctive acoustic properties) all demonstrate the same pattern described below when analyzed separately, suggesting the specific acoustics are perhaps not important.
2. Between the test phase and questionnaires, participants also completed a phase in which they rated their familiarity with each of the character voices (amongst others) and undertook a multidimensional scaling exercise meant to quantify conceptual similarity between these characters. However, given the unexpected outcome of our analyses (described next) these data were neither processed nor analyzed; as a result, we will not discuss them further.
3. We would like to thank Drs. Aaron Newman and Colin MacLeod for independently proposing this test of the cognitive demands account when this work was presented at the 2018 annual meeting of the Canadian Society for Brain, Behaviour and Cognitive Science.
4. We have chosen to present the Experiment 1 data here, too, rather than earlier, owing to the fact that with so little data we felt it better to consume this analysis as a whole. Also, defining "switching" based on condition (voice, aloud, silent) or persona (self, voice – here assuming silent items are self) produces the same pattern.
5. A corollary of the cognitive demand interpretation would be that the present findings imply that production alone is insufficient to effect the memory benefit; should the difference between groups in Experiment 2 be due to cognitive demands masking the production effect early in the voice trials, it implies that attention plays a greater role in the emergence of the production effect than permitted by classic interpretations of the distinctiveness account, because according to that perspective having someone produce a word in an unusual, effortful manner would be sufficient to undermine the effect, despite the fact that production did occur (for discussion of the role of attention, see Fawcett & Ozubko, 2016; Mama & Icht, 2018; Mama et al., 2018; Ozubko et al., 2012).
6. Concurrent to – and independent of – our own efforts to model the production effect using REM, Kelly, Ensor, Liu, MacLeod and Risko (2022) produced their own, comparable implementation using this model. We learnt of each other only during revision of our respective manuscripts.
7. Here, we are focusing on the combined analysis of Experiments 1a and b, with an effect of voice on the production effect being reflected in the interaction between production condition (listen, repeat) and production "instruction" (own voice, imitate) and higher order interactions, including with accent (American, Dutch). Arguably, one could instead consider the interaction between production condition (listen, repeat) and accent (American, Dutch) in Experiment 1b, wherein participants imitated in all cases, as reflecting an interaction between voice and the production effect because many participants were probably American and therefore imitation would be similar to using one's own voice; however, imitation in that case would still necessitate some degree of alteration to one's manner of speaking, and further, it is plausible that not all participants had the same accent as the speaker. Regardless, the conclusions were similar once averaged across intelligibility (easy, hard).
8. We thank an anonymous reviewer for orienting us to this work. Although we were unaware of it, we had spoken to the research team in question preceding their project whilst presenting the current experiments also at the annual meeting of the Canadian Society for Brain, Behaviour and Cognitive Science.
9. Here, one might challenge whether reading a voice as a character rather than one's own voice even manipulates distinctiveness. While distinctiveness is a difficult concept to define without resorting to circular reasoning (e.g., Hunt, 2006), we would argue that reading a word as a character ought to be distinctive, as it is not something that participants often do (meaning that it should stand out in memory) and further that doing so has been shown to activate neural patterns distinct from one's own voice (McGettigan et al., 2013). Further, singing was declared a manipulation of distinctiveness with similar reasoning (e.g., Quinlan & Taylor, 2013).

Acknowledgments

The authors would like to thank Kathleen Prior, Rachel Hewitt, Emily Buchanan, and Michelle Johnson for contributions to data collection. We would also like to thank our voice actor Garrett Martin for narrating the voices used in our audio recordings. Lastly, we would like to thank The Combine music for lending their recording equipment.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

JMF was funded by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant.

Open practices statement and data availability statement

None of the data or materials for the experiments reported here are available online, and neither of the experiments were preregistered. However, all data and materials reported in this article are available from the corresponding author on request.

ORCID

Jason Ozubko http://orcid.org/0000-0003-0351-0957

References

Augustinova, M., & Ferrand, L. (2014). Automaticity of word reading: Evidence from the semantic stroop paradigm. *Current Directions in Psychological Science*, *23*(5), 343–348.

Bailey, L. M., Bodner, G. E., Matheson, H. E., Stewart, B. M., Roddick, K., O'Neil, K., Fawcett, . . ., & M, J. (2021). Neural correlates of the production effect: An fMRI study. *Brain and Cognition*, *152*, 105757. https://doi.org/10.1016/j.bandc.2021.105757

Bodner, G. E., & Taikh, A. (2012). Reassessing the basis of the production effect in memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(6), 1711–1719. https://doi.org/10.1037/a0028466

Brysbaert, M., & New, B. (2009). Moving beyond kucera and francis: A critical evaluation of current word frequency norms and the introduction of a New and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Bürkner, P. (2017). Brms: An R package for Bayesian multilevel models using stan. *Journal of Statistical Software*, *80*(1), 1–28. https://doi.org/10.18637/jss.v080.i01

Bürkner, P. (2018). Advanced Bayesian multilevel modeling with the R package brms. *The R Journal*, *10*(1), 395–411. https://doi.org/10.32614/RJ-2018-017

Cho, K. W., & Feldman, L. B. (2013). Production and accent affect memory. *Phonological and Phonetic Considerations of Lexical Processing*, *8*(3), 295–319. https://doi.org/10.1075/ml.8.3.02cho It can also be found here: https://www.jbeplatform.com/content/journals/10.1075/ml.8.3.02cho

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

Cortese, M. J., & Fugett, A. (2004). Imageability ratings for 3,000 monosyllabic words. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 384–387. https://doi.org/10.3758/BF03195585

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

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

Fawcett, J. M., Lawrence, M. A., & Taylor, T. L. (2016). The representational consequences of intentional forgetting: Impairments to both the probability and fidelity of long-term memory. *Journal of Experimental Psychology: General*, *145*(1), 56–81. https://doi.org/10.1037/xge0000128

Fawcett, J. M., & Ozubko, J. (2016). Familiarity, but not recollection, supports the between-subject production effect. *Canadian Journal of Experimental Psychology*, *70*(2), 99–115. https://doi.org/10.1037/cep0000089

Fawcett, J., Quinlan, C., & Taylor, T. (2012). Interplay of the production and picture superiority effects: A signal detection analysis. *Memory (Hove, England)*, *20*(7), 655–666. https://doi.org/10.1080/09658211.2012.693510

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

Forrin, N. D., Ralph, B. C., Dhaliwal, N. K., Smilek, D., & Macleod, C. M. (2019). Wait for it … performance anticipation reduces recognition memory. *Journal of Memory and Language*, *109*, 104050. https://doi.org/10.1016/j.jml.2019.104050

Hassall, C. D., Quinlan, C. K., Turk, D. J., Taylor, T. L., & Krigolson, O. E. (2016). A preliminary investigation into the neural basis of the production effect. *Canadian Journal of Experimental Psychology/Revue Canadienne De Psychologie Expérimentale*, *70*(2), 139–146. https://doi.org/10.1037/cep0000093

Hopkins, R. H., & Edwards, R. E. (1972). Pronunciation effects in recognition memory. *Journal of Verbal Learning and Verbal Behavior*, *11*(4), 534–537. https://doi.org/10.1016/s0022-5371(72)80036-7

Hourihan, K. L., & Smith, A. R. S. (2016). Production does not improve memory for face–name associations. *Canadian Journal of Experimental Psychology*, *70*(2), 147–153. https://doi.org/10.1037/cep0000091

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

Icht, M., & Algom, S. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology*, *5*, 886. https://doi.org/10.3389/fpsyg.2014.00886

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

Kelly, M. O., Ensor, T. M., Lu, X., MacLeod, C. M., & Risko, E. F. (2022). Reducing retrieval time modulates the production effect: Empirical evidence and computational accounts. *Journal of Memory and Language*, *123*, 104299. https://doi.org/10.1016/j.jml.2021.104299

Liebowitz, M. R. (1987). Social phobia. *Anxiety Modern Trends in Pharmacopsychiatry*, *22*, 141–173. doi:10.1159/000414022.

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

Mama, Y., Fostick, L., & Icht, M. (2018). The impact of different background noises on the production effect. *Acta Psychologica*, *185*, 235–242. https://doi.org/10.1016/j.actpsy.2018.03.002

Mama, Y., & Icht, M. (2018). Production effect in adults With ADHD With and without methylphenidate (MPH): vocalization improves verbal learning. *Journal of the International Neuropsychological Society*, *25*(2), 230–235. https://doi.org/10.1017/s1355617718001017

McGettigan, C., Eisner, F., Agnew, Z. K., Manly, T., Wisbey, D., & Scott, S. K. (2013). T'ain't what You Say, it's the Way that You Say It —left insula and inferior frontal cortex work in interaction with superior temporal regions to control the performance of vocal impersonations. *Journal of Cognitive Neuroscience*, *25*(11), 1875–1886. https://doi.org/10.1162/jocn_a_00427

Ozubko, J. D., Bamburoski, L. D., Carlin, K., & Fawcett, J. M. (2020). Distinctive encodings and the production effect: Failure to retrieve distinctive encodings decreases recollection of silent items. *Memory (Hove, England)*, *28*(2), 237–260. https://doi.org/10.1080/09658211.2019.1711128

Ozubko, J. D., Gopie, N., & MacLeod, C. M. (2012). Production benefits both recollection and familiarity. *Memory & Cognition*, *40*(3), 326–338. https://doi.org/10.3758/s13421-011-0165-1

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

Ozubko, J. D., Major, J., & MacLeod, C. M. (2014). Remebered study mode: Support for the distinctiveness account of the production effect. *Memory (Hove, England)*, *22*(5), 509–524. https://doi.org/10.1080/09658211.2013.800554

Patel, P. (2020). Producing Names with a Bizarre Voice Does Not Improve Memory for Face–Name Pairs (thesis). https://macsphere.mcmaster.ca/handle/11375/25886.

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, *51*(1), 195–203. https://doi.org/10.3758/s13428-018-01193-y

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

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. https://www.R-project.org/.

Roberts, B. R. T., MacLeod, C. M., & Fernandes, M. (2022, May 8). The enactment effect: A systematic review and meta-analysis of behavioral. *Neuroimaging, and Patient Studies*, https://doi.org/10.1037/bul0000360

Scheier, M. F., & Carver, C. S. (1985). The self-consciousness scale: A revised version for use with general populations. *Journal of Applied Social Psychology*, *15*(8), 687–699. https://doi.org/10.1111/j.1559-1816.1985.tb02268.x

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. https://doi.org/10.3758/bf03209391

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder. *Archives of Internal Medicine*, *166*(10), 1092–1097. https://doi.org/10.1001/archinte.166.10.1092

Symons, C. S., & Johnson, B. T. (1997). The self-reference effect in memory: A meta-analysis. *Psychological Bulletin*, *121*(3), 371–394. https://doi.org/10.1037/0033-2909.121.3.371

Taikh, A., & Bodner, G. E. (2016). Evaluating the basis of the between-group production effect in recognition. *Canadian Journal of Experimental Psychology*, *70*(2), 186–194. https://doi.org/10.1037/cep0000083

Tillotson, S., Siakaluk, M., & Pexman, P. (2008). Body—object interaction ratings for 1,618 monosyllabic nouns. *Behavior Research Methods*, *40*(4), 1075–1078. https://doi.org/10.3758/BRM.40.4.1075

Wakeham-Lewis, R. (2019). Does a funny voice make for a distinctive memory trace? (thesis).

Whitridge, J., Huff, M. J., Ozubko, J. D., Lahey, C., & Fawcett, J. M. (2022). *Does the song remain the same? Singing does not necessarily improve memory more than reading aloud.* Manuscript submitted for publication.