Repetition Parallels in Language and Motor Action:

Evidence from Tongue Twisters and Finger Fumblers

Arella E. Gussow[1], Daniel J. Weiss[2], & Maryellen C. MacDonald[1]

[1] Department of Psychology, University of Wisconsin-Madison

[2] Department of Psychology, The Pennsylvania State University

Author Note

Word Count: 12110

Data availability: Materials, data, and analysis code are available at osf.io/ckmgq

Prior Dissemination: The data and ideas included in this manuscript have not been published or presented previously.

Corresponding author:

Arella E. Gussow

Department of Psychology, University of Wisconsin–Madison, 1202 West Johnson St., Madison, WI 53706, United States

Email: gussow@wisc.edu

Abstract

We investigated similarities in language and motor action plans by comparing errors in parallel speech and manual tasks. For the language domain we adopted the "tongue twister" paradigm, while for the action domain we developed an analogous key-pressing task, "finger fumblers". Our results show that both language and action plans benefit from reusing segments of prior plans: when onsets were repeated between adjacent units in a sequence, the error rates decreased. Our results also suggest that this facilitation is most effective when the planning scope is limited, that is, when participants plan ahead only to the next immediate units in the sequence. Alternatively, when the planning scope covers a wider range of the sequence, we observe more interference from the global structure of the sequence that requires changing the order of repeated units. We point to several factors that might affect this balance between facilitation and interference in plan reuse, for both language and action planning. Our findings support similar domain-general planning principles guiding both language production and motor action.

Keywords: language production; motor action; tongue twisters; plan reuse; domain-general processing

Introduction

Goal-directed actions are made up of several subgoals that need to be performed sequentially, often with constraints on the particular order of subgoals within the sequence. When making coffee, for example, one must first reach for the coffee pot before pouring coffee in a cup. Other action subgoals don't have an inherent order (Bernstein, 1967; Wolpert & Landy, 2012); sugar and cream may be added to coffee in either order. There may also be partial overlap in executing subgoals, such as reaching for the cream while adding the sugar.

The need to order subgoals within an action sequence provides support for a central hierarchical plan controlling the serial ordering of subgoals (Lashley, 1951; Rosenbaum et al., 2007). The hierarchical plan maintains the overarching goal during performance, while also monitoring the progression between serial subgoals (Badre, 2008; Cooper & Shallice, 2006; Dell et al., 1997). This includes tracking which subgoals have been performed already, preparing for the next subgoal, and adapting the plan in case the goal changes or a subgoal fails, as when the coffee spills.

Another cognitive domain that requires serial ordering is language production: speakers must carefully sequence the words, phonemes, and other units within their utterance, since the resulting meaning strongly depends on that order. For example, phonemes can combine in several different orders to create a variety of words, e.g., *ant, tan, gnat*. As with research on action, the field of language production has included extensive theorizing about the nature of the utterance plan controlling language production, including memory maintenance and serial ordering (e.g., Dell et al., 1997; MacDonald, 2016).

The idea that both language production and motor action require serial ordering and a hierarchical plan has led several researchers to suggest that some aspects of planning and

ordering in both domains may be domain general, or at least follow similar planning principles (Koranda et al., 2020; Lashley, 1951; Lebkuecher et al., 2022; Logan, 2020; MacDonald, 2013; Rosenbaum et al., 1986). A few researchers have attempted direct comparisons between manual and speech tasks (Anderson & Dell, 2018; Koranda et al., 2020; Lebkuecher et al., 2022; Rosenbaum et al., 1986), but these investigations are relatively rare. This situation is unfortunate, because comparisons of planning in the two domains could contribute to theorizing in each domain and increase our understanding of how serial ordering occurs across different behaviors. Here we consider some progress and limitations in previous investigations of planning in language and action modalities. We then report results from a novel study that compares error patterns in a speech task and a manual task. The tasks were designed to use structurally analogous sequences in each domain, allowing us to explore similarities in sequence planning across language and action.

Error Comparisons in Language and Action

Studies of errors in speech or motor actions can reveal which sequence patterns tend to cause difficulty and which types of errors they elicit. These errors can provide insight into the underlying planning process, or what planning principles guide production in real time (Lashley, 1951; Norman, 1981; Rosenbaum et al., 2007; Wilshire, 1999). Error rates and patterns of errors have been extensively studied in the literature on both spoken language (Acheson & MacDonald, 2009; Dell et al., 1997; Dell & Reich, 1981; Stemberger, 1982; Wilshire, 1999), and motor action (Botvinick & Bylsma, 2005; Drake & Palmer, 2000; Rosenbaum et al., 1986; Starkes et al., 1987).

In a direct comparison between language and action errors, Rosenbaum, Weber, Hazelett, & Hindorff (1986) compared performance on speech and manual tasks using sequences that were designed to elicit errors. In their speech task, participants repeated a series of letters that were marked for whether they were to be given extra emphasis in the utterance. Error rates were found to increase when letters had to be repeated with different stress levels on successive cycles. In a parallel manual task, termed "finger fumblers", repetition of finger-tapping sequences elicited more errors if the number of consecutive taps by the same finger changed from cycle to cycle. From the parallel cause of errors in their speech and manual tasks, Rosenbaum et al. concluded that plans for future productions are prepared by reusing plans that have just been completed (Rosenbaum & Saltzman, 1984), in both language and action. Moreover, they proposed the *Parameter Remapping Effect*: that variable mapping of abstract production "parameters" (e.g., stress, number of taps) to otherwise repeated responses (e.g., letter names, finger taps) increases errors because of interference from the prior plan, which needs to be edited.

The idea that prior plans are modified for subsequent actions is an important one for understanding planning in context for both speech and manual tasks (Koranda et al., 2020; Lebkuecher et al., 2022; MacDonald, 2016; Rosenbaum et al., 1986; Rosenbaum & Saltzman, 1984). Positing abstract plan parameters could clarify how difficulty varies when an action sequence is repeated compared to when it is modified in some way; i.e., when subplans of the sequence are modified to have different parameters. However, a limitation for theorizing about plan modification for subsequent actions is that the key concepts of *plan, subplan* and *parameter* all remain underspecified, for both speech and motor action tasks and across many research groups.

The notion of parameter is clear within Rosenbaum et al.'s (1986) experiments, where responses (or subplans) within a sequence did or did not receive special emphasis, and that emphasis parameter either matched or differed from the prior production of that response. More generally, however, a given action always differs from the prior one in some ways, and it is not clear when some difference would be considered a change in a parameter value. For example, Koranda et al. (2020) investigated benefits of re-using parts of an abstract motor or speech plan in parallel motor and language tasks. They did not examine errors; instead their design investigated whether speech and action sequencing could be modulated, or "primed", by the speech or action produced on a previous trial. In their language task, participants saw a picture that could be described in a sentence with two different word order options; participants tended to use the word order of a preceding prime trial. Similarly, in their manual task, participants were more likely to touch images on a screen in the order they that had been required to use on the preceding trial (left-first or right-first from a start position). This parallel preference for primed orders in language and action tasks indicates that planning in the two domains is similarly affected by the structure of prior productions, consistent with claims that new plans are adapted from prior plans (MacDonald, 2013; Rosenbaum et al., 1986; Rosenbaum & Saltzman, 1984).

It is notable that Koranda et al.'s effects obtained even though the prime trial and the target trial differed in numerous ways. In their language task, virtually all the words to be produced were different in the prime trial and the target trial, and in their action task, the prime and target differed in screen location and extent of hand movements required. The persistence of priming in these cases suggests that the re-use of abstract motor plans is quite robust across modifications of the plan. However, it is not clear how these modifications compare to the modified parameters in Rosenbaum et. al.'s account, or how to equate parameters and/or other

modifications between speech and manual task materials. More generally, the large variation in what is modified versus what is repeated between subsequent plans makes it difficult to compare findings both across studies and between cross-domain tasks within the same study; and this presents a critical challenge for theorizing about plan reuse in language and action sequencing.

Task familiarity

Another important question for examining planning across domains is the role of task familiarity. Language tasks generally require manipulation of familiar units such as words or syllables. In laboratory motor tasks however, task goals are often quite novel for participants even if the required movements might be familiar. Differences in initial task familiarity may lead to different degrees of difficulty or different amounts of learning during the course of the experiment, potentially masking parallels in planning across modalities.

In this regard, it is interesting that Anderson and Dell (2018) used an unfamiliar action task but found substantial parallels to performance on language tasks. Anderson and Dell used a manual button-pressing task that contained statistical regularities in the button sequences, analogous to speech tasks that use phoneme sequences to investigate phonotactic rule learning (e.g., Gaskell et al., 2014; Taylor & Houghton, 2005; Warker & Dell, 2006). Error patterns showed evidence of rapid implicit learning of simple statistical regularities (first-order rules), while more complex regularities (second-order rules) showed significantly better learning after a period of sleep consolidation. Despite the novelty of the button-pressing task, Anderson and Dell's results are similar to learning patterns in speech-based phonotactic rule learning tasks (Gaskell et al., 2014; Warker, 2013; Warker & Dell, 2006). These comparable results suggest that the complexity of statistical regularities affects pattern learning in a way that may be a

domain-general feature of sequence production (see also Fischer-Baum et al., 2021). The work also suggests that in at least some circumstances, marked differences in task familiarity need not prevent findings of analogous behavioral patterns in speech and manual tasks.

These provocative yet sparse findings lead to several unanswered questions concerning sequencing in action and language. One of them concerns the relationship between the difficulty of a particular subplan, such as a syllable or finger movement, and the difficulty of sequencing various subplans into a longer plan. In Anderson and Dell's (2018) task, the various button presses were likely all at similar levels of difficulty. Similarly, in the analogous phonotactic rule-learning tasks (Gaskell et al., 2014; Warker, 2013; Warker & Dell, 2006), the various phonemes to be produced were likely all at similar levels of difficulty. The varying error patterns in these tasks therefore appear to stem from familiarity of the particular sequences; as participants learn the contingencies embedded in the sequences, their errors are more likely to yield familiar sequences. In Rosenbaum et al.'s (1986) study, difficulty arose when the emphasis parameter was changed between repetitions of a given response (syllable or finger). In that case, it is not clear whether added difficulty comes from the particular sequence of individual responses, from changing the individual response (e.g. tap once vs. tap twice), the rareness of the double-tap, or some combination of these. That design provides a test of parameter remapping but does not necessarily address the source(s) of the added difficulty.

Tongue Twister Sequences

With these concerns in mind, the present study compares a speech and a manual task using sequences that contain identical subplans (syllables or finger presses) arranged in different orders, with no emphasized elements. This design allows us to focus on sequencing as a unique

source of planning difficulty, regardless of the particular subplans in the sequence. We modified materials from a previous tongue twister study by Acheson and MacDonald (2009) in order to create sequence patterns that can be used to study both spoken and manual sequencing.

Acheson and MacDonald created tongue twisters using sequences of four monosyllabic non-words characterized by an ABBA pattern of syllable onsets (the consonants before the vowel in a syllable) and a CDCD pattern of syllable offsets or rhymes (the vowel plus any following consonants). An example is ***s**hif **s**eeve **s**if **sh**eeve,* where *sh* and *s* are the syllable onsets and *if* and *eeve* are the offsets. This type of ABBA onset/CDCD offset sequence is also seen in the first four words of the tongue twister *She sells sea shells by the seashore,* and sequences of this sort are known to yield speech errors (Croot et al., 2010; Wilshire, 1999). Acheson and MacDonald compared error rates on this tongue twister condition to a non-tongue twister condition, which contained the same four non-words in a slightly different order, where positions two and four are switched, creating a sequence in which the onsets follow an AABB pattern and the CDCD offset pattern is retained (e.g., ***sh**if **sh**eeve **s**if **s**eeve*). Results showed that the tongue twister condition (ABBA) elicited more errors than the non-tongue twister condition (AABB) in four different tasks: reading aloud, immediate spoken recall, immediate typed recall, and serial recognition. Because the non-words were identical in both conditions and only their order within the sequence differed, these results show how sequencing itself can modulate difficulty. Moreover, the appearance of these sequencing effects in serial recognition task, with no overt production, suggests pre-articulatory internal planning as the locus of these effects (see also Oppenheim & Dell, 2008; Pinet & Nozari, 2018).

Despite the robustness of these and similar tongue twister effects, the particular difficulty of the ABBA pattern is not well understood. Accounts of serial ordering typically use tongue

twisters to study phoneme repetition effects rather than structural patterns per se (e.g., Monaco et al., 2017; Sevald & Dell, 1994; Wilshire, 1999). Most accounts posit spreading activation between phonemes and words within the sequence (Dell et al., 1997; Dell & Reich, 1981; Sevald & Dell, 1994). The higher error rate for repeated onsets can be explained by several mechanisms: competition between targets with shared phonemes (Dell, 1984), suppression of a previously produced target (Stemberger, 2009), or miscuing of targets that share the same onset (Sevald & Dell, 1994).

Interestingly, Sevald and Dell (1994) do distinguish between near repetition (ABBA) and far repetition (ABAB) effects. They find that for whole word repetition, the ABBA pattern results in a *lower* error rate than the ABAB pattern, perhaps because near repetition (ABBA) limits the activation decay compared to far repetition (ABAB). However, Sevald and Dell note that when producing the ABBA sequences repeatedly, participants tended to regroup the sequence into AA and BB chunks by pausing between pairs of words. This finding aligns with other tongue twister studies showing that the third position in the four-word sequences tends to have the largest increase in errors (Kember et al., 2017). Taken together, the particular chunking of words within the sequences appears to affect performance on the ABBA tongue twisters, but this is not readily explained by spreading activation accounts of phoneme repetition.

However, the idea of reusing prior plans, and the interference caused by modifications to prior plans (Rosenbaum et al., 1986; Rosenbaum & Saltzman, 1984), could help explain the ABBA difficulty. Both the tongue twister sequences with the ABBA onset pattern and the non-tongue twister sequences with the AABB onset pattern included repeated onsets and rhymes, providing an opportunity for reuse of subplans in the sequence. However, at a more structural multi-syllable planning level, the ABBA onset pattern requires switching from one initial

segment order (AB) to its reverse (BA) across pairs of nonwords. This may cause interference from the previous order (AB) when trying to plan the new order (BA) within the same sequence, leading to an increase in errors. By contrast, the non-tongue twister condition contains repeated sequence of initial segments, AA followed by another repetition, BB. Moreover, this chunking of AABB into pairs (AA-BB) aligns with the chunking of the offsets into pairs (CD-CD), but the ABBA chunks (AB-BA) do not align with the offset chunking – making sequencing even more challenging. Importantly, under this account the complexity of the ABBA pattern is at an abstract structural sequencing level, so that it should not be specific to language only – but extend to sequence planning in motor actions (Dell et al., 1997; Logan, 2020; Rosenbaum et al., 1986; Rosenbaum & Saltzman, 1984; Sevald & Dell, 1994).

To test this hypothesis, we developed two analogous tasks adapted from Acheson and MacDonald's (2009) tongue twisters, designed to elicit errors in speech production and manual action. We expect that tongue twisters with the ABBA onset pattern will yield more errors than non-tongue twisters, consistent with prior research. If our structurally analogous manual sequences elicit similar error patterns, this result would support the idea of domain-general sequencing principles guiding plans in the two domains. A finding of different error patterns in the manual and speech tasks, however, could suggest distinct planning systems. In either case, comparisons of error patterns across domains and sequencing conditions may be informative about situations in which prior plans may facilitate or interfere with subsequent action, and more generally about cognitive constraints underlying sequential planning across domains.

Experiment 1

In this experiment, we adapted Acheson and MacDonald's (2009) tongue twister elicitation paradigm and designed a key-pressing motor analogue for it. In designing the key-pressing manual task, phoneme or syllable representations from the speech task were assigned to specific keys on a keyboard (for a similar speech-to-manual translation, see Anderson & Dell, 2018).

Method

Participants

English speakers (N = 103; women = 52, men = 49, non-binary = 2, age M = 19.2, SD = 1.84) participated in the experiment in exchange for course credit or 10 USD. Participants self-reported gender and age information, in either a prescreening participant pool survey (for-credit participants), or in a post-experiment survey (paid participants). An additional 10 participants were removed from analyses; for not being right-hand dominant (7), because of technical difficulties (2), or because of failure to follow instructions (1). The study was approved by the UW-Madison IRB board, and all participants provided written informed consent.

Procedure Overview

The experiment began with the manual task, lasting approximately 45 minutes; followed by the speech task, lasting approximately 12 minutes. We begin with the description of tongue twister materials for the speech task because the manual task was designed to mimic the structure of that task.

Materials: speech task

We created 16 consonant-vowel-consonant (CVC) or CCVC non-words by combining four consonant onset options (*z, zl, sh, shp*) with four vowel-consonant offset options (*av, oov, af, oof*). This resulted in pairs of non-words with phonological overlap, with onsets that were either a single consonant or a consonant cluster including that consonant (e.g., *zoov, zloov; shaf, shpaf*). In designing these materials, we aimed to make the non-words pronounceable but quite unfamiliar to English speakers, to make the speech task more difficult and more aligned with the difficulty of the novel manual task. To establish that the sequences were unfamiliar in these materials, the phonotactic probabilities of the segments in the 16 non-words were extracted from a web-based interface to calculate phonotactic probability for words and nonwords in English (Vitevitch & Luce, 2004). The values are very low: Mean biphone frequency (i.e., segment to-segment co-occurrence probability of sounds within a word) is .0009 (*SD* = .0007), mean positional segment frequency (i.e., how often a particular segment occurs in a certain position in a word) is 0.02 (*SD* = .04).

Trial lists were created by arranging non-words with phonological overlap into sequences of four non-words each. For Tongue twister (TT) trials, the four non-words were ordered such that their onsets followed an ABBA pattern while offsets alternated as a CDCD pattern (e.g., *zav, zloof, zlav, zoof)*. Non-tongue twister (non-TT) trials were created by switching the second and the fourth non-words in TT items, such that the resulting onset pattern was AABB, and offsets again alternated as a CDCD pattern (e.g., *zav, zoof, zlav, zloof*). By arranging different groups of overlapping non-words into different orders we were able to create 16 TT trials and 16 non-TT trials. In addition, we created Control trials that were designed to minimize overlap between non-words within the sequence. To do so, we grouped together non-words with no identical onsets or offsets, and arranged them into orders that avoided similar onsets in adjacent positions (e.g., *zav,*

shoov, zloof, shpaf). Note that the term "Control" here means that the sequences had minimal overlap between phonemes and no particular repetition pattern (see also Monaco et al., 2017; Wilshire, 1999), but not that these trials should necessarily be the least difficult. Because there were more options for Control trials than for the other trial types (since their pattern did not require phonological overlap), a subset of 16 Control trials was chosen pseudo-randomly, ensuring a similar number of occurrences for each non-word across the set of Control trials. The remaining Control trials were used as practice trials in the familiarization phase (see below). Taken together, the full list of test stimuli consisted of 16 TT trials, 16 non-TT trials, and 16 Control trials; see Table 1 for example stimuli.

Table 1

Examples of tongue twister stimuli.

Trial Type	Example Item
Tongue Twister (TT)	*zav, zloof, zlav, zoof*
Non-Tongue Twister (Non-TT)	*zav, zoof, zlav, zloof*
Control	*zav, shoov, zloof, shpaf*

Materials: manual task

The manual task required participants to press sequences of keys on a standard keyboard. For the experiment we used keyboards where the letter keys were all in black with little visible markings on the keys, except that the designated experiment keys had white stickers indicating where participants should place their fingers. Participants placed their left-hand fingers on keys *1,2,3,4* and their right-hand fingers on keys *7,8,9,0*. On screen, they were presented with images

of templates made up of eight small gray boxes corresponding spatially to each of these keys (see Figure 1). For each template, the specific keys to be pressed were colored in black.

In creating these stimuli, left-hand keys were designed to be analogous to the linguistic onsets and were to be pressed first in a trial, immediately followed by the right-hand keys which were analogous to the linguistic offsets. Participants were unaware of this correspondence (recall that they completed the manual task before the spoken task) nor of the relationship between the two tasks. Sixteen templates were created by combining four onsets (pressing number keys *1, 12, 3, 34*) with four offsets (*7, 78, 9, 90*), and a whole trial consisted of a sequence of four templates. For each template, participants first pressed down on the onset (left hand; consisting of either one key or two keys pressed simultaneously), released immediately, and then pressed down on the offset (right hand; consisting of either one key or two keys pressed simultaneously), and released immediately. Key numbers are mentioned in the method section for the benefit of the readers; participants never saw numbers on the keys and placed their hands on keys with blank stickers. The gray and black box stimuli prompting responses were entirely nonverbal.

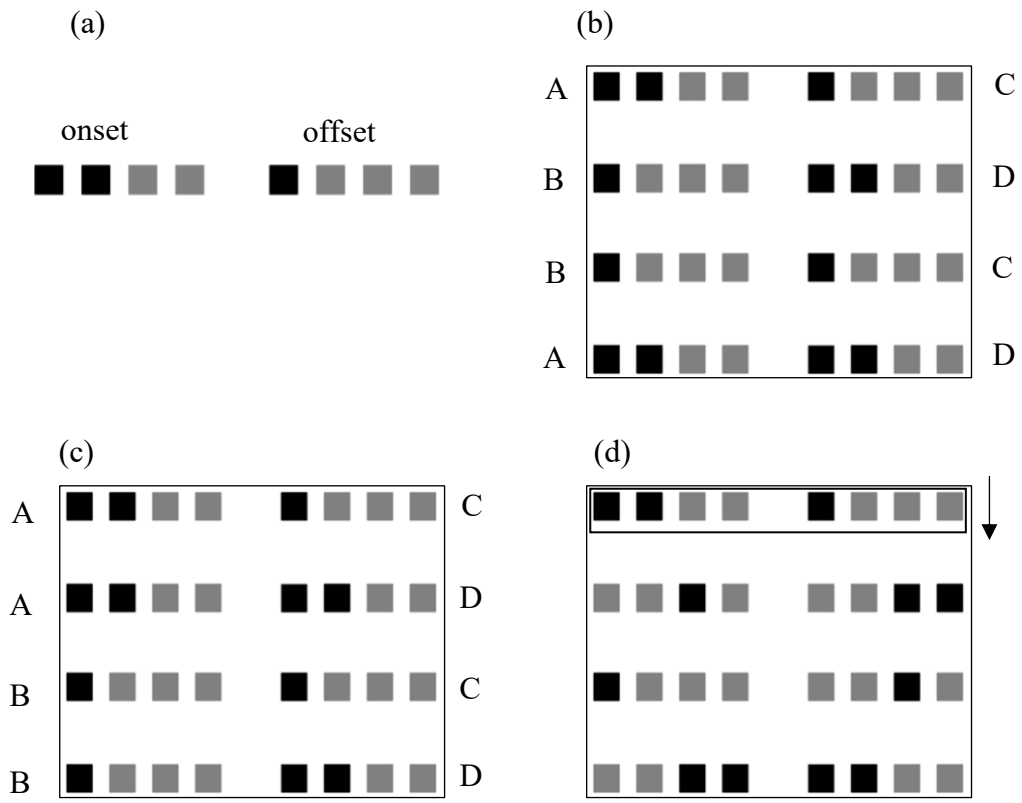

Figure 1. (a) Example of a single template. Each template presented as a row of eight small boxes, and participants were to press the keys corresponding to boxes colored in black. The left side of the template represented onset keys, while the right side represented offset keys. Participants were to press the onset (left hand) first, immediately followed by the offset (right hand). Panels (b), (c), and (d) present examples of full trial sequences. Each trial consisted of four templates presented vertically, ordered according to one of the three trial types: (b) Finger Fumbler (FF), (c) non-FF, (d) Control. Onsets followed an ABBA pattern in FF trials and an AABB pattern in non-FF trials, while offsets followed a CDCD pattern in both trial types. Control trials did not repeat onsets or offsets and did not follow any pattern. During each trial, a rectangle outlining a single template scrolled down from the top of the screen, as illustrated on the first template in panel (d). The rectangle paced the speed of the task by indicating which

template should be pressed. Letter notation and the illustrated arrow did not appear on participants' screens.

Finger fumbler (FF) trials were created by arranging groups of four templates such that their onsets followed an ABBA pattern (e.g., key numbers ***1****+78*, ***12****+7, **12**+78, **1**+7,* where onsets are shown in bold). Non-finger fumbler (non-FF) trials were created by switching the second and the fourth templates, such that the resulting onset pattern was AABB (e.g., ***1****+78*, ***1****+7, **12**+78, **12**+7*). Offsets followed a CDCD pattern in both FF and non-FF trials. The parallel of phonological similarity from the tongue twisters feature was defined here as either pressing a single key (e.g., *1*), or simultaneously pressing that key together with an adjacent key (e.g., *12*). Control trials were sequences of templates with no identical onsets or offsets, arranged into orders that avoid similar onsets in adjacent positions (e.g., *12+7*, *3+90, 1+9, 34+78*). As in the language task, there were more options for Control trials than for the other trial types (since their patterns did not require key overlap), and thus a subset of 16 trials was chosen pseudo-randomly, ensuring a similar number of occurrences for each template across the set of Control trials. The remaining Control trials were used as practice trials in the familiarization phase (see below). The full list of test stimuli consisted of 16 FF trials, 16 non-FF trials, and 16 Control trials.

Procedure

Manual Task. Participants sat in front of a screen in a sound-proof booth with a keyboard and microphone. An experimenter showed them how to place their fingers on the designated keys. An instruction screen explaining how to interpret the templates was presented together with an example template. Participants were told that for each template, they should press the left-

hand keys first, followed immediately by the right-hand keys. They were also told that when two keys needed to be pressed by the left or right hand, the two keys should be pressed simultaneously.

Participants next had three types of practice blocks to familiarize themselves with pressing keys in response to templates. The three blocks of practice together lasted approximately 25 minutes with frequent opportunities for taking a short break. This practice sequence was developed in pilot testing. The stimuli for the practice phases resembled the control conditions in the main experiment, as described above. That is, the stimuli consisted of the 16 templates used in the experimental trials, but never in the order of the test FF and non-FF trials.

For the first practice block, participants viewed one template displayed on screen and pressed the indicated keys in response. They proceeded to respond to one template at a time until they had correctly responded to each of the sixteen stimulus templates. If the response to a template was incorrect, participants saw a message on screen instructing them to try again. The participant repeated that template until a correct response was provided. If there was a lag of more than 300ms between pressing the left-hand and the right-hand keys for a given template, participants received a message on the screen instructing them to try to respond faster.

In the second practice block, participants practiced responding to a sequence of four templates, in self-paced presentation. These trials began with a single template appearing at the top of the screen. Once participants pressed the corresponding keys, that template disappeared and the next template appeared beneath its location. This continued until completion of the four templates in the sequence. If a response was incorrect, a red ‘X’ appeared at the bottom right of the screen, but the trial continued. Participants completed four self-paced trials.

Finally, participants completed 40 practice trials that had the same display and similar pacing as experimental trials. These trials began with a self-paced presentation stage as in the second practice phase, after which a timed trial began. In timed trials, the same four templates appeared simultaneously, vertically oriented on the screen, and after 550ms a rectangle at the top of the screen began scrolling down. Participants were instructed to press the correct keys for each template once the rectangle reached that template. They were told to be as fast and accurate as possible, and that they could press the keys from the moment the rectangle overlapped with the template even partially but not after it moved on. The speed of the rectangle created a window of 1.2s within which participants had the opportunity to press the corresponding keys. When the rectangle scrolled out of the bottom of the screen, it reappeared at the top to begin another repetition of that trial, until three repetitions were completed. If a response was incorrect, a red 'X' briefly appeared at the bottom right of the screen and the trial continued. After every trial (one self-paced sequence and three timed repetitions), a fixation cross appeared for 500ms before beginning the next trial.

Following the three practice blocks and a short participant-timed break (usually less than one minute), participants began the experimental trials. These trials were identical in structure to the full practice trials (one self-paced and three timed repetitions). However, no feedback was provided, and the timed trials required faster responses than practice trials. The rectangle began scrolling down 420ms after the templates first appeared on the screen with a scrolling rate that allowed 900ms to respond to each template. The first trial in the test phase was an additional practice trial and was not included in later analyses. All participants then saw all 48 test trials (16 FF, 16 non-FF, 16 Control) with order randomized for each participant. The test phase lasted approximately 20 minutes with frequent opportunities to take a short break. Accuracy was

automatically coded within the experimental script. An error was defined as any deviation from the keypresses that participants were supposed to produce, coded separately for the onset (e.g., pressing *1+7* instead of *12+7*) and the offset (e.g., pressing *1+78* instead of *1+7*) of each template.

Speech Task. After completing the manual task, participants began the speech task. First, they were familiarized with all non-words to be used in the experiment. The non-words appeared one by one on the screen for 3 seconds each. They were separated by a fixation cross lasting 1s and participants read each word out loud. After being exposed to all 16 non-words, an instruction screen appeared to explain the task procedure.

Each trial began with a fixation cross for 1s, after which all four words appeared on the screen one beneath the other. Words appeared in black, with a gray rectangle frame surrounding each word; see Figure 2. Frames were of equal size regardless of word length. At the beginning of each trial, a red circle appeared on the left of the topmost word for 2s allowing participants a brief preview of the non-words in the sequence. Next, the red circle disappeared and a rectangle began scrolling down from the top of the screen. The rectangle was a cue – every time it reached a word, participants were supposed to speak that word aloud. The speed of the rectangle created a window of approximately 525ms for responding to each word. This speed was determined based on previous tongue twister studies (Acheson & MacDonald, 2009; Dell et al., 1997; Wilshire, 1999) and additional piloting with our stimuli. After the rectangle scrolled off the bottom of the screen it reappeared at the top to begin another repetition of that trial until three repetitions were completed. The first five trials were practice trials, comprised of the leftover Control trials that were not used as test trials. Each participant then completed all 48 test trials

(16 TT, 16 non-TT, 16 Control) in a randomized order. The entire task lasted approximately 12 minutes.

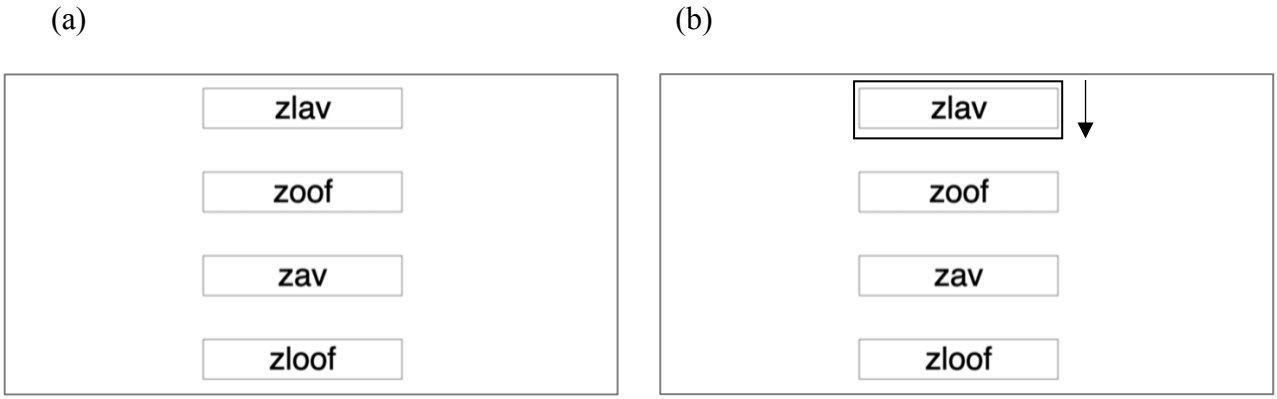

Figure 2. (a) Example of a tongue twister display. Each trial consisted of four non-words presented vertically, and each non-word was framed by a gray rectangle. (b) During each trial, a black rectangle frame scrolled down from the top of the screen, pacing the speed of the task by indicating which non-word should be produced. The illustrated arrow in panel (b) did not appear on the experiment screen.

Analysis

For both tasks, we analyzed errors in the onsets separately from the offsets. Only the first error per repetition was coded, and all further observations of that repetition were removed from the data for that trial. This choice was motivated by evidence that production of an error tends to lead to further errors due to hysteresis (Nooteboom & Quené, 2015), self-correction (Page et al., 2007), learned associations between error and stimuli (Beth Warriner & Humphreys, 2008), or the participant simply giving up.

For the speech task analysis, recordings of participants' speech were transcribed by trained research assistants using the phoneme conventions from the CMU Pronouncing Dictionary (http://www.speech.cs.cmu.edu/cgi-bin/cmudict), and an automated script then

compared transcriptions with the stimulus list and coded for accuracy. An error was defined as any deviation from the phonemes that participants were supposed to produce, coded separately for the onset (e.g., saying *zlav* instead of *zav*) and the offset (e.g., saying *zoof* instead of *zav*) of each non-word.

In the speech task, word positions where participants did not produce any phoneme at all were removed from analyses (5%), and an additional 2% of the remaining observations were removed due to degraded audio files or indecipherable productions. In the manual task, template positions where participants did not press any key at all were removed from analyses (3%).

For the speech task analyses, we ran a mixed-effects logistic regression model regressing Error (0,1) on trial Type (TT, non-TT, Control), word Position (First, Second, Third, Fourth), and the interaction between Position and Type. Position was tested using successive difference contrasts, such that each level was compared to the subsequent one. This resulted in the comparisons of position [Two vs One]; [Three vs Two], [Four vs Three]. Type was coded using the same scheme, resulting in the comparisons of [TT vs non-TT] and [Control vs TT]. The random effects structure (Barr, Levy, Scheepers, & Tily, 2013) included random intercepts for subjects, items (whole four non-word sequence), and words (non-word used across items). A by-subject random slope was included for the interaction between Type and Position, and a by-item random slope was included for Position. For the by-subjects random slope we chose to include only the interaction term, and not the main effects, as this has previously been shown not to affect the type-I error rate (Barr, 2013) and allowed model simplification. Coefficients are reported in log-odds.

Analysis of the manual task was identical to that of the speech task, where "template" in the manual task has the status of "non-word" in the speech task. Here we follow the parallel

terminology used in the methods description: “onsets” refer to the left side of every template (pressed first, with the left hand) while “offsets” refer to the right side (pressed second, using the right hand); FF and non-FF trial types are analogous to TT and non-TT trials types, respectively, in that both FF and TT conditions paired ABBA onset patterns and CDCD offset patterns, while the non-FF and non-TT conditions both had AABB onset patterns and CDCD offsets. In both tasks, Control trials had no repeating pattern of onsets or offsets.

Transparency and Openness

Data, analysis scripts, and stimuli are freely available online (https://osf.io/ckmgq/). The study was not preregistered. Experiment design, analysis, and procedure information are described thoroughly throughout the paper, including the motivation behind our methods and analysis choices.

Results

Rather than analyzing error rates in entire non-words and key-press templates, we conducted separate analyses on onsets and offsets, because these units may convey different information about the difficulty across conditions. Onset analyses reflect the comparative difficulty of the ABBA onset pattern in the TT and FF conditions versus the AABB pattern in the non-TT/non-FF conditions. The offsets in these conditions always had the CDCD pattern and thus are identical across TT/non-TT and FF/non-FF conditions. Any differences in difficulty observed in offsets would therefore reflect the influences of the onset sequences. Because the Controls did not follow any structural pattern but included similar onsets (e.g., *z, zl*) within the sequence, comparing onset errors of the Controls versus the other conditions could help

disentangle effects of the sequence structure from effects of phonological similarity within the sequence (Jaeger et al., 2012; Sevald & Dell, 1994; Sullivan & Riffel, 1999). Similarly, differences in the offset error rates between the Controls and the other conditions could reflect the effect of the regular CDCD offset pattern that both the TT/FF and the non-TT/non-FF conditions followed but the Controls did not.

Onsets: speech task

The mean onset error rate was $M = 0.33$, 95% CI [0.31, 0.34]. Table 2 shows the mean proportion of errors and 95% confidence intervals at each Position as a function of Trial Type, computed from the by-subject means. Figure 3 plots the corresponding model predictions, extracted from the regression analysis. The error rate for TT trials was significantly higher than for non-TT trials ($b = .35$, $SE = .08$, $z = 4.31$, $p < .001$). There was no significant difference between Control and TT trials ($b = .09$, $SE = .09$, $z = 1.09$, $p > .1$). None of the Position comparisons were significant, and none of the interactions between Position and Type were significant (ps > .1). Taken together, these results resemble previous tongue twister results and support our main prediction that TT trials elicit more errors than non-TT trials.

Table 2

Means and within-subject confidence intervals for onset errors as a function of trial Type and Position in Exp 1

		Position							
		1		2		3		4	
Task	Trial Type	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
Speech Task									
	TT	0.37	[0.35, 0.39]	0.36	[0.33, 0.39]	0.3	[0.26, 0.35]	0.3	[0.24, 0.35]
	non-TT	0.31	[0.28, 0.33]	0.25	[0.22, 0.28]	0.3	[0.26, 0.34]	0.23	[0.18, 0.28]
	Control	0.35	[0.33, 0.37]	0.26	[0.23, 0.3]	0.42	[0.38, 0.47]	0.52	[0.45, 0.58]
Manual Task									
	FF	0.1	[0.09, 0.11]	0.12	[0.11, 0.13]	0.08	[0.06, 0.09]	0.1	[0.08, 0.12]
	non-FF	0.14	[0.12, 0.15]	0.07	[0.05, 0.09]	0.14	[0.12, 0.16]	0.09	[0.07, 0.11]
	Control	0.14	[0.12, 0.16]	0.12	[0.1, 0.14]	0.1	[0.08, 0.11]	0.11	[0.09, 0.13]

Note. M and 95% CI represent mean and 95% confidence intervals, respectively; computed from the by-subject means.

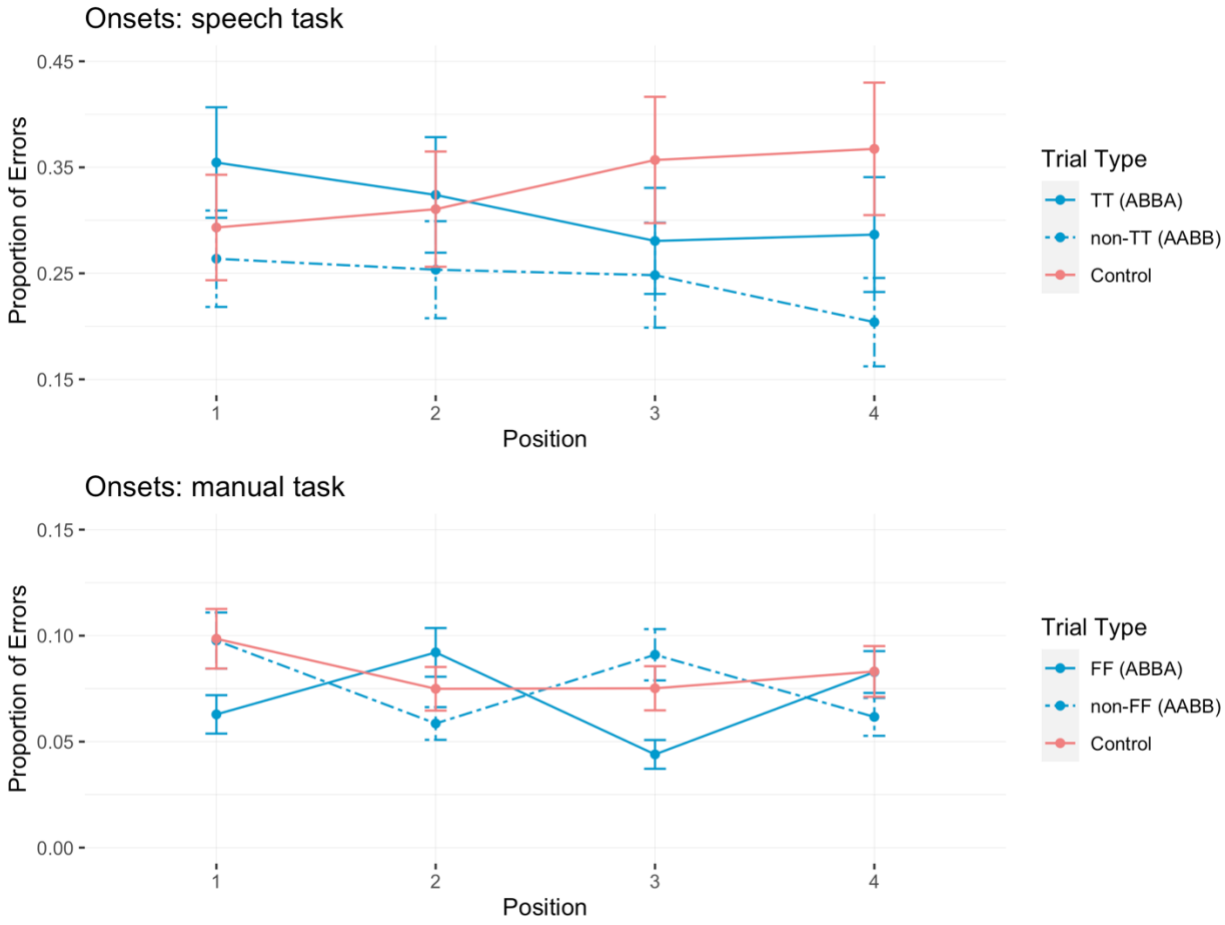

Figure 3. Model predictions for the proportion of onset errors at each word Position as a factor of trial Type, for the speech (top panel) and manual (bottom panel) tasks in Experiment 1. Error bars show 95% confidence intervals. Note the different y-axis scales.

Onsets: manual task

The mean onset error rate was $M = 0.11$, 95% CI [0.1, 0.12]. Table 2 shows the mean proportion of errors and 95% confidence intervals at each Position as a function of Trial Type, computed from the by-subject means. Figure 3 plots the corresponding model predictions. There

was a significant interaction between the comparisons of Type [FF vs non-FF] and Position [Two vs One], indicating the error rate for FF increased between Positions One and Two, while the error rate for non-FF decreased (b = .97, SE = .17, z = 5.58, p < .001). The same pattern of interaction – an increase for FF but decrease for non-FF – was seen between Positions Three and Four (b = 1.09, SE = .2, z = 5.46, p < .001), while the opposite – a decrease for FF while non-FF increased – was found between Positions Two and Three (b = -1.27, SE = .17, z = -7.35, p < .001). As apparent in Figure 3, the pattern arising from these interactions is that repetition of an onset between adjacent positions resulted in a reduced error rate. That is, we do not find a global trial Type effect of the sequence pattern as we had expected, but instead we find local effects from one position to the next.

There was also a significant interaction between the comparison of Type [Control vs FF] and Position [Two vs One], as the error rate for FF increased between Positions One and Two, while the error rate for Controls decreased (b = -.71, SE = .17, z = -4.14, p < .001). Between Positions Two and Three, the error rate decreased for FF but not for Controls (b = .79, SE = .16, z = 4.96, p < .001), while between Positions Three and Four it increased for FF but not for Controls (b = -.57, SE = .18, z = -3.06, p < .01). No other effects were significant (ps > .05). Taken together, the overall error rate for Controls was similar to the other trial types but remained rather stable across positions.

Offsets: speech task

The mean offset error rate was M = 0.27, 95% CI [0.26, 0.28]. Table 3 shows the mean proportion of errors and 95% confidence intervals at each Position as a function of Trial Type, computed from the by-subject means. Figure 4 plots the corresponding model predictions. The

error rate for Control trials was significantly higher than for TT trials (b = .33, SE = .1, z = 3.31, p < .001), and the error rate at position Three was significantly lower than at position Two (b = -.44, SE = .09, z = -4.57, p < .001). No other comparisons or interactions were significant (ps > .05). This suggests that the offset errors were not affected by the patterns of our onset manipulation, though the error rate for Controls was overall higher than the other trial types.

Offsets: manual task

The mean offset error rate was M = 0.1, 95% CI [0.09, 0.1]. Table 3 shows the mean proportion of errors and 95% confidence intervals at each Position as a function of Trial Type, computed from the by-subject means. Figure 4 plots the corresponding model predictions. The error rate for Control trials was significantly higher than for TT trials (b = .69, SE = .1, z = 7.11, p < .001), and the error rate at position Two was significantly lower than at position One (b = -.32, SE = .11, z = -2.97, p < .01). No other comparisons or interactions were significant (ps > .1). As with the speech task, the offset errors were not affected by the patterns of our onset manipulation, but the error rate for Control offsets was overall higher than the other trial types.

Table 3

Means and within-subject confidence intervals for offset errors as a function of trial Type and Position in Exp 1

		Position							
		1		2		3		4	
Task	Trial Type	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
Manual Task									
	FF	0.1	[0.09, 0.12]	0.08	[0.07, 0.1]	0.06	[0.05, 0.08]	0.07	[0.06, 0.09]
	non-FF	0.1	[0.09, 0.12]	0.07	[0.06, 0.09]	0.06	[0.05, 0.08]	0.07	[0.05, 0.08]
	Control	0.15	[0.13, 0.17]	0.11	[0.09, 0.13]	0.13	[0.11, 0.15]	0.13	[0.11, 0.16]
Speech Task									
	TT	0.3	[0.28, 0.32]	0.29	[0.25, 0.33]	0.23	[0.18, 0.27]	0.19	[0.13, 0.24]
	non-TT	0.28	[0.26, 0.31]	0.3	[0.28, 0.33]	0.2	[0.16, 0.25]	0.18	[0.13, 0.23]
	Control	0.3	[0.27, 0.32]	0.37	[0.33, 0.4]	0.27	[0.22, 0.32]	0.3	[0.23, 0.38]

Note. M and 95% CI represent mean and 95% confidence intervals, respectively; computed from the by-subject means.

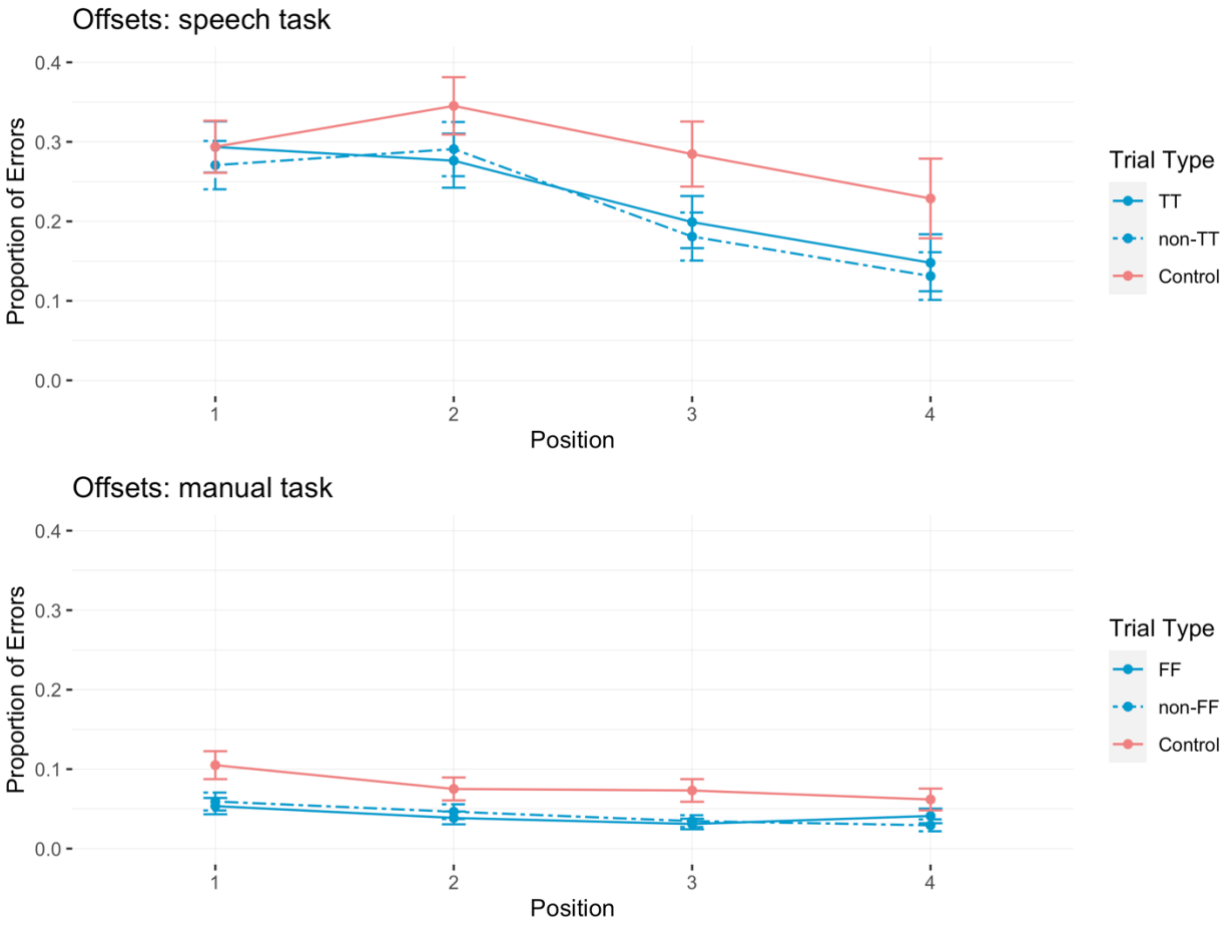

Figure 4. Model predictions for the proportion of offset errors at each word Position as a factor of trial Type, for the speech (top panel) and manual (bottom panel) tasks in Experiment 1. Error bars show 95% confidence intervals.

Individual Differences

Individual differences are often used to assess cross-domain processing similarities, as correlations between participants' performance on two different tasks might indicate a shared underlying processing mechanism (Kidd et al., 2018; Ou & Law, 2017; Vogel & Awh, 2008; Wardlow, 2013). We therefore calculated Pearson's correlation coefficient for the relationship between participants' overall error rate on the manual versus the speech task. There was a

significantly positive correlation between error rates, $R = 0.26$, $p < .01$, see Figure 5. However, note that we did not have a control task to rule out effects of motivation, fatigue, or other situational confounds. Moreover, because we see differences in the error patterns between tasks – potentially indicating different performance strategies (see Experiment 1 Discussion), it is unclear whether we should expect cross-task correlations, or what they might reflect. We therefore hesitate to interpret these results, although they are consistent with our hypothesis.

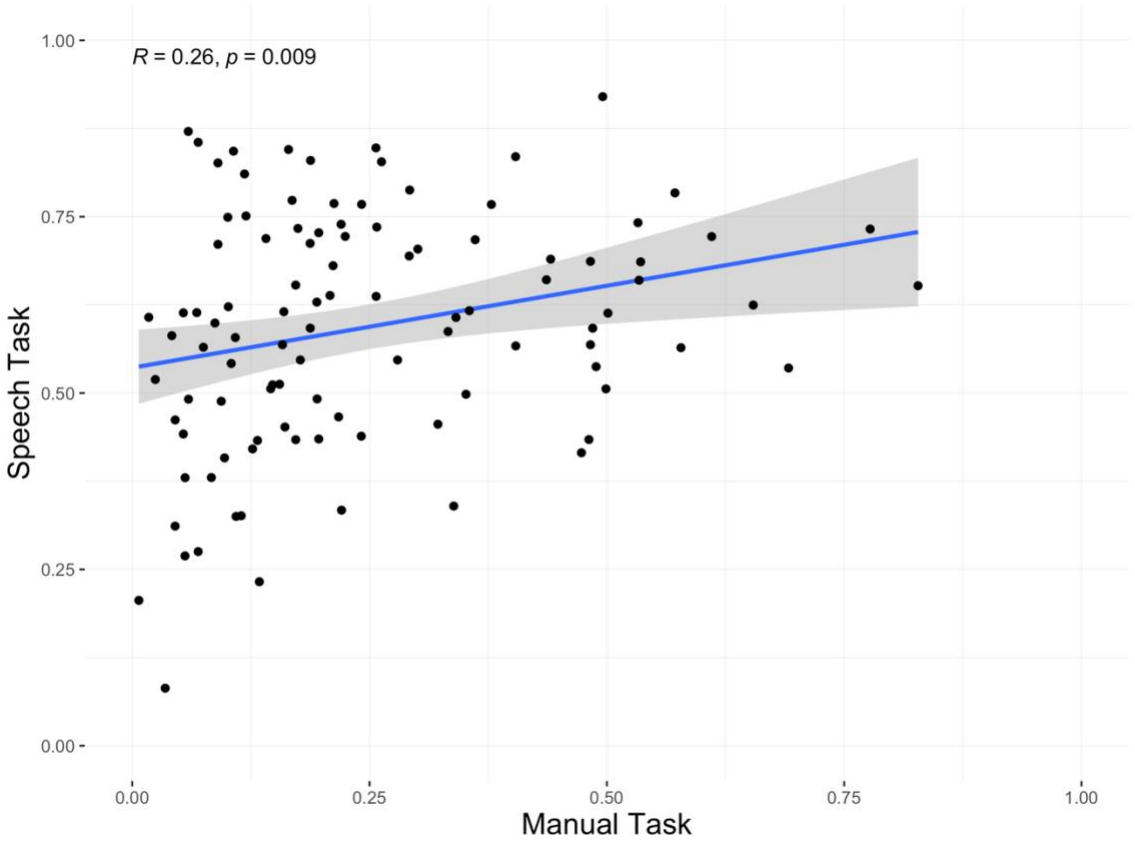

Figure 5. Scatterplot of the correlation between the proportion of errors on the manual task (x-axis) and the speech task (y-axis). Each dot represents one participant. Error bands show 95% confidence interval.

Experiment 1 Discussion

The results of Experiment 1 showed some similarities and some differences across the two tasks. The clearest similarities were evidenced in the offsets errors. In both tasks, the offset error rate for Control trials was higher than TT/FF and non-TT/non-FF trials, with no difference between the latter two. The overall higher error rate for Control offsets may stem from the fact that there was no repetition of offsets within a sequence, in contrast with the other trial types that followed a CDCD pattern of offsets and could benefit from the repetition of subplans. In other words, the contrast between the Control condition and the other conditions may reflect benefits of plan reuse in both modalities. The offsets in the TT/FF and non-TT/non-FF followed the identical CDCD pattern, and the lack of difference here suggests that offset error rates were not affected by the varying onset patterns between conditions.

By contrast, onset errors show two distinct patterns across the two modalities: the tongue twisters elicited the expected effect of trial Type, with an increased error rate for the ABBA onset pattern (TT). The manual task, however, elicited a Type by Position interaction. Specifically, onset repetition between adjacent positions resulted in lower error rates in the manual task: for the A<u>A</u>B<u>B</u> pattern (non-FF) there was a decrease in errors between positions One and Two and between positions Three and Four; while in the AB<u>B</u>A pattern (FF) there was a decrease in errors between positions Two and Three. Although this result is not what we had anticipated, we suggest it can be explained by plan reuse between adjacent onsets in the sequence. Before turning to a direct comparison between the language and action results, however, we first consider differences between the two tasks and their consequences for cross-task comparison.

Although the speech and manual tasks had very similar structure, participants' performance in the two tasks differed in important ways. First, the mean error rate in the speech task (onsets: 0.33, offsets: 0.27) was much higher than in the manual task (onsets: 0.11, offsets: 0.1), and higher than in similar previous tongue twister studies (e.g., Acheson & MacDonald, 2009). This suggests that our attempt to minimize experience differences between language and action may have been too extreme. The difficulty of the speech task seems to stem from the fast production speed and the difficult pronunciation of our unusual non-words. In fact, transcribers noted that several participants tended to add short schwa vowels within the consonant clusters at the onset (e.g., saying *zuh-loof* for zloof). Vowel insertion is a typical error when speakers produce consonant sequences that do not exist in their language (Davidson, 2006; Davidson et al., 2015); this is particularly true of fricative-stop clusters such as those used in the current experiment (e.g., "shp") for English speakers (Wilson & Davidson, 2013). Thus difficulty differences in the two tasks may have complicated attempts to observe commonalities between them.

Second, the pattern of onset errors in the manual task – showing local effects from one position to the next but no effects of the global sequence pattern – brought to our attention a crucial difference between the tasks' stimuli: non-words in the speech task were monosyllabic, and each non-word appeared on the screen as a single unit, with no clear divide between the onset and the offset (see also Dell et al., 2021; Fischer-Baum et al., 2021, for a discussion of the tight organization of the syllable unit). In the manual task, however, onset and offset components were clearly separated on the left versus right sides of the screen and were performed with separate effectors (left hand, right hand). This difference in stimulus encoding and corresponding actions may have led to a different representation of units in the production plan, encouraging a

different performance strategy. Specifically, participants may have limited their planning scope in the manual task to just the current template (instead of planning all four templates in the sequence) since there were more units to represent, and general difficulty in representing such unfamiliar templates. In the speech task, however, planning a sequence of four monosyllabic non-words is not very difficult given participants' prior experience with language (Lindsley, 1975), and this grain of planning was probably preferred in order to keep up with the speed of the task (Wilshire, 1999). Arguably, such differences in stimulus grouping and the resulting planning scope would affect performance on the ABBA pattern, since the pattern's difficulty depends on the global structure of the sequence, across all four nonwords or templates.

To make the speech and manual tasks more parallel in planning, we had two general options. We could create a manual task that was more speech like—more continuous between onsets and offsets and encouraging a broader scope of planning across the four templates. Alternatively, we could create a speech task more like the manual task, with a clear delineation between onsets and offsets. Given that the former would require participants to complete many hours of practice, we chose to adjust the speech task in Experiment 2. We therefore introduced more separation between the onsets and the offsets in each non-word in the speech task. We hypothesized that if the pattern of local Position by Type interactions in the manual task was due to the separability between onsets and offsets in our templates, then a language task that introduces such separation in the non-word stimuli would yield a similar pattern of results – local facilitation when an onset is repeated. This would suggest that the difference between the patterns seen in the Experiment 1 tasks could reflect different planning scopes, which are known to depend on task demands and participants' abilities (Ferreira & Swets, 2002; Snyder & Logan, 2014), and not necessarily an inherent difference between language and action sequencing.

Experiment 2: disyllabic tongue twisters

In Experiment 2 we aimed to create non-word stimuli that were more separable into their onset and offset components, both in their appearance on the screen and in production (temporal and linguistic distance). This was expected to encourage more distinct onset-offset parsing as in the finger fumbler materials, and therefore a more limited planning scope. If so, the error patterns in the speech task of Experiment 2 should bear closer resemblance to the manual task in Experiment 1, compared to the original speech task of Experiment 1. Because we were testing a hypothesis specifically about the speech task, we did not test Experiment 2 participants on the manual task. To get a measure of inter-transcriber agreement, we asked the main transcriber of Experiment 1 to transcribe data from five random subjects in Experiment 2. Onset error agreement was at 84%, similar to other speech error studies (Kittredge & Dell, 2016; Warker et al., 2009).

Method

Participants

Thirty-five native English speakers (women = 15, men = 20, age $M = 18.31$, $SD = 0.57$) participated in the experiment in exchange for course credit. An additional 5 participants were removed from analyses because of failure to follow task instructions (3) or technical difficulties (2). This smaller sample size was sufficient for Experiment 2 given that we used only the well-known tongue twister paradigm, which shows robust effects across studies (e.g., Acheson & MacDonald, 2009; Wilshire, 1999), as opposed to Experiment 1 in which we opted for a larger sample size to accommodate the novel manual task. Moreover, a power analysis using PANGEA

(Westfall, 2016) showed that for our design with an expected medium effect size (d = 0.45), our sample of 35 participants would provide 90% power. This expected effect size was determined based on the medium effect sizes found in the manual task of Exp 1, obtained from the odds ratio estimates in our model output. The study was approved by the UW-Madison IRB board, and all participants provided written informed consent.

Materials

We created a new set of tongue twisters where every non-word was comprised of two syllables appearing on the screen with a hyphen between them (e.g., *zlib-gipe*). Creation of these stimuli necessitated new phonemes and phoneme combinations than those used in Experiment 1, as every non-word now included more phonemes. Four options for the first syllable (*zib, zlib, fub, frub*) were combined with four options for the second syllable (*gipe, gine, keet, keem*), for a total of 16 disyllabic non-words. Trials were created by grouping non-words into sequences of four disyllabic non-words, creating 16 trial items for each trial type (TT, non-TT, Control) as in Experiment 1. Within a given trial, non-words varied only in the onset of the first syllable and the coda of the second syllable; see Table 4 for example stimuli. The onsets of the first syllable were either a single consonant or a consonant cluster, creating phonologically similarity (e.g., *zib-gipe, zlib-gipe*) as in Experiment 1.

Table 4

Examples of disyllabic tongue twister Stimuli (Experiment 2)

Trial Type	Item

TT	*zib-gipe, zlib-gine, zlib-gipe, zib-gine*
non-TT	*zib-gipe, zib-gine, zlib-gipe, zlib-gine*
Control	*zib-gipe, fub-keem, zlib-gine, frub-keet*

Because the two syllables within every non-word were designed to correspond to our onsets and offsets in the finger fumbler templates of Experiment 1, from here on we will refer to the entire first syllable as the "onset" (CVC or CCVC; e.g., *zlib* in *zlib-gipe*), and the entire second syllable as the "offset" (CVC; e.g., *gipe* in *zlib-gipe*).

Procedure

Participants sat in front of a screen in a sound-proof booth with a microphone for recording. First, they were familiarized with all non-words to be used in the experiment. Non-words were presented with a hyphen between the two syllables. Each non-word appeared on the screen for 3 seconds, separated by a fixation cross lasting 1s. To ensure that participants understood how to pronounce the disyllabic non-words, visual presentation of each non-word was accompanied by an audio recording of it being pronounced by a female native speaker of American English; and participants repeated the non-word after hearing it. Participants wore headsets for the familiarization phase and removed them before the test phase. The test phase procedure was identical to the procedure in Experiment 1, except that the moving rectangle cue allowed more time per non-word, approximately 810ms. The whole task lasted approximately 20 minutes. Responses were transcribed as in Experiment 1.

Results

Analyses were identical to that of Experiment 1, with "onset errors" here referring to errors on the first syllable within each non-word (e.g., *zlib* in *zlib-gipe*), and "offset errors" referring to errors on the second syllable (e.g., *gipe* in *zlib-gipe*). Word positions where participants did not produce any phoneme at all were removed from analyses (4%), and an additional .003% of the remaining observations were removed due to degraded audio files or indecipherable productions.

The mean onset error rate was $M = 0.15$, 95% CI [0.14, 0.16]. This rate is similar to the manual task in Experiment 1, and lower than the speech task in that study. Table 5 shows the mean proportion of onset and offset errors in Exp 2, and respective 95% confidence intervals, computed from the by-subject means. Figure 6 plots the corresponding model predictions for the proportion of errors at each Position as a function of trial Type. There was a significant interaction between the comparisons of Type [TT vs non-TT] and Position [Three vs Two], understood by the error rate decreasing between Positions Two and Three for TT trials, but not for non-TT ($b = -.81$, $SE = .37$, $z = -2.18$, $p < .05$). The opposite pattern – an increase for TT but decrease for non-TT – was seen between Positions Three and Four ($b = 1.05$, $SE = .38$, $z = 2.77$, $p < .01$). The [non-TT vs TT] difference between Positions Two and One was not significant ($p > .1$). None of the interactions between [Control vs TT] and Position were significant, ps > .1, but overall Control trials showed a significantly lower error rate than TT ($b = -.43$, $SE = .13$, $z = -3.25$, $p < .01$). Additionally, the error rate at Position Three was significantly lower than at Position Two ($b = -.31$, $SE = .15$, $z = -2.15$, $p < .05$), and the TT error rate was higher than non-TT ($b = .21$, $SE = .08$, $z = 2.55$, $p < .05$).

Table 5

Means and within-subject confidence intervals for onset and offset errors as a function of trial Type and Position in Exp 2

		Position							
		1		2		3		4	
Task	Trial Type	*M*	95% CI	*M*	95% CI	*M*	95% CI	*M*	95% CI
Onsets									
	TT	0.19	[0.15, 0.23]	0.19	[0.16, 0.23]	0.1	[0.06, 0.14]	0.19	[0.13, 0.24]
	non-TT	0.16	[0.13, 0.19]	0.13	[0.11, 0.16]	0.16	[0.13, 0.19]	0.09	[0.05, 0.12]
	Control	0.15	[0.12, 0.18]	0.16	[0.11, 0.2]	0.11	[0.08, 0.14]	0.16	[0.1, 0.21]
Offsets									
	TT	0.13	[0.1, 0.16]	0.14	[0.1, 0.17]	0.07	[0.03, 0.1]	0.1	[0.07, 0.13]
	non-TT	0.14	[0.12, 0.17]	0.12	[0.1, 0.15]	0.06	[0.04, 0.09]	0.08	[0.04, 0.13]
	Control	0.16	[0.13, 0.18]	0.15	[0.12, 0.18]	0.18	[0.13, 0.22]	0.13	[0.09, 0.18]

Note. M and 95% CI represent mean and 95% confidence intervals, respectively; computed from the by-subject means.

As apparent in Figure 6, the onset results suggest a decrease in error rates when onsets were repeated between adjacent positions. Notably, the difference between positions One and Two did not show the expected interaction with trial Type, but this seems to be driven by the high error rate for TT trials at position One, which may reflect a general difficulty in the transition between repetitions – in particular considering the eye-movement needed between the bottom of the screen back to the top. This difficulty may be exacerbated in the disyllabic speech task because of the longer strings of phonemes presented on screen.

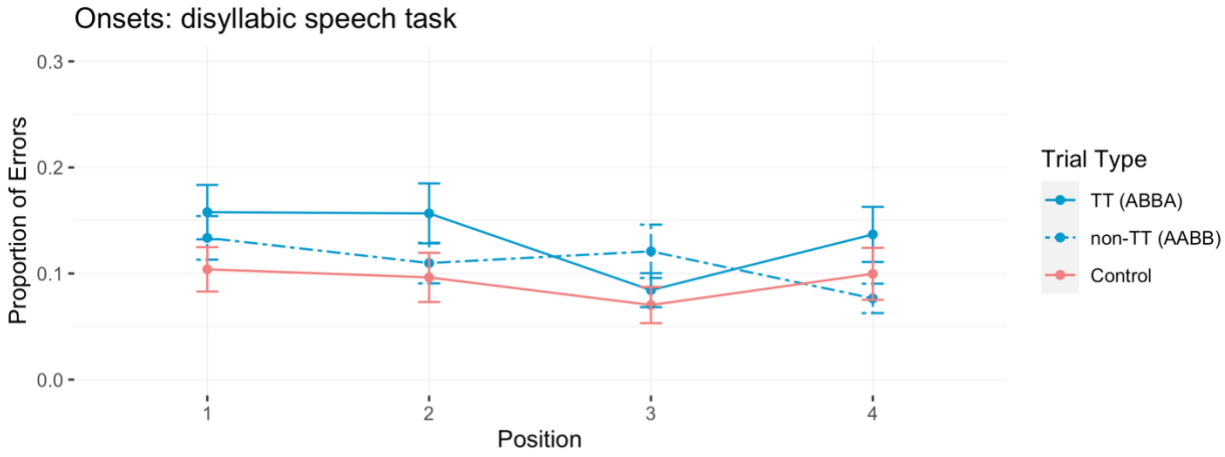

Figure 6. Model predictions for the proportion of onset errors at each word Position as a factor of trial Type in Experiment 2. Error bars show 95% confidence intervals.

The mean error rate at the offsets was $M = 0.12$, 95% CI [0.11, 0.14]. Figure 7 plots the model predictions for the proportion of offset errors at each Position as a function of trial Type. There was a significant interaction between the comparisons of Type [Control vs TT] and Position [Three vs Two], which could be understood by the error rate for Control increasing between Positions Two and Three, while the error rate for TT decreased ($b = .7$, $SE = .28$, $z =$

2.45, $p < .05$). Additionally, the error rate at Position Three was significantly higher than at Position Two ($b = -.27$, $SE = .13$, $z = -2.08$, $p < .05$), and the overall Control error rate was higher than the TT error rate ($b = .34$, $SE = .12$, $z = 2.71$, $p < .01$). Results from the offsets therefore show no difference between TT and non-TT trial types (as in Experiment 1), while the Control trials show an increase in errors between positions Two and Three – unseen in the other trial types.

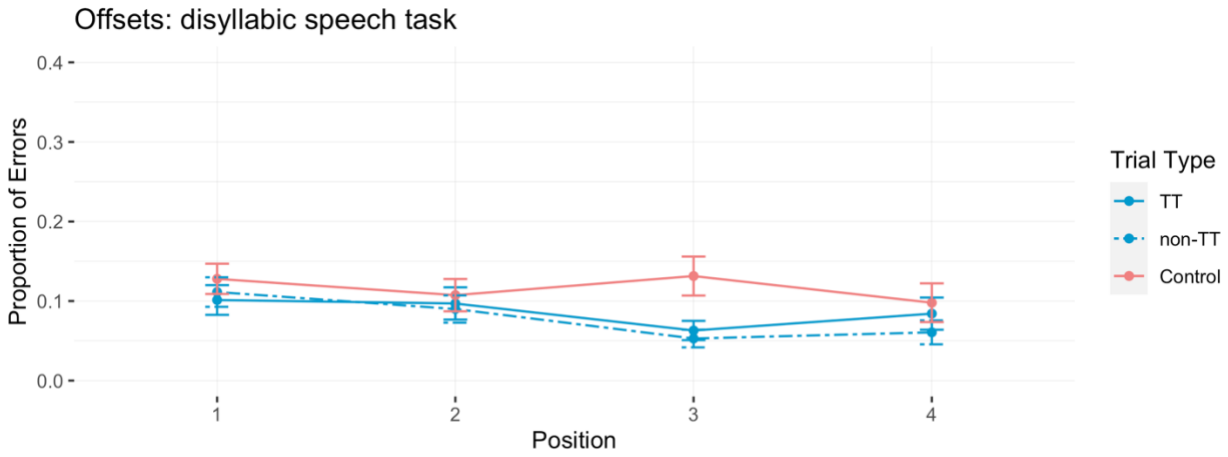

Figure 7. Model predictions for the proportion of offset errors at each word Position as a factor of trial Type in Experiment 2. Error bars show 95% confidence intervals.

Cross-experiment Comparisons

Onset results from the disyllabic tongue twisters in Experiment 2 appear similar to those from the manual task in Experiment 1: repeated onsets between adjacent positions led to facilitation. To further examine these similarities, we ran an additional analysis on combined data from the manual task in Exp 1 and the disyllabic speech task in Exp 2. We regressed onset errors on trial Type (TT/FF, non-TT/non-FF, Control), word Position (First, Second, Third, Fourth), and Experiment (Exp 1 manual task, Exp 2 disyllabic speech task), including their

interactions. Random effects included random intercepts for subjects and items, and a by-subject random slope for the interaction between Type and Position. We then examined the results of the three-way interactions, in particular for the contrasts of TT/FF vs non-TT/non-FF trial types. Finding a significant three-way interaction would suggest between-experiment differences in the interaction between Position and trial Type. In the case of no interactions, it might suggest similar cross-experiment effects. Note that we do not necessarily expect complete equivalence of effects between experiments, but rather similarity of error patterns – significant increases and decreases at the same critical positions, as a function of trial type. However, this cross-task analysis could shed light on which effect sizes differ between experiments.

There was a significant three-way interaction between trial Type [TT/FF vs non-TT/non-FF] and Experiment [manual vs disyllabic speech] for the contrast of positions [2-1], b = -0.75, SE = 0.22, z = -3.38, $p < .001$. This suggests that the Type by Position interaction effect was different between the two datasets at the contrast of positions [2-1]. This difference was expected based on the results of our initial analyses: in the manual task, there was a significant interaction between Type and Position for the contrast of positions [2-1], but not in the disyllabic speech task. Importantly, the three-way interaction was not significant at positions [3-2], b = 0.39, SE = 0.32, sz = 1.21, p = 0.23, nor at positions [4-3], b = 0.04, SE = 0.36, z = 0.1, p = .9. Although null results have limited interpretability, this lack of interaction is at least consistent with the similar patterns of onset errors we observed when comparing the two datasets in separate analyses.

In an additional post-hoc analysis to clarify the nature of the facilitation effects across tasks, we examined the difference in error rates between adjacent positions across all three tasks. For each task we calculated the by-subject mean differences in the proportion of onset errors between adjacent positions in both FF/TT and non-FF/TT trials (position One subtracted from

Two, Two subtracted from Three, Three subtracted from Four). Control trials were excluded because they had no repeated elements within the trial.

As expected, the difference in errors was negative when onsets were repeated ("no-switch" positions) – indicating facilitation, but positive when they were changed ("switch" positions). This pattern was found for the manual task (switch: $M = .04$, no-switch: $M = -.05$) and the Experiment 2 speech task (switch: $M = .04$; no-switch: $M = -.07$), and surprisingly, even for the Experiment 1 speech task (switch: $M = .02$; no-switch: $M = -.05$). To directly compare the effect across tasks, we ran a linear mixed-effects model regressing Difference on Switch (no-switch, switch), Task (Experiment 1 manual task, Experiment 1 disyllabic speech task, Experiment 2 monosyllabic speech task), the interaction between Switch and Task, and controlling for trial Type (FF/TT or non-FF/non-TT). The maximal random effects structure to converge included by-subject random slopes for Switch and trial Type. Results showed a significant effect of Switch ($b = .07$, $SE = .02$, $t = 3.95$, $p < .001$), and no other effects were significant (ps $> .05$); see Figure 8 for the plotted model predictions.

The significance of this Switch effect, above and beyond Task and trial Type, suggests that repeated onsets resulted in similar facilitation across the action and language tasks in both ABBA and AABB conditions. This pattern was not initially apparent in the monosyllabic speech task of Experiment 1, where our main analysis only showed a global trial type effect but no local interactions. However, this additional post-hoc analysis suggests that there may have been some facilitation between repeated onsets even in Experiment 1. We return to this point in the General Discussion, suggesting that repetition effects are modulated by several factors at different levels of the hierarchical plan. Together, the additional post-hoc analysis provides more evidence of plan reuse in all three of our tasks.

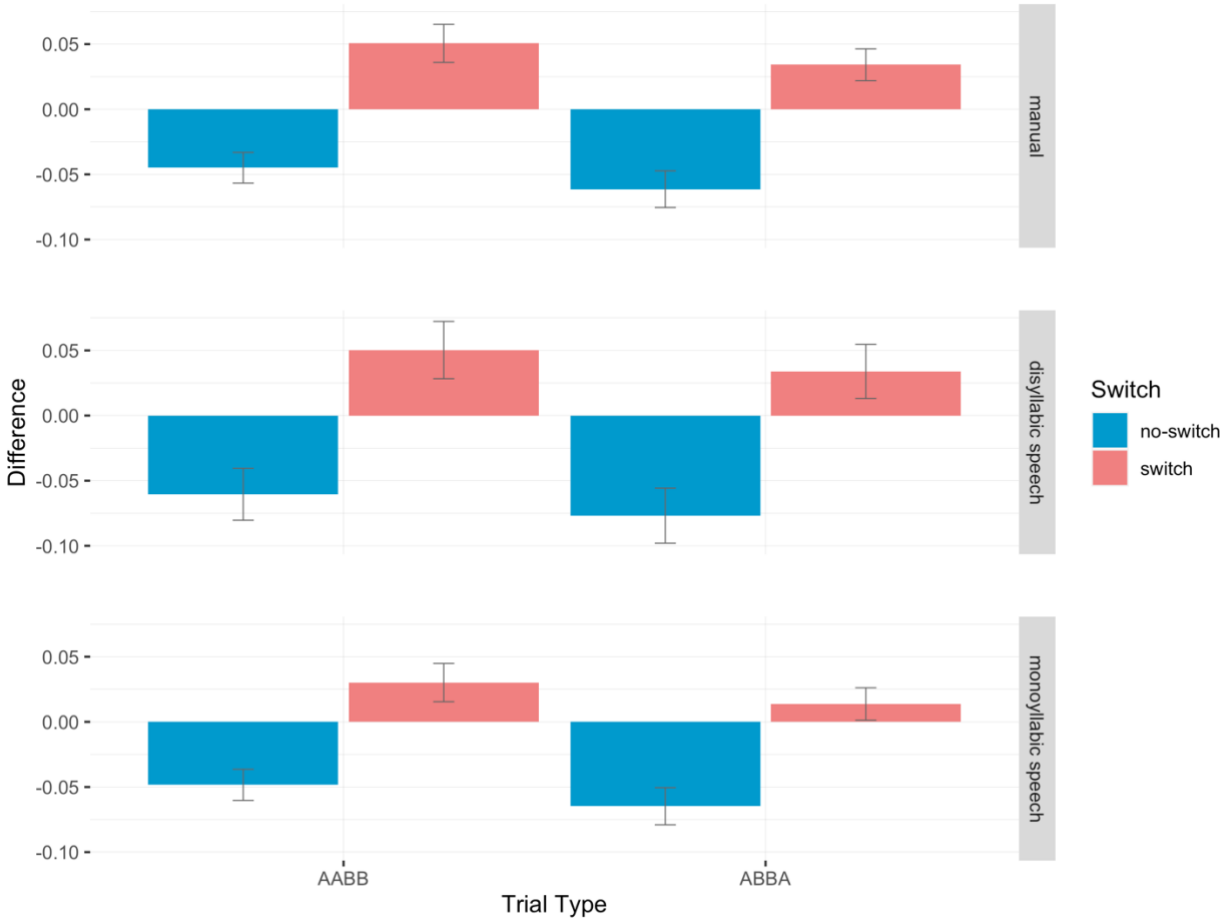

Figure 8. Model predictions for the difference in the proportion of onset errors between adjacent positions. Pink bars show "switch" positions – where the onset changed, while blue bars show "no-switch" positions – where the onset was repeated. Data are further separated by trial type – AABB pattern for non-TT/non-FF trials, ABBA pattern for TT/FF trials. Top panel shows Experiment 1 manual task, middle panel shows Experiment 2 disyllabic speech task, bottom panel shows Experiment 1 monosyllabic speech task. A negative difference indicates a decrease in error rates between adjacent positions, i.e., facilitation, while a positive difference indicates and increase in error rates. Error bars represent ±1 standard error.

Results from the offsets in Experiment 2 again show no difference between TT and non-TT trial types (as in both tasks of Experiment 1). However, the error rate for offsets in the Control trials shows a large increase between positions Two and Three, not seen in the other trial types. A potential explanation for the increase specifically at position Three could be that participants learn to expect a pattern of CDCD offsets because that is always the case for both TT and non-TT trial types. When they reach the third position in the Control trials, however, there is no repetition, so the unexpected divergence from the CDCD pattern leads to an increase in errors. This effect may be more influential in the language task because participants could hear their speech productions, and the CDCD rhyming alternation may have been particularly salient, making them more likely to develop expectations for that rhyming scheme. While this hypothesis could be investigated in future studies, the important point for the current study is that results from the offsets across all three tasks were not affected by the manipulated onset patterns (ABBA versus AABB). We will therefore concentrate our discussion on results from the onsets – the locus of our manipulation and the target of our hypotheses – which were clearly affected by the main manipulation.

Finally, following a reviewer's suggestion, we also examined the distribution of error subtypes across the three tasks. Although we did not have particular hypotheses about the distributions of errors, these have previously provided insights into the nature of language and action planning (Acheson & MacDonald, 2009; Botvinick & Bylsma, 2005; Dell, 1984; Wilshire, 1999). Error coding was automatized using a python script on the existing phoneme-by-phoneme transcriptions (or key-pressing data). For every onset error analyzed, the script checked whether the erroneous onset appeared elsewhere in the target sequence, and determined

whether that source was another onset, vowel, or coda. For errors where the source was another onset in the sequence, we then determined whether the error was preservatory – if the source position was before the error position; or anticipatory – if the source position was after the error, forthcoming in the target sequence.

Across all three tasks, when the erroneous onset appeared elsewhere in the target sequence, the error source was more likely to be another onset (manual task: 79%, monosyllabic speech: 98%, disyllabic speech: 99%) than a coda, while no error sources were vowels. This aligns with previous findings that errors tend to maintain their relative position within syllables (onset, vowel, coda), suggesting that participants learned the distributional properties of our non-familiar stimuli – in a way that is perhaps domain-general (Anderson & Dell, 2018; Fischer-Baum et al., 2021). Anticipation errors were far more common than perseveration errors in the speech tasks (monosyllabic speech: 74%, disyllabic speech: 86%), and only slightly more common in the manual task (55%). Anticipatory errors are typically seen as evidence for advance planning – suggesting that upcoming phonemes can intrude in speech because they are already part of the speech plan, while perseveratory errors are interpreted as difficulty inhibiting a prior component of the plan (Dell et al., 1997; Monaco et al., 2017). Moreover, the relative frequency of anticipation versus perseveration errors appears to depend on skill: increased task experience yields a higher proportion of anticipation errors (Dell et al., 1997). The higher rate of anticipatory errors in our speech tasks thus aligns with these findings, given that language is a much more practiced skill compared to our manual task. It also aligns with our interpretation of Exp 1 results: a wider planning scope in the speech task allows for more anticipatory errors, where the source of the error is in the upcoming speech plan.

General Discussion

In this study, we investigated similarities in language and motor action plans by comparing errors in parallel speech and manual tasks. A key feature of our design was comparing conditions that contained identical items and varied only in the order of the items in TT/FF versus non-TT/non-FF conditions. This property allowed us to home in on effects of serial ordering in conceptually similar speech and manual tasks.

We hypothesized that if sequence planning operates via similar principles in language and action, then parallel sequences in the two domains should yield similar error patterns. However, our manual and speech tasks in Experiment 1 displayed qualitative differences in error patterns. While the manual task showed a global trial type effect, the speech task showed local interaction effects from one position within a trial to the next. We suspected this was because of differences in the planning scope available for each task. Specifically, the clear division between onsets and offsets in the manual task likely encouraged limited planning ahead, resulting in local facilitation effects between adjacent units when the onset was repeated. In contrast, the continuous production of monosyllabic non-words in the speech task likely encouraged planning of the entire sequence, introducing interference effects from the global ABBA sequence structure. In attempt to better equate the planning scope in our cross-domain comparison, we ran a follow-up speech task (Experiment 2) using disyllabic non-words with greater separation between onsets and offsets, both visually (appearance on the screen) and in production (temporal and linguistic distance). Results of this disyllabic speech task resembled those of the manual task: repetition of an onset between adjacent positions led to facilitation.

The parallel facilitation effects in our language and action tasks indicate that in both domains, production planning is facilitated when a new plan shares features with a prior plan (Koranda et al., 2020; Lashley, 1951; Logan, 2020; MacDonald, 2013; Rosenbaum et al., 1986; Rosenbaum & Saltzman, 1984). Our results may also shed light on the level of detail within plans. We found the benefit for plan reuse can occur between segments (onsets or offsets) within nonwords or templates. That is, what caused facilitation was not just simple repetition of an entire non-word or template. It was repetition of the onset segment from one nonword/template to the next, even though there was an intervening offset segment that did not repeat. This suggests that these onset and offset segments could be thought of as separate subplans within the larger non-word/template unit, which itself is a subplan within the larger sequence of four non-words/templates – giving rise to the complex hierarchical plan.

The ability to separate and substitute subplans within an action plan is a key feature of hierarchical planning (Lashley, 1951; Rosenbaum et al., 2007) that would not be possible if sequencing was based on automatic associations between adjacent units (Washburn, 1916; Watson, 1920). It is arguably owing to this hierarchical structure that plan reuse can function as an efficient planning principle: if plans had no internal structure, then a repetition benefit would arise only if the entire sequence was repeated. By contrast, in a hierarchical plan, there can be benefits from repetition of subplans such as repeating onsets across nonwords/templates in our studies. This point underscores the reasoning behind our hypothesis that language and action plans should draw on similar sequencing principles: despite the differences between domains, both our speech and motor tasks involve hierarchical planning, and so both should see similar benefits of subplan repetition.

Repetition Costs and Benefits

The effects of repetition, however, are modulated by several other factors (Crump & Logan, 2010; Fournier et al., 2014; Jaeger et al., 2012; Logan, 2020; Snyder & Logan, 2014), and sometimes present conflicting results. Previous research on phonological onset overlap between prime and target words has shown both facilitation (e.g., Meyer & Schriefers, 1991; Smith & Wheeldon, 2004) and inhibition (e.g., Sevald & Dell, 1994; Sullivan & Riffel, 1999) effects, depending on the temporal interval between presentation of the prime and the target word (Jaeger et al., 2012). Similar disparities have been found between typing and speaking studies (Snyder & Logan, 2014), leading to the suggestion that task demands, and in particular task speed and the producer's skill, will affect the strategy chosen for the task and the resulting outcome (Logan, 2020). This echoes our own motivation for Experiment 2, in which we found that task demands and the producer's skill can affect the scope of the plan and the resulting error pattern. Indeed, prior research has shown that the planning scope is under some strategic control and varies with task demands (Ferreira & Swets, 2002), and we suspected it would modulate the effects of repetition in our sequences.

The comparison of our monosyllabic and disyllabic speech tasks provides the most direct evidence that repetition effects are modulated by planning scope. Although both tasks were in the language domain and used the same sequence structures, the shorter non-words in the monosyllabic task allow for the planning scope to cover a wider range of the sequence. In that case, although there might be facilitation between adjacent repeated onsets, at another level of the hierarchical plan – the global sequence of four words – the order of onsets is reversed across pairs of non-words (the ABBA pattern). This reversal requires editing of the prior plan which can cause interference (Rosenbaum et al., 1986), potentially obscuring the facilitation effects of

repetition. In contrast, the disyllabic speech task had longer non-words intended to limit the planning scope, and results showed facilitation effects between repeated onsets in adjacent positions. Notably, there was still an overall higher error rate for TT than the non-TT trials, as in the monosyllabic speech task. It is likely that even in the disyllabic speech task there was still some representation of the global ABBA sequence (given the auditory feedback of speech and participants' extensive reading experience), but not strong enough to obscure the local facilitation effects.

The different outcomes of repetition also reflect a certain tension between facilitation and interference effects of plan reuse (MacDonald, 2013): reusing a production plan can minimize computations and improve performance, but significant changes to the plan will come at a cost (Rosenbaum et al., 1986). The factors that determine this balance warrant further research. Our results suggest that the hierarchical nature of production plans must be considered in future work, because repetition at one level of the hierarchy might result in a difficult change at another level of the hierarchy. These dynamics can be explored by manipulating task demands, as we did here with limiting the planning scope for the disyllabic tongue twisters. In fact, manipulating task demands is a crucial step for cross-domain comparisons, where there are stark differences between domains in task demands and participants' prior experience (Koranda et al., 2020). By better aligning the tasks across domains, it becomes possible to uncover parallels such as the benefits of plan reuse that we report here.

Constraints on Generality

As noted earlier, the error patterns in our tasks are likely to vary across materials and task demands. This is already evident in the comparison of the language tasks in Experiments 1 and 2

that yielded different error patterns when we varied the difficulty of the tongue twisters, specifically, by varying the length of the non-words and whether they conform to English phonotactic constraints. But these changes were in a predictable direction as hypothesized when we created the new stimuli, suggesting the principles we outlined here would make it possible to predict what effects would emerge for a given set of stimuli for at least some future variations in stimuli. In addition, our sample was a convenience sample of participants in a Midwestern university community, all native English speaking young adults, and likely with substantial typing experience (though not in a key-pressing task like ours). We would therefore expect our results to generalize to similar samples, but they would likely differ with speakers of other languages, or people with no typing or key-pressing experience. Throughout the discussion we underscore several factors that could affect the strategy used for performance in our tasks and the resulting error patterns. We argue that manipulating these factors in future research could uncover which planning principles guide production, and which contexts elicit varying production strategies. We have no reason to believe that the results depend on other characteristics of the participants, materials, or context.

Conclusion

Our findings add to a limited number of studies showing parallels between language and action production, whether in learning patterns (Anderson & Dell, 2018), priming effects (Koranda et al., 2020; Lebkuecher et al., 2022), behavioral performance (Rosenbaum et al., 1986), computational modeling (Logan, 2020), or neural activation (Casado et al., 2018; Cona & Semenza, 2017). Cross-domain comparisons are particularly challenging given that each domain has unique output forms, different amounts of experience participants bring to the task, and

potentially different available strategies (Koranda et al., 2020; Lebkuecher et al., 2022). The questions are therefore not simply whether we will observe similar or different outcomes for parallel tasks, but rather *which* principles guide strategy and production choices, and what interactions they exhibit above and beyond particularities in output or skill. Identifying such principles – such as plan reuse discussed here – allows targeting them directly in experimental manipulations, for a more focused comparison of sequencing behavior across domains.

Context of the Research

This study contributes to research investigating parallels in language and action planning (e.g., Koranda et al., 2020; Lebkuecher et al., 2022). The idea that both language production and motor action require serial ordering and a hierarchical plan has led several researchers to suggest that some aspects of planning and ordering in both domains may be domain general, or at least follow similar planning principles (Koranda et al., 2020; Lashley, 1951; Lebkuecher et al., 2022; Logan, 2020; MacDonald, 2013; Rosenbaum et al., 1986). The goal of our research is to make direct comparisons between analogous manual and speech tasks, in order to uncover cross-domain similarities or differences in planning. Comparisons of planning in the two domains not only contributes to theorizing within each domain but could also increase our understanding of which domain-general principles guide planning across different behaviors.

Public Significance Statement

This study suggests that there are cognitive similarities in the way people prepare language productions (e.g., words, sentences) and motor actions (e.g., movements, key presses).

In particular, it highlights the benefit of repetitions in both speech and action – it is easier to produce a component that has been produced before, even if it appears within a new sequence of words or actions.

Author contributions statement

A. E. Gussow: Conceptualization, Methodology, Software, Investigation, Analysis, Writing—original draft. D. J. Weiss: Conceptualization, Writing—review and editing. M. C. MacDonald: Conceptualization, Methodology, Funding acquisition, Supervision, Writing-review and editing.

Acknowledgments

This work was supported by NSF Grant number 1849236 and a UW2020 grant from the University of Wisconsin Office of the Vice Chancellor for Research and Graduate Education, M.C. MacDonald, PI. We thank research assistants at the Language and Cognitive Neuroscience Lab for help with data collection, transcription, and coding.

References

Acheson, D. J., & MacDonald, M. C. (2009). Twisting tongues and memories: Explorations of the relationship between language production and verbal working memory. *Journal of Memory and Language*, *60*(3), 329–350. https://doi.org/10.1016/j.jml.2008.12.002

Anderson, N. D., & Dell, G. S. (2018). The role of consolidation in learning context-dependent phonotactic patterns in speech and digital sequence production. *Proceedings of the National Academy of Sciences*, *115*(14), 3617–3622. https://doi.org/10.1073/pnas.1721107115

Badre, D. (2008). Cognitive control, hierarchy, and the rostro–caudal organization of the frontal lobes. *Trends in Cognitive Sciences*, *12*(5), 193–200. https://doi.org/10.1016/j.tics.2008.02.004

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00328

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bernstein, N. A. (1967). *The co-ordination and regulation of movements.* Pergamon. https://www-worldcat-org.ezproxy.library.wisc.edu/title/co-ordination-and-regulation-of-movements/oclc/598328285

Beth Warriner, A., & Humphreys, K. R. (2008). Short Article: Learning to Fail: Reoccurring Tip-of-the-Tongue States. *Quarterly Journal of Experimental Psychology*, *61*(4), 535–542. https://doi.org/10.1080/17470210701728867

Botvinick, M. M., & Bylsma, L. M. (2005). Distraction and action slips in an everyday task: Evidence for a dynamic representation of task context. *Psychonomic Bulletin & Review*, *12*(6), 1011–1017. https://doi.org/10.3758/BF03206436

Carota, F., & Sirigu, A. (2008). Neural Bases of Sequence Processing in Action and Language. *Language Learning*, *58*(s1), 179–199. https://doi.org/10.1111/j.1467-9922.2008.00470.x

Casado, P., Martín-Loeches, M., León, I., Hernández-Gutiérrez, D., Espuny, J., Muñoz, F., Jiménez-Ortega, L., Fondevila, S., & de Vega, M. (2018). When syntax meets action: Brain potential evidence of overlapping between language and motor sequencing. *Cortex*, *100*, 40–51. https://doi.org/10.1016/j.cortex.2017.11.002

Cona, G., & Semenza, C. (2017). Supplementary motor area as key structure for domain-general sequence processing: A unified account. *Neuroscience & Biobehavioral Reviews*, *72*, 28–42. https://doi.org/10.1016/j.neubiorev.2016.10.033

Cooper, R. P., & Shallice, T. (2006). Hierarchical schemas and goals in the control of sequential behavior. *Psychological Review*, *113*(4), 887–916. https://doi.org/10.1037/0033-295X.113.4.887

Croot, K., Au, C., & Harper, A. (2010). Prosodic structure and tongue twister errors. *Laboratory Phonology*, *10*, 433–461.

Crump, M. J. C., & Logan, G. D. (2010). Hierarchical control and skilled typing: Evidence for word-level control over the execution of individual keystrokes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*(6), 1369–1380. https://doi.org/10.1037/a0020696

Davidson, L. (2006). Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics*, *34*(1), 104–137. https://doi.org/10.1016/j.wocn.2005.03.004

Davidson, L., Martin, S., & Wilson, C. (2015). Stabilizing the production of nonnative consonant clusters with acoustic variability. *The Journal of the Acoustical Society of America*, *137*(2), 856–872. https://doi.org/10.1121/1.4906264

Dell, G. S. (1984). Representation of serial order in speech: Evidence from the repeated phoneme effect in speech errors. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *10*(2), 222–233. https://doi.org/10.1037//0278-7393.10.2.222

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review*, *104*(1), 123–147. https://doi.org/10.1037/0033-295X.104.1.123

Dell, G. S., Kelley, A. C., Hwang, S., & Bian, Y. (2021). The adaptable speaker: A theory of implicit learning in language production. *Psychological Review*, *128*(3), 446–487. https://doi.org/10.1037/rev0000275

Dell, G. S., & Reich, P. A. (1981). Stages in sentence production: An analysis of speech error data. *Journal of Verbal Learning and Verbal Behavior*, *20*(6), 611–629. https://doi.org/10.1016/S0022-5371(81)90202-4

Drake, C., & Palmer, C. (2000). Skill acquisition in music performance: Relations between planning and temporal control. *Cognition*, *74*(1), 1–32. https://doi.org/10.1016/S0010-0277(99)00061-X

Ferreira, F., & Swets, B. (2002). How Incremental Is Language Production? Evidence from the Production of Utterances Requiring the Computation of Arithmetic Sums. *Journal of Memory and Language*, *46*(1), 57–84. https://doi.org/10.1006/jmla.2001.2797

Fischer-Baum, S., Warker, J. A., & Holloway, C. (2021). Learning phonotactic-like regularities in immediate serial recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *47*(1), 129–146. https://doi.org/10.1037/xlm0000815

Fournier, L. R., Gallimore, J. M., Feiszli, K. R., & Logan, G. D. (2014). On the importance of being first: Serial order effects in the interaction between action plans and ongoing actions. *Psychonomic Bulletin & Review*, *21*(1), 163–169. https://doi.org/10.3758/s13423-013-0486-0

Gaskell, M. G., Warker, J., Lindsay, S., Frost, R., Guest, J., Snowdon, R., & Stackhouse, A. (2014). Sleep Underpins the Plasticity of Language Production. *Psychological Science*, *25*(7), 1457–1465. https://doi.org/10.1177/0956797614535937

Jaeger, F. T. F., Furth, K., & Hilliard, C. (2012). Incremental Phonological Encoding during Unscripted Sentence Production. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00481

Kember, H., Connaghan, K., & Patel, R. (2017). Inducing speech errors in dysarthria using tongue twisters. *International Journal of Language & Communication Disorders*, *52*(4), 469–478. https://doi.org/10.1111/1460-6984.12285

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Sciences*, *22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Kittredge, A. K., & Dell, G. S. (2016). Learning to speak by listening: Transfer of phonotactics from perception to production. *Journal of Memory and Language*, *89*, 8–22. https://doi.org/10.1016/j.jml.2015.08.001

Koranda, M. J., Bulgarelli, F., Weiss, D. J., & MacDonald, M. C. (2020). Is Language Production Planning Emergent From Action Planning? A Preliminary Investigation. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.01193

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior* (pp. 112–131). New York, NY: Wiley.

Lebkuecher, A. L., Schwob, N., Kabasa, M., Gussow, A. E., MacDonald, M. C., & Weiss, D. J. (2022). Hysteresis in motor and language production. *Quarterly Journal of Experimental Psychology*, 17470218221094568. https://doi.org/10.1177/17470218221094568

Lindsley, J. R. (1975). Producing simple utterances: How far ahead do we plan? *Cognitive Psychology*, *7*(1), 1–19. https://doi.org/10.1016/0010-0285(75)90002-X

Logan, G. D. (2020). *Serial order in perception, memory, and action.* Psychological Review; US: American Psychological Association. https://doi.org/10.1037/rev0000253

MacDonald, M. C. (2013). How language production shapes language form and comprehension. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00226

MacDonald, M. C. (2016). Speak, act, remember: The language-production basis of serial order and maintenance in verbal memory. *Current Directions in Psychological Science*, *25*(1), 47–53. https://doi.org/10.1177/0963721415620776

Meyer, A. S., & Schriefers, H. (1991). Phonological facilitation in picture-word interference experiments: Effects of stimulus onset asynchrony and types of interfering stimuli.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *17*(6), 1146–1160. https://doi.org/10.1037/0278-7393.17.6.1146

Monaco, E., Pellet Cheneval, P., & Laganaro, M. (2017). Facilitation and interference of phoneme repetition and phoneme similarity in speech production. *Language, Cognition and Neuroscience*, *32*(5), 650–660. https://doi.org/10.1080/23273798.2016.1257730

Nooteboom, S., & Quené, H. (2015). Word-onsets and Stress Patterns: Speech Errors in a Tongue-Twister Experiment. *In ICPhS.*

Norman, D. A. (1981). Categorization of action slips. *Psychological Review*, *88*(1), 1–15. https://doi.org/10.1037/0033-295X.88.1.1

Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, *106*(1), 528–537. https://doi.org/10.1016/j.cognition.2007.02.006

Ou, J., & Law, S.-P. (2017). Cognitive basis of individual differences in speech perception, production and representations: The role of domain general attentional switching. *Attention, Perception, & Psychophysics*, *79*(3), 945–963. https://doi.org/10.3758/s13414-017-1283-z

Page, M. P. A., Madge, A., Cumming, N., & Norris, D. G. (2007). Speech errors and the phonological similarity effect in short-term memory: Evidence suggesting a common locus☆. *Journal of Memory and Language*, *56*(1), 49–64. https://doi.org/10.1016/j.jml.2006.09.002

Pinet, S., & Nozari, N. (2018). "Twisting fingers": The case for interactivity in typed language production. *Psychonomic Bulletin & Review*, *25*(4), 1449–1457. https://doi.org/10.3758/s13423-018-1452-7

Rosenbaum, D. A., Cohen, R. G., Jax, S. A., Weiss, D. J., & van der Wel, R. (2007). The problem of serial order in behavior: Lashley's legacy. *Human Movement Science*, *26*(4), 525–554. https://doi.org/10.1016/j.humov.2007.04.001

Rosenbaum, D. A., & Saltzman, E. (1984). A Motor-Program Editor. In W. Prinz & A. F. Sanders (Eds.), *Cognition and Motor Processes* (pp. 51–61). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-69382-3_4

Rosenbaum, D. A., Weber, R. J., Hazelett, W. M., & Hindorff, V. (1986). The parameter remapping effect in human performance: Evidence from tongue twisters and finger fumblers. *Journal of Memory and Language*, *25*(6), 710–725. https://doi.org/10.1016/0749-596X(86)90045-8

Sevald, C. A., & Dell, G. S. (1994). The sequential cuing effect in speech production. *Cognition*, *53*(2), 91–127. https://doi.org/10.1016/0010-0277(94)90067-1

Smith, M., & Wheeldon, L. (2004). Horizontal Information Flow in Spoken Sentence Production. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*(3), 675–686. https://doi.org/10.1037/0278-7393.30.3.675

Snyder, K. M., & Logan, G. D. (2014). The problem of serial order in skilled typing. *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1697–1717. https://doi.org/10.1037/a0037199

Starkes, J. L., Deakin, J. M., Lindley, S., & Crisp, F. (1987). Motor versus Verbal Recall of Ballet Sequences by Young Expert Dancers. *Journal of Sport Psychology*, *9*(3), 222–230. https://doi.org/10.1123/jsp.9.3.222

Stemberger, J. P. (1982). Syntactic errors in speech. *Journal of Psycholinguistic Research*, *11*(4), 313–345. https://doi.org/10.1007/BF01067585

Stemberger, J. P. (2009). Preventing perseveration in language production. *Language and Cognitive Processes*, *24*(10), 1431–1470. https://doi.org/10.1080/01690960902836624

Sullivan, M. P., & Riffel, B. (1999). The nature of phonological encoding during spoken word retrieval. *Language and Cognitive Processes*, *14*(1), 15–45. https://doi.org/10.1080/016909699386365

Taylor, C. F., & Houghton, G. (2005). Learning artificial phonotactic constraints: Time course, durability, and relationship to natural constraints. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *31*(6), 1398–1416. https://doi.org/10.1037/0278-7393.31.6.1398

Vitevitch, M. S., & Luce, P. A. (2004). A Web-based interface to calculate phonotactic probability for words and nonwords in English. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 481–487. https://doi.org/10.3758/BF03195594

Vogel, E. K., & Awh, E. (2008). How to Exploit Diversity for Scientific Gain: Using Individual Differences to Constrain Cognitive Theory. *Current Directions in Psychological Science*, *17*(2), 171–176. https://doi.org/10.1111/j.1467-8721.2008.00569.x

Wardlow, L. (2013). Individual differences in speakers' perspective taking: The roles of executive control and working memory. *Psychonomic Bulletin & Review*, *20*(4), 766–772. https://doi.org/10.3758/s13423-013-0396-1

Warker, J. A. (2013). Investigating the Retention and Time-Course of Phonotactic Constraint Learning From Production Experience. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(1). https://doi.org/10.1037/a0028648

Warker, J. A., & Dell, G. S. (2006). Speech errors reflect newly learned phonotactic constraints. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *32*(2), 387–398. https://doi.org/10.1037/0278-7393.32.2.387

Warker, J. A., Xu, Y., Dell, G. S., & Fisher, C. (2009). Speech errors reflect the phonotactic constraints in recently spoken syllables, but not in recently heard syllables. *Cognition*, *112*(1), 81–96. https://doi.org/10.1016/j.cognition.2009.03.009

Washburn, M. F. (1916). *Movement and Mental Imagery: Outlines of a Motor Theory of the Complexer Mental Processes*. Houghton Mifflin.

Watson, J. B. (1920). IS THINKING MERELY ACTION OF LANGUAGE MECHANISMS1? (V.). *British Journal of Psychology. General Section*, *11*(1), 87–104. https://doi.org/10.1111/j.2044-8295.1920.tb00010.x

Westfall, J. (2016). *PANGEA: Power ANalysis for GEneral Anova designs (working paper)*.

Wilshire, C. E. (1999). The "Tongue Twister" Paradigm as a Technique for Studying Phonological Encoding. *Language and Speech*, *42*(1), 57–82. https://doi.org/10.1177/00238309990420010301

Wolpert, D. M., & Landy, M. S. (2012). Motor control is decision-making. *Current Opinion in Neurobiology*, *22*(6), 996–1003. https://doi.org/10.1016/j.conb.2012.05.003