Running head: ERROR BIASES IN INNER SPEECH

Error Biases in Inner and Overt Speech: Evidence from Tonguetwisters

Martin Corley Paul H. Brocklehurst H. Susannah Moat

Psychology, PPLS, University of Edinburgh

Martin Corley
Psychology
School of Philosophy, Psychology, and Language Sciences
University of Edinburgh
Edinburgh EH8 9JZ, UK

(tel) +44 131 650 6682; (fax) +44 131 650 3461; Martin.Corley@ed.ac.uk

Abstract

In order to compare the properties of inner and overt speech, Oppenheim and Dell (2008) counted participants' self-reported speech errors when reciting tonguetwisters either overtly or silently, and found a bias towards substituting phonemes which resulted in words in both conditions, but a bias towards substituting similar phonemes only when speech was overt. Here, we report three experiments that revisit their conclusion, that inner speech remains underspecified at the subphonemic level, which they simulated within an activation-feedback framework. In two experiments, participants recited tonguetwisters which could result in the errorful substitutions of similar or dissimilar phonemes to form real words or nonwords. Both experiments included an auditory masking condition, to gauge the possible impact of loss of auditory feedback on the accuracy of self-reporting of speech errors. In Experiment 1 the stimuli were composed entirely from real words, whereas in Experiment 2 half of the tokens used were nonwords. Although masking did not have any effects, participants were more likely to report substitutions of similar phonemes in both experiments, in inner as well as overt speech. This pattern of results was confirmed in a third experiment using the real word materials from Oppenheim and Dell (in press). In addition to these findings, a lexical bias effect found in Experiments 1 and 3 disappeared in Experiment 2. Our findings support a view in which plans for inner speech are indeed specified at the feature level, even when there is no intention to articulate words overtly, and in which editing of the plan for errors is implicated.

Keywords: Inner Speech; Speech Errors; Phonemic Similarity; Lexical Bias; Tonguetwisters

Error Biases in Inner and Overt Speech: Evidence from Tonguetwisters

Inner speech plays a key role in a variety of different cognitive activities, including writing, personal thought, reasoning and memorization (e.g., Baddeley & Hitch, 1974; Ellis, 1988; Sokolov, 1972; Vygotsky, 1986). Although the intent to articulate is not a prerequisite of inner speech (see MacKay, 1992; Sokolov, 1972, for detailed reviews), subjective accounts suggest that it frequently resembles overt speech, in that it it appears to be sound-based and can vary in tempo, pitch and rhythm (MacKay, 1992). Nonetheless, it has been suggested that inner speech without articulation is often attenuated at the surface level, lacking phonological (Dell & Repka, 1992; Oppenheim & Dell, 2008, in press) or phonetic (Wheeldon & Levelt, 1995) detail.

This conclusion appears to be supported in a recent study by Oppenheim and Dell (2008, Experiment 2). In this study, participants were asked to repeat a series of four-word tonguetwisters aloud, and report each occasion that they made an error. The tonguetwisters were manipulated such that an onset substitution would result in either a word or a nonword. The experiment replicated the well-known lexical bias effect (Baars, Motley, & MacKay, 1975; Dell, 1986; Hartsuiker, Corley, & Martensen, 2005): Participants were more likely to make errors where a real word ensued. In addition to the lexical manipulation, the onset phonemes of subsequent words differed by either one or two phonological features, and participants were more likely to substitute phonemes that differed by one feature, showing that phonological detail affected the production of errors (cf. Dell & Reich, 1981; Levitt & Healy, 1985; Nooteboom, 2005a, 2005b). This pattern of findings was simulated by Oppenheim and Dell using activation feedback between the levels of the speech production system. Phonemes activated in error feed back to representations for words, but not to nonwords, since the latter do not occur in the mental lexicon, promoting the likelihood of uttering words in error. Where features are activated by phonemes, representations for phonemes accrue feedback activation to the extent that they share features with the intended phoneme, promoting the likelihood of uttering similar phonemes in error (see Dell, 1986, 1988, for detailed explanations of the role of activation feedback in speech errors).

In order to investigate the properties of inner speech, Oppenheim and Dell included a second condition in which participants were asked to repeat the tonguetwisters silently, reporting the errors they detected in their inner speech. In this condition, the lexical bias found for overt speech was replicated. However, there were no effects of phonological detail: participants were no more likely to report substitutions of phonemes that differed from each other by one feature than of those that differed by two. Oppenheim and Dell (2008) concluded that the lack of a phonological similarity effect in the inner speech condition could most likely be attributed to the fact that inner speech is impoverished at the feature level, in which case there would be no feedback of activation from feature to phoneme levels of representation, and thus no bottom-up activation of competitor phonemes.

However, two issues are raised by these conclusions and are investigated in the present paper. The first is that the view that inner speech is impoverished requires an assumption that participants are able to attend to their own inner speech (Levelt, 1983, 1989), and that they can successfully detect and report all of the errors they make under these circumstances. As Oppenheim and Dell (2008) acknowledge, a plausible alternative account of the differences between conditions is that participants are better able to *perceive* certain types of error in speech that is overt. In fact, Oppenheim and Dell's participants self-reported marginally more errors when speaking aloud (averaged across two analyses, 54% of participants' reports were of errors in overt speech). This difference is in line with previous research (Postma & Noordanus, 1996), which also showed that more errors at the phonemic level were detected when speech was overt. Where there is no overt speech, single-feature errors may be particularly susceptible to underreporting, because there is no motor feedback to confirm that an error has been made (e.g., Borden, 1979; Lackner & Tuller, 1979). On this interpretation, inner speech could be routinely specified at a featural level, irrespective of whether or not there is an intention to articulate it overtly.

The second issue raised by Oppenheim and Dell's conclusions is that they simulate the patterns of errors reported in terms of feedback between levels of representation. An alternative view is that speakers monitor their speech plans and edit out errors that would otherwise have been produced (Levelt, 1983, 1989; Levelt, Roelofs, & Meyer, 1999), such as nonwords (Baars et al., 1975). This view is based on the premise that speakers are able to detect, and covertly repair, a phonological speech error before articulation. Unlike feedback, the editing of the speech plan is adaptive (Baars et al., 1975; Hartsuiker et al., 2005): Accidentally-produced nonwords are only filtered out where the demand characteristics of the task in hand include a requirement to produce words (see Hartsuiker, 2006, for further discussion).

In the present paper, we address both of these issues. In two experiments, participants produce tonguetwisters either overtly or silently, either with or without auditory masking. If Oppenheim and Dell's finding that there is no phonemic similarity bias in inner speech reflects underspecification of inner speech, we expect the phonemic similarity effect to diminish only in the silent conditions, where there is no overt articulation. On the other hand, if it reflects underreporting of errors that cannot be heard, we expect the masked and silent conditions to pattern together. To investigate the role of activation feedback, we vary the demand characteristics that an editor would be sensitive to across experiments: in Experiment 1 participants produce tonguetwisters that consist entirely of real words; in Experiment 2, 50% of the 'words' in the material set are nonwords. If activation feedback provides the best account of the error patterns reported, a clear lexical bias should be observed in each experiment; differences in the lexical bias between experiments would suggest that editing plays a role. A third experiment replicates the results of Experiment 1 using the tonguetwisters originally used by Oppenheim and Dell (in press), based on those used by Oppenheim and Dell (2008) with one substituted item.

Experiment 1

Experiment 1 was closely modelled on Oppenheim and Dell's (2008) study. Participants produced a series of four-word tonguetwisters, designed such that substituting the onsets of the third words with the phonemically similar or dissimilar fourth-word onsets would result in either words or nonwords. We anticipated that, as for Oppenheim and Dell, participants would report more errors which resulted in real words, and more substitutions of similar than of dissimilar phonemes. In a silent speech condition, we expected the lexical bias to remain, but the effect of phonemic similarity to disappear.

In our experiment, participants produced half of the tonguetwisters (silently or overtly) under conditions of auditory masking. If, as Oppenheim and Dell (2008) claim, the differences between overt and silent conditions were because inner speech remains underspecified when there is no intention to speak aloud, we expect there to be no differences between the numbers of self-reported errors in the masked and unmasked conditions. If, on the other hand, the lack of phonemic similarity effects in inner speech could be attributed to the difficulty of detecting single-feature errors in the absence of auditory feedback, the phonemic similarity effect could be expected to diminish for overt speech under conditions of auditory masking.

Method

Participants. Thirty-two native speakers of English were recruited from the Edinburgh student population for course credit. Ages ranged from 18 to 32. In this and in the following experiment, participants reported no speech, language, hearing or visual impairments.

Materials. Forty-eight matched sets of four-word tonguetwister sequences were generated following Oppenheim and Dell (2008). Candidate sequences were generated automatically from lists of CVC(C) words with CELEX frequencies greater than 1 per million (Baayen, Piepenbrock, & Gulikers, 1995). Pronunciations were checked using BEEP (Robinson, 1997) and also by hand. Words with ambiguous pronunciations were not used.

Each set comprised four sequences. To induce onset-phoneme substitution errors, the onset consonants of each sequence followed an ABBA pattern; however the onsets of words 1 and 4 (the 'A' words) varied in each set, whereas those of words 2 and 3 did not change. In two of the four sequences the onsets of the 'A' words were phonologically *similar* to those of the 'B's, differing by a single feature. In the other two they were *dissimilar*, differing by two or more features.

In addition to the phonological similarity manipulation, the tonguetwister sequences were also manipulated within each set to vary the lexicality of error outcomes. This was achieved by varying the coda of word 3 (traditionally the most susceptible to onset substitutions in ABBA tonguetwisters, e.g., Wilshire, 1999) such that, if the onset of word 3 was substituted with an 'A' onset, the outcome would either be a *word* or a *nonword.*

A sample set of four tonguetwister sequences is shown in Table 1. In this example, /k/ differs from /p/ by one feature but from /b/ by two; substituting the onset of word 3 with that of word 4 would yield *patch* or *batch* in the word outcome conditions, but *pab* or *bab* for nonword outcomes. Words in position 3 were frequency-matched across the experiment: mean log frequency for the word outcome conditions was 0.97, and for nonword outcome conditions 0.97; $t(47) = 0.01$, $p = .99$. We also matched the frequencies of the primed-for real-word outcomes (e.g., *patch* and *batch*) across similar and dissimilar conditions: mean log frequency for the similar conditions was 1.14, and for dissimilar conditions also 1.14; $t(47) = 0.00$, $p < 1$.

Insert Table 1 about here

Four lists were drawn up from the 48 matched sets of word sequences. Each list contained only one sequence from each of the 48 sets, arranged such that there were equal numbers of sequences in each condition in each list. Within each list, half of the sequences were assigned to the auditory masking condition. Auditory masking was blocked, and four versions of each original list were drawn up such that all sequences appeared in masked and unmasked, and masking-first and masking-last conditions. Finally, each sequence in each of the resultant 16 lists was marked for either overt or silent recitation, such that there were equal numbers of each across other experimental conditions. This pattern was reversed to create an additional 16 lists, resulting in 32 lists of experimental items in a fully counterbalanced design.

Auditory masking was achieved using computer-generated pink noise, delivered through a set of Panasonic RP-HT225 stereo headphones. Participants' responses were captured on a Zoom H2 digital recorder and analyzed using Praat software (Boersma & Weenink, 2009).

Procedure. The procedure was closely modelled on that of Oppenheim and Dell (2008), with three differences: (a) we used visual timing cues (white dashes on the screen) instead of a auditory metronome to pace participants' repetitions of the word sequences (necessary because a metronome would not have been audible in conditions with auditory masking); (b) the tonguetwister sequence was not visible on the screen during experimental recitations, to eliminate potential orthographic interference; (c) the timing cues stopped automatically after each recitation, to ensure that participants had time to report errors.

Prior to beginning the experiment, participants underwent a computer-led tutorial and practise session, which included full instructions concerning the inner speech and overt speech procedures, with particular discouragement from attempting to 'mouth' sequences silently in the inner speech condition (preventing a possible effect of silent articulation: cf.Oppenheim & Dell, in press). At the beginning of each masked block, participants were instructed to adjust the fit and volume of the headphones to ensure that the loudness of the pink noise prevented them from hearing the sound of their own voice.

Tonguetwister sequences were presented in a random order on a 17" computer monitor. For each sequence, participants underwent a *familiarization* phase followed by a *performance* phase. In the familiarization phase, the tonguetwister sequence appeared in the centre of the screen, above an icon prompting participants to speak overtly (a mouth). Three seconds later, a series of four dashes appeared (one every second), acting as a visual metronome for the repetition of the words in the sequence. In the masked condition, pink noise began as the first dash appeared, and lasted until the last of the four dashes disappeared. The dashes and mouth icon were then replaced by a single dot, which remained onscreen for an additional second before the mouth icon reappeared and the dash sequence started again. The dash sequence was repeated so that participants repeated each sequence aloud four times before the performance phase began.

During familiarization, participants were not aware whether repetition of the sequence during the subsequent performance phase would be silent or out loud. Once familiarization had ended, the sequence was moved to the top of the screen and required activity was indicated centre-screen by means of the mouth icon (as used in familiarization), or a face icon representing silent repetition. At the same time the words "press ENTER to continue" appeared below the icon. Pressing ENTER caused all text to disappear from the screen, leaving only the mouth or face icon visible. After 200 ms, a four-dash sequence began at a rate of one dash every 500 ms, acting as a visual metronome for the (overt or silent) repetition of the tonguetwister sequence, and (in the blocks with auditory masking) pink noise started to play over the participant's headphones. 500 ms after the appearance of the fourth dash, the dashes disappeared, the pink noise (if any) ended, and the tonguetwister sequence reappeared at the top of the screen, together with an instruction to "report any errors and then press ENTER to continue" at the bottom. Participants were instructed to report each error aloud, as fully as possible, saying for example "I said *pat cap patch pad*, whereas I meant to say *pad cap catch pad*". Once errors, if any, had been reported, pressing ENTER started the next four-dash sequence. Each performance phase included four repetitions of the four dashes, before familiarization for the next word sequence began.

Transcriptions were made of participants' self-reports of their in inner and overt speech errors. Additionally, the transcriber identified errors in the overt speech condition irrespective of participants' reports. Errors were coded as *onset substitutions*, *correct pronunciations*, or *other errors*. In order to be considered as an onset exchange, an error had to consist of the substitution of the onset of a 'B' word with that of an 'A' word, with no concomitant change in the coda. For example, *catch* → *patch* was considered to be an onset exchange, but *catch* → *pan* was not (in practise, many of the latter type of error were indistinguishable from word-order errors, e.g., *catch* → *pad*). In cases where an error was followed by an overt self-repair, only the original error was coded for analysis.

Analyses. Analyses were carried out using logit mixed-effects models (Breslow & Clayton, 1993; DebRoy & Bates, 2004) using the lme4 package (Bates & Maechler, 2009) in R (R Development Core Team, 2009). This approach allowed us to investigate the contributions of experimentally-manipulated variables to the likelihood of making onset substitution errors relative to correct pronunciations (other errors were discarded from all analyses). For each dependent variable of interest we generated a base model which included an intercept, and random by-participant and by-item variation. We then proceeded to add predictors stepwise to the model. Selection of models was based on two criteria. First, we assessed whether the fit of the model to the data was improved by the addition of a given predictor using log-likelihood ratio tests, calculated as $-2(l_1 - l_0)$, where l_0 and l_1 represent the maximized likelihoods of models without and with the predictor of interest respectively. This difference can be evaluated using a χ^2 test because it has a null distribution approximating that of χ^2, with degrees of freedom representing the difference in the number of model parameters. Predictors were retained if the model was improved, but removed from consideration if they did not improve the current 'best' model. Second, where two or more predictors each significantly improved the current model, we selected the model which had the smallest log-likelihood. Once predictors representing the experimental manipulations and their interactions had been exhaustively explored, the resulting model represented the 'best fit' to the data, being a model which could not be improved by the addition of further predictors.

Each model includes coefficients representing the intercept and any effects of predictors. Where models were selected, the Wald statistic, calculated from each estimated coefficient and its standard error, was used to determine whether the coefficients differed significantly from zero (see Agresti, 2002).

Results

Out of a total of 6144 experimental recitations (48 tongue twisters, each repeated four times by 32 participants), participants self-reported a total of 851 errors of any type, of which 510 were in overt and 341 in inner speech.

Primed-for errors. 77 (40 overt, 37 inner speech) self-reported errors were cases in which the onset of word 3 was substituted with that of an ‘A’ word. Table 2 gives a breakdown of errors by experimental condition.

Insert Table 2 about here

Analyses included predictors of *masking* (whether or not there was auditory masking), *overtness* (whether or not participants were speaking aloud), *lexicality* (whether an onset substitution would result in a word), and *similarity* (whether the substituted phonemes differed by one feature or more). The best fit model of self-reported errors included effects of lexicality and similarity, but was not improved by including their interaction ($\chi^2(1) = 2.12$, $p = .15$). There were no effects of, or interactions with, masking (all $\chi^2(1) \leq 2.42$, $p\text{s} \geq .12$).

Surprisingly, there were no discernible effects of, or interactions with, overtness (all $\chi^2(1) \leq 0.26$, $p\text{s} \geq .61$). Table 3 gives the coefficients of the model, and the probabilities that they differ from zero. According to the model, participants were approximately 2.3 ($= e^{0.85}$) times as likely to report errors when the outcome was a word, and 2.2 times as likely when the substituted phoneme differed by a single feature, with other differences between conditions attributable to random variance.

Insert Table 3 about here

Because of the unexpected lack of overtness effects, we ran two additional analyses. The first of these explored the inner speech condition in isolation, as it was in inner speech that Oppenheim and Dell reported the absence of a phonemic similarity effect. In inner speech, the fit of a model including lexicality and similarity is improved by the addition of their interaction ($\chi^2(1) = 4.13$, $p = .04$), but the tendency for participants to report less errors with word outcomes where the substituted phoneme is similar is not reliable (see Table 3). Importantly, the main effect of similarity remains significant; participants are more likely to report errors with similar phonemes in inner speech alone. The second additional analysis allowed a direct comparison with Oppenheim and Dell's interaction analysis for phonemic similarity, which combined errors reported on words 2 and 3 (since word 2 was not manipulated for lexicality of outcome, we followed Oppenheim and Dell in ignoring lexicality). In inner speech, participants reported substitutions of 16 dissimilar and 28 similar phonemes; in overt speech, 12 and 34. There was no effect of overtness ($\chi^2(1) = 0.12$, $p = .73$); a further explicit test confirmed that there was no interaction of overtness with the substantial effect of phonemic similarity ($\chi^2(2) = 1.46$, $p = .48$): Once random variance was accounted for, participants were 2.3 times as likely to report substitutions of phonemes that differed by a single feature as those that differed by two or more, regardless of whether or not they were speaking aloud. Table 3 gives the model coefficients.

Accuracy of self-reporting. 632 errors were transcribed from recordings of participants' overt speech. Of these, 66 were cases in which the onset of word 3 was substituted with that of word 4. To evaluate the accuracy with which participants reported their own errors in overt speech, we ran an additional set of analyses comparing participants' self-reported errors to those transcribed from the recordings. As well as random effects of items and participants, these models included a random variable for the rater, with 33 levels representing each of the individuals who identified errors: 32 participants, plus one independent rater. This variable was designed to control for differences in individuals' propensities to report errors. As well as the experimental predictors discussed above, we also included a *rater type* predictor, with 2 levels (self-rating or independent rater). The best fitting model included lexicality and similarity as predictors (coefficients are given in Table 3). The inclusion of rater type further improved the model ($\chi^2(1) = 5.01$, $p = .03$), showing that, overall, the independent rater reported more primed-for errors than did the participants. Irrespective of who was doing the rating, errors with lexical outcomes, and those involving the substitutions of phonemes which differed by single features, were reported in significantly greater numbers.

Discussion

In Experiment 1 a lexical bias was found in both overt and inner speech conditions, as also reported by Oppenheim and Dell (2008). Because the tonguetwisters in this experiment consisted entirely of words, this bias is compatible both with Oppenheim and Dell's activation feedback simulation, and with an account in which the lexical bias is due wholly or in part to lexical editing (Baars et al., 1975; Hartsuiker et al., 2005). A similarity effect was also found, but strikingly, this did not interact with overtness. Contrary to Oppenheim and Dell (2008), participants reliably reported more substitutions of similar than of dissimilar phonemes at word 3 regardless of overtness, and the phonemic similarity effect held true when inner speech was considered in isolation. When word 2 was included in analyses, the same pattern emerged. The observed patterns of errors do not seem to be likely to be due to participants' misreporting of the errors that they made, given the correspondence between participants' own responses to their overt speech, and those of an independent rater.

Given that the effect of phonemic similarity was not affected by overtness, it is perhaps unsurprising that auditory masking made no difference either. However, the lack of a masking effect does suggest that any differences between studies cannot be attributed to participants' abilities to 'hear' the errors they are making.

Because the pattern of results in Experiment 1 had not been predicted, a primary motivation for Experiment 2 was to replicate our findings using an entirely new set of materials. We varied the design for this experiment, by creating tonguetwisters such that on word 3, the primed-for errors (rather than word 3 itself) were identical across similar and dissimilar conditions. This was done to eliminate any serendipitous effects that might be associated with the particular errors produced in Experiment 1. In an attempt to distinguish an account based solely on activation feedback from one which included editing, we also changed the makeup of the tonguetwisters such that they comprised a mixture of nonwords and real words. On the editing account, the production of tonguetwisters in a mixed context should reduce (or even eliminate) the adaptive utility of lexical editing (cf. Hartsuiker et al., 2005), suggesting that the lexical bias in Experiment 2 should weaken or disappear.

For comparison with Experiment 1, an auditory masking condition was also included in Experiment 2.

Experiment 2

Experiment 2 was a replication of Experiment 1 which differed in the following respects: (a) A novel set of materials was used, in which half of the targets were nonwords; (b) Materials were generated such that within each set of four tonguetwisters there was only one word and one nonword outcome.

Method

Participants. Thirty-two native speakers of English were recruited from the Edinburgh student population for course credit. Ages ranged from 18 to 26.

Materials. Forty-eight matched sets of four-word tonguetwister sequences were generated. Candidate sequences were generated automatically from two lists: one of CVC(C) words as in Experiment 1, and another of CVC(C) phonotactically legal nonwords obtained from the ARC Nonword Database (Rastle, Harrington, & Coltheart, 2002). Within each 4-word tonguetwister the number of nonwords varied from one to three; nonwords could appear in any of the four positions. Across the 48 sets, exactly 50% of sequence members were words, and 50% nonwords.

Each set comprised four tonguetwister sequences with ABBA onset patterns. In contrast to Experiment 1, words 1 and 4 (the ‘A’ words) did not change. Instead words 2 and 3 (the ‘B’ words) were manipulated to vary phonemic similarity of onsets, and lexicality of potential outcome. In two of the four sequences, the onsets of the ‘B’ words were phonologically similar to the ‘A’ onsets, differing by a single feature, and in two they were dissimilar, differing by two or more features. As in Experiment 1, the coda of word 3 was manipulated such that a substitution with an ‘A’ word onset would result in either a word or a nonword. Where they were real words, words in position 3 were frequency-matched: mean log frequency of words for the word outcome conditions was 1.39, and for nonword outcome conditions 1.35; $t(12) = 0.21$, $p = .84$. Mean log frequency for word 3 in the similar condition was 1.33, and in the dissimilar condition, 1.40; $t(12) = 0.27$, $p = .78$. Since the ‘B’ words were manipulated to ensure that the outcomes were the same, the outcome frequencies were exactly matched across similarity (mean log frequency = 1.03). A sample set of four sequences is given in Table 4.

Insert Table 4 about here

Construction of lists and counterbalancing proceeded as for Experiment 1.

Procedure. The procedure was identical to that of Experiment 1.

Analyses. Four mixed-effects analyses of primed-for errors were carried out, corresponding to those carried out in Experiment 1: inner vs. overt speech (word 3); inner speech only (word 3); inner vs. overt speech (words 2 and 3 combined); overt speech only (participants vs. independent rater).

Results

Out of a total of 6144 experimental recitations, participants self-reported a total of 950 errors of any type, of which 556 were in overt and 394 in inner speech.

Primed-for errors. 116 (58 overt, 58 inner speech) reported errors were cases in which the onset of word 3 was substituted with that of the ‘A’ word. Table 5 gives a breakdown of errors by experimental condition.

Insert Table 5 about here

As for Experiment 1, each of three analyses of participants’ self-reported errors included predictors of *masking*, *overtness*, *lexicality* and *similarity.* Coefficients for each model, and the probabilities that they differ from zero, are given in Table 6. The model for word 3 was improved by the addition of similarity. Excluding random variance, participants were 2.2 ($= e^{0.77}$) times more likely to report errors when the substituted phoneme differed by a single feature. No other predictor significantly improved the model (for masking, $\chi^2(1) = 2.76$, $p = .10$; for other predictors, $\chi^2(1) \leq 0.35$, ps $> .55$).

In the analysis of the inner speech condition in isolation, the fit of the model was improved only by the addition of similarity ($\chi^2(1) = 5.91$, $p = .02$), corresponding to a 2.0-fold increase in the likelihood of reporting errors with similar phonemes. In the third set of analyses, which combined errors reported on words 2 and 3 (ignoring lexicality), participants reported substitutions of 29 dissimilar and 50 similar phonemes in inner speech; for overt speech, 21 and 53. As in Experiment 1, there was no effect of overtness ($\chi^2(1) = 0.01$, $p = .93$), and overtness did not significantly interact with the substantial effect of phonemic similarity ($\chi^2(2) = 1.37$, $p = .50$). Accounting for random differences, participants were 2.1 times more likely to report substitutions of phonemes that differed by a single feature than those that differed by two or more, regardless of whether or not they were speaking aloud.

Insert Table 6 about here

Accuracy of self-reporting. 732 errors were transcribed from recordings of participants' overt speech. Of these, 137 were cases in which the onset of word 3 was substituted with that of word 4. As in Experiment 1, we ran a set of analyses comparing participants' self-reported errors to those transcribed from the recordings. These analyses included an additional random variable representing the rater. The best fitting model included independent effects of *rater type*, *masking* and *similarity* (coefficients are given in Table 6). Overall, the independent rater was more likely to report primed-for errors than were the participants. Irrespective of who was doing the rating, errors were significantly more likely to be reported under masked conditions; and primed-for errors involving the substitutions of phonemes which differed by single features were reported in significantly greater numbers.

Discussion

Experiment 2 replicated the phonemic similarity effect found in Experiment 1, and showed again that the effect did not interact with overtness of speech. As in Experiment 1, the effect of phonemic similarity was also found when errors reported in inner speech were considered separately. Once again, there were no effects of masking. However, contrary both to Experiment 1 and to Oppenheim and Dell (2008), there was no effect of lexicality: Participants were no more likely to report errors which resulted in real words than they were those that resulted in nonwords. We turn our attention to the lack of lexical bias in this experiment in the General Discussion below.

When participants' self-ratings of their overt speech were compared to those of the independent rater, effects of masking and of rater were also found. Taking both ratings together, more errors were reported when participants listened to pink noise, indicating that more overt errors were made in the masked than in the unmasked condition. This increase can be accounted for by an increased tendency of participants in the masked conditions to repeat the same error in subsequent iterations of the same tonguetwister. This suggests that auditory feedback helped participants to detect errors and take steps to prevent them reoccurring in subsequent iterations. Masking aside, the independent rater was also more likely to report errors overall, perhaps suggesting that participants experienced a degree of difficulty in detecting errors in an experiment where nonwords comprised half of the material. However, neither of these effects can account for the fact that phonemic similarity reliably affects the likelihood of substituting an onset in inner speech.

Over two experiments, we can find no evidence that there is a difference between inner and overt speech, contrary to the findings reported by Oppenheim and Dell (2008). A potential interpretation is that the differences between experiments lie in (as yet unspecified) differences between the materials used, despite the care taken by the present authors in material construction. In order to rule out an artefactual explanation, Experiment 3 is therefore a replication of Oppenheim and Dell's (2008) original experiment, using a version of the materials originally used in that study with one substituted item (Oppenheim & Dell, in press).

Experiment 3

Experiment 3 was a replication of Experiments 1 and 2, using the thirty-two tonguetwisters originally used by Oppenheim and Dell (in press). Because there were fewer tonguetwisters than in the previous experiments reported here we used a larger number of participants (48) to maintain equivalent power; because no interesting effects of masking had been found in either of Experiments 1 or 2 we did not manipulate masking in this experiment. In all other respects, Experiment 3 was identical to the previous two experiments.

Method

Participants. Forty-eight native speakers of English were recruited from the Edinburgh student population for course credit. Ages ranged from 18 to 23.

Materials. We used the thirty-two matched sets of four-word tonguetwister sequences originally used by Oppenheim and Dell (in press; thirty-one of these were identical to those used by Oppenheim & Dell, 2008). All sequences used real words, with onsets arranged in an ABBA pattern; as in Experiment 1, the onsets of the 'A' words were manipulated such that they were either phonemically similar or dissimilar to the 'B' onsets, and the coda of word 3 was manipulated such that the substitution of the onset of word 3 with an 'A' onset would result in either a word or a nonword.

Counterbalancing proceeded as for Experiment 1. Since there was no manipulating of auditory masking in the present experiment, the counterbalancing procedure resulted in 16 experimental lists.

Procedure. The procedure was identical to that of Experiment 1, with the exception that participants were not required to wear headphones, since all tonguetwisters were recited without auditory masking.

Analyses. As for Experiments 1 and 2, we carried out mixed-effects analyses of primed-for errors in inner vs. overt speech (word 3); inner speech only (word 3); inner vs. overt speech (words 2 and 3 combined); overt speech only (participants vs. independent rater). A final analysis considered Experiments 1–3 together, increasing the statistical power to establish whether there was any evidence in our experiments to support Oppenheim and Dell's (2008) finding that the phonemic similarity effect was reliably reduced when words 2 and 3 were considered together.

Results

Out of a total of 6144 experimental recitations, participants self-reported a total of 586 errors of any type, of which 361 were in overt and 225 in inner speech.

Primed-for errors. 75 (44 overt, 31 inner speech) self-reported errors were cases in which the onset of word 3 was substituted with that of an 'A' word. Table 7 gives a breakdown of errors by experimental condition.

Insert Table 7 about here

Each of three analyses of participants' self-reported errors included predictors of *overtness*, *lexicality* and *similarity.* Coefficients for each model, and the probabilities that they differ from zero, are given in Table 8. The best fit model for word 3 included factors of lexicality and similarity, but was not improved by including their interaction ($\chi^2(1) = 2.60$, $p = .11$). There was no effect of overtness ($\chi^2(1) = 3.14$, $p = .08$), and, importantly, overtness did not interact with either lexicality or phonemic similarity ($\chi^2(1) < 0.08$, ps $> .78$). Regardless of overtness, and taking random effects into account, participants were approximately 2.7 ($= e^{0.99}$) times as likely to report errors when the outcome was a word, and 2.7 times as likely when the substituted phoneme differed by a single feature.

When inner speech was analyzed in isolation, the best fit model included effects of lexicality ($\chi^2(1) = 5.78$, $p = .02$) and of similarity ($\chi^2(1) = 5.53$, $p = .02$), but not their interaction. Participants were 2.9 times as likely to report errors resulting in words, and 2.5 times as likely to report the substitutions of similar phonemes. In the analysis of words 2 and 3, participants reported substitutions of 13 dissimilar and 27 similar phonemes in inner speech; for overt speech, 14 and 43. A marginal effect of overtness ($\chi^2(1) = 3.76$, $p = .05$) did not reflect a reliable difference in likelihood, and did not interact with similarity ($\chi^2(1) = 0.63$, $p = .43$), showing that participants were once again much more likely (here, by a factor of 2.7) to substitute similar than dissimilar phonemes, regardless of whether the speech was overt or not.

Insert Table 8 about here

Accuracy of self-reporting. 826 errors were transcribed from recordings of participants' overt speech. Of these, 67 were cases in which the onset of word 3 was substituted with that of word 4. The analyses comparing self-reports to transcribed errors included an additional random variable representing the rater. The best fitting model included independent effects of *rater type*, *lexicality* and *similarity* (coefficients are given in Table 8). Overall, the independent rater was more likely to report primed-for errors than were the participants. Primed-for errors involving the substitutions of phonemes which differed by single features, or which resulted in real words, were reported in significantly greater numbers.

Discussion

The tonguetwisters in Experiment 3 consisted entirely of words. Consistent with Experiment 1 and with Oppenheim and Dell (2008), we found a lexical bias in both overt and inner speech conditions. However, there was no interaction of phonemic similarity with overtness. This finding is consistent with the results of Experiments 1 and 2, but stands in contrast to the results obtained using a highly similar material set by Oppenheim and Dell (2008).[1] Before considering the implications of our findings further, we turn to a final set of analyses in which the results of all three experiments reported above are considered together.

Meta-analyses

Over three experiments, we have shown that overtness does not interact with similarity in predicting the likelihood of an onset substitution. However, there are some small signs that there may be numerical trends in this direction; and since Experiment 3 is a direct replication of a study where an interaction was previously reported, it is particularly important to exhaustively test for an interaction. To this end, we report two further analyses below, which include the data from all three of the experiments reported above, collapsed across masking where appropriate. The first focuses on word 3 only, following up on the main analyses reported above, and investigating the effects of lexicality, similarity, and overtness. In the analysis of word 3 errors we built additional models, allowing us to explore the role of lexical frequency. We included two types of frequency in our analyses: first, the target frequency (excluding items from Experiment 2 where the target was a nonword); and second, the frequency of the potential outcome of an onset substitution (again only including cases where real words resulted).

The second analysis follows Oppenheim and Dell (2008), investigating the effects of similarity and overtness on onset substitutions in words 2 and 3 combined. Both analyses included an additional ‘experiment’ factor, but since there were no effects of, or interactions with, experiment once other predictors had been included in the models, the details are omitted below (but see footnote 3).

Across the three experiments, onset substitutions on word 3 were affected by lexicality ($\chi^2(1) = 18.5$, $p < .01$) and similarity ($\chi^2(1) = 41.5$, $p < .01$). There was no effect of overtness ($\chi^2(1) = 1.82$, $p = .18$); an explicit test for an interaction of similarity with overtness also failed to reach significance ($\chi^2(2) = 2.41$, $p = .29$). In this analysis, we also built models which focused on the frequencies of lexical targets and outcomes. Perhaps unsurprisingly, there was a reliable effect of (log) frequency: High-frequency words were less error-prone, such that for each 10-fold increase in word 3’s frequency, participants were 0.7 times as likely to make an error involving that word ($\chi^2(1) = 560$, $p < .01$). Importantly, there was no interaction between the effect of frequency and those of either lexicality or similarity ($\chi^2(1) \leq .86$, ps $> .35$). In the best fitting model, participants were 2.6 times as likely to report errors resulting in words, and 2.4 times as likely to report substitutions of similar phonemes, whether in inner or in overt speech. Model coefficients are given in Table 9.

Unlike the frequency of word 3, the (log) frequency of the lexical outcome of a potential onset substitution had no effect on the likelihood of making that substitution ($\chi^2(1) < .36$, $p \geq .55$). Once again, there was an effect of phonemic similarity, and frequency of word 3 improved the model significantly but had a marginal effect on likelihood. Coefficients are given in Table 9. Explicit testing showed that there were no effects of, or interactions with, experiment.

Insert Table 9 about here

Taking words 2 and 3 together, the best-fitting model was one in which participants were 2.3 times as likely to report substitutions of similar phonemes as those of dissimilar phonemes, with other differences between conditions attributable to random variance. Once again, there were no effects of overtness ($\chi^2(1) = 1.35$, $p = .25$); nor, when we explicitly tested for an interaction with overtness, was there a significant effect ($\chi^2(2) = 4.91$, $p = .09$). Explicit testing revealed no effects of, or interactions with, experiment. Taking data from 18,432 total recitations of four-word tonguetwisters by 112 participants, no evidence could be found that any numerical difference in the likelihood of substituting similar phonemes in inner compared to overt speech was reliable.

General Discussion

The experiments reported above were predicated on the claim that a phonemic similarity bias in speech errors is found in overt, but not inner, speech (Oppenheim & Dell, 2008). In two experiments modelled closely on that of Oppenheim and Dell, we manipulated auditory masking, to determine whether single-feature errors were underreported in inner speech because they were harder for speakers to detect without access to the acoustic signal corresponding to their own speech. However, we did not find any effects of masking. Instead, in our experiments, phonemic similarity consistently influenced the likelihood of reporting errors to a similar extent in inner speech as it did in overt speech; moreover, in each experiment, the phonemic similarity effect was still clearly evident when inner speech was considered independently. Perhaps most surprisingly, when we replicated our experiments using Oppenheim and Dell's (in press) materials, the results were consistent with our two earlier experiments: Once again, the effects of phonemic similarity were manifest whether or not speech was overt.

The only differences between the experiments reported here and that reported by Oppenheim and Dell (2008) were the inclusion of masking as an additional experimental variable in two experiments, and the minor procedural alterations detailed above. Masking did not interact with any of the other variables under consideration, weakening any suggestion that participants in our experiments were somehow differentially able to detect errors in inner speech. The results of Experiment 3 (without a masking manipulation) did not differ from those of the first two experiments. We have no specific reason to suspect that the procedural differences should have had an impact on the distribution across conditions of reported errors, although we note that Oppenheim and Dell's (2008) use of an auditory rather than a visual metronome may have impacted on participants' ability to monitor their speech; and requiring them to speak when prompted by a metronome (as opposed to initiating each of their own tonguetwister recitations) may also have had an effect. Although we are not able to fully account for Oppenheim and Dell's (2008) findings, we believe that we have demonstrated in three stringently-controlled experiments that it is perfectly possible to find a phonemic similarity effect in inner speech.

The most straightforward interpretation of this evidence is that inner speech is not (always) impoverished at the featural level. Instead, feedback of activation from the feature to the phoneme levels of representation supports the bottom-up activation of competitor phonemes in inner speech as it does in overt speech, resulting in both cases in a tendency to substitute similar phonemes for one another. Note that this is not the same as claiming that the representations of inner and overt speech are indistinguishable: For example, Wheeldon and Levelt (1995) have suggested that some details are not represented in inner speech. Their evidence comes from a series of experiments in which participants were asked to monitor for probe phonemes in their silent translations into Dutch of visually-presented English words; recognition latencies varied with syllable count, but crucially, not with the time it took to speak the Dutch words aloud. Wheeldon and Levelt concluded that their participants were monitoring abstract (timeless) syllabified phonological representations, at least in the absence of any need to plan for articulation.

This interpretation requires the assumption that the representation of phonetic duration in the speech plan affects the rate at which that plan is scanned, although there does not appear to be any strong reason to presume that such a relationship holds.[2] However, even if Wheeldon and Levelt's view is accepted, it does not rule out featural representations, and is fully compatible with evidence that inner speech is fully phonologically represented.

In contrast to the consistently reliable effects of phonemic similarity, the lexical bias effect found in Experiments 1 and 3 disappeared in Experiment 2. In Experiments 1 and 3 the tonguetwisters were made up of real words; in Experiment 2, they contained both words and nonwords. On a standard interactive account (e.g., Dell, 1986, 1988) lexical errors should predominate in both experiments, because the errorful production of real words is supported by activation feedback from the phonemic to the lexical level of representation. In a series of SLIP experiments, Dell (1990) provided support for this view by showing that target words with low frequencies were more likely to be affected by onset errors than were those with high frequencies, but that high-frequency error outcomes were no more likely to be produced than their low-frequency counterparts. In our own analyses we were able to use a comparatively large dataset to directly model the effects of frequency on the likelihood of producing an error, and confirm both of Dell's findings: Increased target frequency reduced the likelihood of an errorful onset substitution, but increased outcome frequency had no reliable effect. Dell argued that his findings were compatible with a model in which frequency was represented at the lemma level (although the contention that frequency is represented in this way has been challenged: e.g., Jescheniak & Levelt, 1994). Importantly, Dell (1990) argued that the lack of an outcome frequency effect militated against the view that the bias towards producing words could be attributed to an editor. According to the editing view (Baars et al., 1975; Levelt, 1989; Levelt et al., 1999), speakers are able to monitor their speech plans prior to articulation, using the comprehension system to edit out potential problems. Whereas there is little doubt that speakers are able to do this (e.g., Motley, Camden, & Baars, 1982), Dell argued that an editor based on the comprehension system (Levelt, 1983, 1989; see Postma, 2000, for alternatives) should be sensitive to the frequencies of the erroneously-planned words it detected: an effect he (and we) failed to find.

It may be that outcome frequency effects are simply not detectable in the present study, given the relatively small numbers of errors observed. Moreover, the failure to observe an increase in the numbers of errors reported without masking in Experiments 1 and 2 (cf. Postma & Noordanus, 1996) suggests that the editor does not rely on access to auditory feedback. However, the target frequency effects discussed above do not interact with experiment (or more generally, with lexicality) in any way, and the editor still has access to inner speech. Our findings, therefore, do not lead to a straightforward non-editing account of the differences between Experiment 2 and Experiments 1 and 3 in terms of lexical bias. In more recent formulations of the editing account, the self-monitor is adaptive (cf. Hartsuiker et al., 2005; Hartsuiker, 2006). That is, it edits out problems based on functional criteria. Where only real words are to be uttered (such as in Experiments 1 and 3), a nonword in the speech plan is informative, and it is useful for the monitor to employ a lexicality criterion. Other criteria may also apply in particular circumstances, for example, where it is undesirable to utter taboo words (Motley et al., 1982). But in cases where participants are asked to pronounce intermixed sequences of words and nonwords (e.g., Experiment 2), the lexical status of entries in the speech plan is uninformative, and it would not be useful to edit items on the basis of their lexical status. Unless the monitor has access to the intended utterance (e.g., Nooteboom, 2005a, 2005b), there are no useful editing criteria to apply. In this way, the differences between experiments reported here fall out as a consequence of the types of materials used.

The adaptive self-monitor account implies that the lack of lexical bias seen in Experiment 2 reflects the underlying position prior to editing. In previous work, mixed word-nonword lists such as that in Experiment 2 have resulted in a lexical bias and, because there are no adaptive editing criteria, this bias has been attributed to activation feedback from lexical representations (Hartsuiker et al., 2005; see also Nooteboom & Quené, 2008, but note that Baars et al., 1975 attribute this pattern to editing). In Experiment 2 there is no evidence that feedback plays a role, perhaps because the speed of tonguetwister recitation is too fast to allow activation to spread (cf. Dell, 1986), or because of the different way in which errors are elicited: Previous work has relied on the SLIP technique, in which onset exchanges are primed across trials, potentially increasing the amount of top-down activation in a production network. Another possibility is that errors in tonguetwisters are caused by a tendency to repeat gestures (Pouplier, 2008), although this seems an unlikely candidate given that errors are found in inner as well as overt speech.

Without activation feedback, the lexical bias in Experiment 1 must be wholly attributed to editing. This explanation implies that the monitor must be able to edit out some errors before they can be detected and reported by participants, even in inner speech (see Laver, 1980, for a similar suggestion), a process which would occur much faster than the 'strategic lexical editing' investigated by Nozari and Dell (2009). However, the pattern of results across Experiments 1–3 is wholly consistent with this view. Across the three experiments, the numbers of real words produced in error were approximately equal (54, 61, and 54, in order of experiment). However, the numbers of nonwords produced were substantially lower in Experiments 1 and 3 (24, 55, 21). A compelling interpretation of this pattern is that nonword errors were edited out in these experiments, relative to a baseline error rate.[3]

Suggesting that the lexical bias observed in Experiments 1 and 3 may be attributed, at least in part, to editing rather than to activation feedback raises the question of whether feedback is still necessitated in a full account of our findings. Although we have followed Oppenheim and Dell (2008) in attributing the phonemic similarity effect to activation feedback, it may be that a more parsimonious complete account of our findings could be built around editing, for example following Nooteboom's (2005a, 2005b) suggestion that the monitor (and thus the editing process) has access to the intended utterance. On this account, phoneme substitutions are more likely to be detected and edited out if they are saliently different from the intended utterance; in other words, substitutions of similar phonemes should be more likely to escape editing. We have however commented elsewhere that this seems to be an unlikely possibility: In order to compare an existing speech plan to what was originally intended, an editor would have to maintain details of the intended utterance. If these were maintained, it is not clear why an incorrect plan would be generated (McMillan & Corley, submitted).

In fact, activation feedback is not required to account for the phonemic similarity effect, regardless of whether or not the speech plan is edited. The claim that feedback is implicated rests on the assumption that the units output for speech are phonemes (Dell, 1986), but several recent studies have demonstrated the existence of speech errors that do not consist of whole-phoneme substitutions (e.g., Frisch & Wright, 2002; Goldrick & Blumstein, 2006; Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007). If features, rather than phonemes, are the units of output, then the phonemic similarity effect can be attributed within a feedforward model to the simple fact that a noise-driven error in the activation of a single feature is more probable than errors in multiple features.

Taking all of our findings into consideration, the conclusions of the present study are quite different from those of Oppenheim and Dell (2008). Across three experiments, there is no specific evidence that activation feedback causes the errors that participants report. The lexical bias in Experiments 1 and 3 may be, at least in part, due to editing; and the phonemic similarity effect found in all conditions can be accounted for using a feedforward model with featural output. Whether or not the underlying mechanisms of speech production include activation feedback, however, inner speech must be specified at the subphonemic level for phonemic similarity to exert an influence on the likelihood of reporting an accidental phoneme substitution. And whether or not editing is implicated in the reduction of errors that would have resulted in nonwords, it appears to apply equally across overt and covert speech conditions. The three experiments reported in the present paper suggest that, far from being underspecified, our 'inner voice' sounds much like our overt speech, and is produced in much the same way, whether overtly articulated or not.

References

Agresti, A. (2002). *Categorical data analysis.* Hoboken, NJ: Wiley.

Baars, B. J., Motley, M. T., & MacKay, D. G. (1975). Output editing for lexical status in artificially elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *14*, 382–391.

Baayen, H. R., Piepenbrock, R., & Gulikers, L. (1995). *The CELEX lexical database. release 2 (CD-ROM).* Philadelphia, Pennsylvania: Linguistic Data Consortium, University of Pennsylvania.

Baddeley, A. D., & Hitch, G. J. (1974). Working memory. *Psychology of Learning and Motivation*, *8*, 47–89.

Bates, D., & Maechler, M. (2009). lme4: Linear mixed-effects models using s4 classes [Computer software manual]. Available from `http://CRAN.R-project.org/package=lme4` (R package version 0.999375-31)

Boersma, P., & Weenink, D. (2009). Praat: Doing phonetics by computer (version 5.1.05) [Computer software manual]. Available from `http://www.praat.org/`

Borden, G. J. (1979). An interpretation of research on feedback interruption in speech. *Brain and Language*, *7*, 307–319.

Breslow, N. E., & Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Society*, *88*, 9–25.

DebRoy, S., & Bates, D. M. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis*, *91*, 1–17.

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

Dell, G. S. (1988). The retrieval of phonological forms in production: Tests of predictions from a connectionist model. *Journal of Memory and Language*, *27*, 124–142.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*, 313–349.

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*, 611–629.

Dell, G. S., & Repka, R. J. (1992). Errors in inner speech. In B. J. Baars (Ed.), *Experimental slips and human error: Exploring the architecture of volition* (pp. 237–262). New York: Plenum Press.

Ellis, A. (1988). Normal writing processes and peripheral acquired dysgraphias. *Language and Cognitive Processes*, *3*, 99–127.

Frisch, S. A., & Wright, R. (2002). The phonetics of phonological speech errors: An acoustic analysis of slips of the tongue. *Journal of Phonetics*, *30*, 139–162.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, *21*(6), 649 - 683.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*, 386–412.

Hartsuiker, R. J. (2006). Are speech error patterns affected by a monitoring bias? *Language and Cognitive Processes*, *21*, 856–891.

Hartsuiker, R. J., Corley, M., & Martensen, H. (2005). The lexical bias effect is modulated by context, but the standard monitoring account doesn't fly: Related beply to baars et al. (1975). *Journal of Memory and Language*, *52*, 58–70.

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

Lackner, J. R., & Tuller, B. H. (1979). Role of efference monitoring in the detection of self-produced speech errors. In W. E. Cooper & E. C. T. Walker (Eds.), *Sentence processing: Psycholinguistic studies presented to Merrill Garrett* (pp. 281–294).

Hillsdale, NJ: Erlbaum Associates.

Laver, J. D. M. (1980). Monitoring systems in the neurolinguistic control of speech production. In V. A. Fromkin (Ed.), *Errors in linguistic performance: Slips of the tongue, ear, pen, and hand* (pp. 287–305). New York: Academic Press.

Levelt, W. J. M. (1983). Monitoring in self-repair in speech. *Cognition*, *14*, 41–104.

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

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Levitt, A. G., & Healy, A. F. (1985). The roles of phoneme frequency, similarity, and availability in the experimental elicitation of speech errors. *Journal of Memory and Language*, *24*, 717–733.

MacKay, D. G. (1992). Awareness and error detection: New theories and research paradigms. *Consciousness and Cognition*, *1*, 199–225.

McMillan, C. T., & Corley, M. (submitted). *Subphonemic influences on the production of phonemes: Evidence from articulation.*

Motley, M. T., Camden, C. T., & Baars, B. J. (1982). Covert formulation and editing of anomalies in speech production; evidence from experimentally elicited slips of the tongue. *Journal of Verbal Learning and Verbal Behavior*, *21*, 578–594.

Nooteboom, S. (2005a). Lexical bias revisited: Detecting, rejecting and repairing speech errors in inner speech. *Speech Communication*, *47*, 43–58.

Nooteboom, S. (2005b). Listening to one-self: Monitoring speech production. In R. J. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech* (pp. 167–186). Hove, UK: Psychology Press.

Nooteboom, S., & Quené, H. (2008). Self-monitoring and feedback: A new attempt to find the main cause of lexical bias in phonological speech errors. *Journal of Memory and Language*, *58*, 837–861.

Nozari, N., & Dell, G. S. (2009). More on lexical bias: How efficient can a "lexical editor" be? *Journal of Memory and Language*, *60*, 291–307.

Oppenheim, G. M., & Dell, G. S. (2008). Inner speech slips exhibit lexical bias, but not the phonemic similarity effect. *Cognition*, *106*, 528–537.

Oppenheim, G. M., & Dell, G. S. (in press). Motor movement matters: The flexible abstractness of inner speech. *Memory & Cognition*.

Postma, A. (2000). Detection of errors during speech production: A review of speech monitoring models. *Cognition*, *77*, 97–131.

Postma, A., & Noordanus, C. (1996). Production and detection of speech errors in silent, mouthed, noise-masked, and normal auditory feedback speech. *Language and Speech*, *39*, 375–392.

Pouplier, M. (2008). The role of a coda consonant as error trigger in repetition tasks. *Journal of Phonetics*, 114–140.

R Development Core Team. (2009). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Available from `http://www.R-project.org` (ISBN 3-900051-07-0)

Rastle, K., Harrington, J., & Coltheart, M. (2002). 358,534 nonwords: The ARC nonword database. *Quarterly Journal of Experimental Psychology*, *55A*, 1339–1362.

Robinson, A. (1997). *British English Example Pronunciation dictionary (BEEP version 1.0).* Retrieved 1 August 2009 from `ftp://svr-ftp.eng.cam.ac.uk/pub/comp.speech/dictionaries/`.

Sokolov, A. (1972). *Inner speech and thought.* London: Plenum Press.

Vygotsky, L. S. (1986). *Thought and language.* Cambridge, MA: MIT Press.

Wheeldon, L. R., & Levelt, W. J. M. (1995). Monitoring the time course of phonological encoding. *Journal of Memory and Language*, *34*, 311–334.

Wilshire, C. E. (1999). The "tongue twister" paradigm as a technique for studying phonological encoding. *Language and Speech*, *42*, 57–82.

Appendix A

Materials for Experiment 1

Outcome and similarity for word 3 onset substitution			
word, similar	nonword, similar	word, dissimilar	nonword, dissimilar
fan van vat fad	fan van valve fad	man van vat mad	man van valve mad
pole coast cope poke	pole coast comb poke	soul coast cope soak	soul coast comb soak
till kid kin tinge	till kid kiln tinge	bill kid kin binge	bill kid kiln binge
seep heath heel scene	seep heath heave scene	keep heath heel keen	keep heath heave keen
rig link limb rip	rig link limp rip	dig link limb dip	dig link limp dip
pat cap catch pad	pat cap cab pad	bat cap catch bad	bat cap cab bad
busk puff puck bunk	busk puff pub bunk	musk puff puck monk	musk puff pub monk
cob golf gone cot	cob golf goth cot	yob golf gone yacht	yob golf goth yacht
finch ship shin fill	finch ship shift fill	pinch ship shin pill	pinch ship shift pill
meal bead beak mean	meal bead beach mean	weal bead beak wean	weal bead beach wean
dove gulf gull dump	dove gulf gut dump	love gulf gull lump	love gulf gut lump
wail range rake waist	wail range race waist	tale range rake taste	tale range race taste
pink bid bit pick	pink bid bib pick	kink bid bit kick	kink bid bib kick
come tut tub cuff	come tut tuck cuff	hum tut tub huff	hum tut tuck huff
conk toss top cog	conk toss tongs cog	honk toss top hog	honk toss tongs hog
reap leap leach reef	reap leap leash reef	beep leap leach beef	beep leap leash beef
dock tod tot dodge	dock tod tom dodge	lock tod tot lodge	lock tod tom lodge
peck ketch keg pet	peck ketch kelp pet	beck ketch keg bet	beck ketch kelp bet
gust cusp cut gum	gust cusp cup gum	rust cusp cut rum	rust cusp cup rum
face vein vale feign	face vein vague feign	race vein vale cane	race vein vague cane
pang tank tack patch	pang tank tap patch	hang tank tack hatch	hang tank tap hatch
hunk thump thug hump	hunk thump thud hump	junk thump thug jump	junk thump thud jump
rot watt wad rob	rot watt was rob	not watt wad knob	not watt was knob

continued

word, similar	nonword, similar	word, dissimilar	nonword, dissimilar
tape pain pale take	tape pain paid take	nape pain pale knave	nape pain paid knave
tut done duck tug	tut done dove tug	mutt done duck mug	mutt done dove mug
dock knock knot dodge	dock knock notch dodge	lock knock knot lodge	lock knock notch lodge
sill tick tip sick	sill tick tint sick	chill tick tip chick	chill tick tint chick
deck wreck wren dead	deck wreck realm dead	tech wreck wren ted	tech wreck realm ted
run duck dub rum	run duck dud rum	son duck dub some	son duck dud some
wench wreck red well	wench wreck rev well	bench wreck red bell	bench wreck rev bell
kale gauge gape cake	kale gauge gait cake	shale gauge gape shake	shale gauge gait shake
bag dad dash back	bag dad damp back	sag dad dash sack	sag dad damp sack
mull buff buck much	mull buff bulge much	dull buff buck dutch	dull buff bulge dutch
roam lone lope role	roam lone loaf role	dome lone lope dole	dome lone loaf dole
wade range reign wait	wade range wraith wait	maid range reign mate	maid range wraith mate
sit zing zip sick	sit zing zinc sick	knit zing zip nick	knit zing zinc nick
puff buff bunch punk	puff buff bulge punk	huff buff bunch hunk	huff buff bulge hunk
rip width witch rim	rip width wish rim	hip width witch hymn	hip width wish hymn
dock toss tot dosh	dock toss top dosh	wok toss tot wash	wok toss top wash
delve wreck ref dead	delve wreck realm dead	shelve wreck ref shed	shelve wreck realm shed
wreck wet west wren	wreck wet wedge wren	peck wet west pen	peck wet wedge pen
fame safe sail fade	fame safe sage fade	maim safe sail maid	maim safe sage maid
bell peg pet beck	bell peg pep beck	knell peg pet neck	knell peg pep neck
pad tank tack patch	pad tank tab patch	mad tank tack match	mad tank tab match
teem seep seek teach	teem seep siege teach	beam seep seek beach	beam seep siege beach
hub thump thug hush	hub thump thud hush	rub thump thug rush	rub thump thud rush
jug chuck chump just	jug chuck chub just	lug chuck chump must	lug chuck chub must
rot loft lock rob	rot loft loll rob	not loft lock knob	not loft loll knob

Appendix B

Materials for Experiment 2

Outcome and similarity for word 3 onset substitution			
word, similar	nonword, similar	word, dissimilar	nonword, dissimilar
sote zone zoal soap	sote zone zote soap	sote loan loal soap	sote loan lote soap
gulk dump dull gulf	gulk dump duck gulf	gulk lump lull gulf	gulk lump luck gulf
bish mill mitt bid	bish mill miss bid	bish hill hit bid	bish hill hiss bid
feel veal veat fieve	feel veal veam fieve	feel keel keat fieve	feel keal keme fieve
tod cog cock tonch	tod cog cob tonch	tod log lock tonch	tod log lob tonch
chilk jiv jick chin	chilk jiv jiz chin	chilk viv vik chin	chilk viv viz chin
jog chon chosh jonch	jog chon chof jonch	jog fon fosh jonch	jog fon fof jonch
leat reef reach leed	leat reef ream leed	leat beef beach leed	leat beef beam leed
make naist nace mane	make naist nabe mane	make saist sace mane	make saist sabe mane
beech meeg meast beeve	beech meeg meald beeve	beech keeg keast beeve	beech keeg keald beeve
gosh kolf cod gone	gosh kolf cop gone	gosh solf sod gone	gosh solf sop gone
song zof zolve soft	song zof zolf soft	song bof bolve soft	song bof bolf soft
yon watt wob yacht	yon watt wom yacht	yon shot shob yacht	yon shot shom yacht
kiv ghyst gick kiln	kiv ghyst gish kiln	kiv zist zick kiln	kiv zist zish kiln
fack valt vadd fab	fack valt vam fab	fack nalt nadd fab	fack nalt gnam fab
sag zap zadd sash	sag zap zav sash	sag rap rad sash	sag rap rav sash
bail gate gade beige	bail gate gaif beige	bail kate kade beige	bail kate kaif beige
fob thoft thox phon	fob thoft thomp phon	fob zoft zocs phon	fob zoft zomp phon
bung pub puzz bund	bung pub puv bund	bung sub suzz bund	bung sub suv bund
rug yull yust rulp	rug yull yumf rulp	rug tull tust rulp	rug tull tumf rulp
jost choss chot jog	jost choss chom jog	jost thos thot jog	jost thos thom jog
baint paich pain bail	baint paich pave bail	baint waich wane bail	baint waich wave bail
chess jep jec chel	chess jep jebb chel	chess fep phek chel	chess fep feb chel

continued

word, similar	nonword, similar	word, dissimilar	nonword, dissimilar
taste daith daik tape	taste daith daide tape	taste yaith yake tape	taste yaith yade tape
cod gop gonk cotch	cod gop golve cotch	cod rop ronk cotch	cod rop rolve cotch
beg mep meck bed	beg mep mem bed	beg vep veck bed	beg vep vem bed
pipe tibe tyke pife	pipe tibe tight pife	pipe lybe like pife	pipe lybe light pife
hag faz fache hat	hag faz falp hat	hag yaz yache hat	hag yaz yalp hat
vak zaf zatt van	vak zaf zatch van	vak waf wat van	vak waf wach van
puff cud kuck pumb	puff cud kutch pumb	puff thud thuck pumb	puff thud thuch pumb
heath sheen sheal heap	heath sheen sheace heap	heath jean jeel heap	heath jean jeace heap
roof noosh nool rooch	roof noosh noog rooch	roof soosh sool rooch	roof soosh soog rooch
nuv dutt dumb nund	nuv dutt dug nund	nuv chut chum nund	nuv chut chug nund
vote zome zole vose	vote zome zope vose	vote yome yoal vose	vote yome yope vose
muck nunk nuch mull	muck nunk nuzz mull	muck wunk wuch mull	muck wunk wuzz mull
lid rilk rim lizz	lid rilk rich lizz	lid hilk him lizz	lid hilk hitch lizz
dive gike gyne dime	dive gike gite dime	dive thike thyne dime	dive thike thite dime
namn dank dap nag	namn dank das nag	namn thank thap nag	namn thank thass nag
cap gab ghan cash	cap gab ghav cash	cap shab shan cash	cap shab shav cash
den nem neff dead	den nem nech dead	den sem seff dead	den sem sech dead
tud cuff cub tug	tud cuff come tug	tud huff hub tug	tud huff hum tug
bech delp dench beg	bech delp denth beg	bech selp sench beg	bech selp centh beg
keep teeth tiege keen	seep teeth teeve keen	seep keith keege keen	seep keith keeve keen
move boost boon moop	move boost boom moop	move roost rune moop	move roost room moop
ship fill fin shid	ship fill fib shid	ship bill bin shid	ship bill bib shid
tup duff dutch tuck	tup duff dunk tuck	tup huff hutch tuck	tup huff hunk tuck
sheave scene seef sheek	sheave scene seech sheek	sheave keen keaf sheek	sheave keen keech sheek
beve deal deke bean	beve deal deeth bean	beve veal veek bean	beve veal veath bean

Author Note

This work was completed while Brocklehurst was a student at Edinburgh and was in part sponsored by ESRC studentships ES/G01230X/1 to Brocklehurst, and PTA/031/2004/00279 to Moat. Portions of this work were presented at AMLaP 2009. The authors wish to thank Gary Dell, Gary Oppenheim, and an anonymous reviewer for helpful comments on earlier versions of this manuscript, and Gary Oppenheim for making the materials used in Experiment 3 available to us.

Footnotes

[1] Oppenheim and Dell (in press) did not test overt speech, and therefore an explicit comparison with results obtained using identical materials is not possible.

[2] A version of this argument was made by Sieb Nooteboom and Hugo Quené in an unpublished manuscript.

[3] The shape of this interaction was confirmed in a further logit mixed model analysis across experiments. There was no main effect of experiment ($\chi^2(1) = 2.91$, $p = .23$) but the interaction of experiment and lexicality significantly improved the model fit ($\chi^2(3) = 30.8$, $p < .01$). Compared to Experiment 1, participants were 2.2 times as likely to make nonlexical errors in Experiment 2 ($B = 0.79$, $p = .03$), but the likelihood of making lexical errors remained constant ($B = .04$, $p = .90$). The likelihoods of making errors in Experiment 3 did not differ statistically from those in Experiment 1.

Table 1

A matched set of tonguetwister sequences from Experiment 1

	Similar onsets				Dissimilar onsets			
Word outcome	pat	cap	catch	pad	bat	cap	catch	bad
Nonword outcome	pat	cap	cab	pad	bat	cap	cab	bad

Table 2

Experiment 1: Onset substitutions on word 3

	Similar Onsets			Dissimilar Onsets		
	Unmasked	Masked	(Total)	Unmasked	Masked	(Total)
Self-reports (inner speech)						
Word	9	7	(16)	7	5	(12)
Nonword	3	6	(9)	0	1	(1)
Self-reports (overt speech)						
Word	8	10	(18)	4	4	(8)
Nonword	4	6	(10)	1	3	(4)
Independent rater (overt speech)						
Word	14	12	(26)	6	5	(11)
Nonword	4	11	(15)	2	4	(6)

Table 3

Experiment 1: Model coefficients and probabilities for best-fitting models. All intercepts represent unmasked conditions; there were no reliable effects of masking in any analysis.

Predictor	Value	Coefficient	Std. Error	p(coefficient = 0)
Inner vs. overt speech, word 3				
(Intercept)	Nonword, Dissimilar, Inner	−5.59	0.31	< .001
Lexicality	Word	0.85	0.26	< .001
Similarity	Similar	0.79	0.26	.002
Inner speech only, word 3				
(Intercept)	Nonword, Dissimilar	−7.10	1.10	< .001
Lexicality	Word	2.53	1.12	.025
Similarity	Similar	2.24	1.14	.048
Lex×Sim	Word *and* Similar	−1.94	1.21	.110
Inner vs. overt speech, words 2 and 3				
(Intercept)	Dissimilar, Inner	−5.67	0.24	< .001
Similarity	Similar	0.83	0.24	< .001
Overt speech (participants vs. independent rater), word 3				
(Intercept)	Nonword, Dissimilar, Ppts	−6.26	0.27	< .001
Lexicality	Word	0.44	0.17	.012
Similarity	Similar	0.93	0.19	< .001
Rater	Independent Rater	0.54	0.17	.001

Table 4

A matched set of tonguetwister sequences from Experiment 2

	Similar onsets				Dissimilar onsets			
Word outcome	gulk	dump	dull	gulf	gulk	lump	lull	gulf
Nonword outcome	gulk	dump	duck	gulf	gulk	lump	luck	gulf

Table 5

Experiment 2: Onset substitutions on word 3

	Similar Onsets			Dissimilar Onsets		
	Unmasked	Masked	(Total)	Unmasked	Masked	(Total)
Self-reports (inner speech)						
Word	12	9	(21)	4	6	(10)
Nonword	8	9	(17)	3	7	(10)
Self-reports (overt speech)						
Word	12	14	(26)	1	3	(4)
Nonword	4	10	(14)	6	8	(14)
Independent rater (overt speech)						
Word	14	27	(41)	2	4	(6)
Nonword	11	23	(34)	8	7	(15)

Table 6

Experiment 2: Model coefficients and probabilities for best-fitting models. All intercepts represent unmasked conditions.

Predictor	Value	Coefficient	Std. Error	p(coefficient = 0)
Inner vs. overt speech, word 3				
(Intercept)	Nonword, Dissimilar, Inner	−4.86	0.25	< .001
Similarity	Similar	0.77	0.21	< .001
Inner speech only, word 3				
(Intercept)	Nonword, Dissimilar	−4.99	0.33	< .001
Similarity	Similar	0.69	0.30	.022
Inner vs. overt speech, words 2 and 3				
(Intercept)	Dissimilar, Inner	−5.21	0.22	< .001
Similarity	Similar	0.74	0.18	< .001
Overt speech (participants vs. independent rater), word 3				
(Intercept)	Nonword, Dissimilar, Ppts	−5.56	0.23	< .001
Similarity	Similar	0.94	0.13	< .001
Rater	Independent rater	0.66	0.12	< .001
Masking	Masked	0.51	0.12	.014

Table 7

Experiment 3: Onset substitutions on word 3

	Similar Onsets	Dissimilar Onsets
Self-reports (inner speech)		
Word	16	7
Nonword	6	2
Self-reports (overt speech)		
Word	20	11
Nonword	12	1
Independent rater (overt speech)		
Word	24	16
Nonword	14	1

Table 8

Experiment 3: Model coefficients and probabilities for best-fitting models.

Predictor	Value	Coefficient	Std. Error	p(coefficient = 0)
Inner vs. overt speech, word 3				
(Intercept)	Nonword, Dissimilar	−6.31	0.39	< .001
Lexicality	Word	0.99	0.27	< .001
Similarity	Similar	0.99	0.27	< .001
Inner speech only, word 3				
(Intercept)	Nonword, Dissimilar, Inner	−6.24	0.54	< .001
Lexicality	Word	1.08	0.49	.027
Similarity	Similar	0.90	0.43	.038
Inner vs. overt speech, words 2 and 3				
(Intercept)	Dissimilar, Inner	−6.04	0.29	< .001
Similarity	Similar	0.98	0.24	< .001
Overt speech (participants vs. independent rater), word 3				
(Intercept)	Nonword, Dissimilar, Ppts	−6.98	0.36	< .001
Rater	Independent rater	0.36	0.17	.030
Lexicality	Word	0.88	0.19	< .001
Similarity	Similar	1.06	0.18	< .001

Table 9

Meta-analyses: Model coefficients and probabilities for best-fitting models of data across 3 experiments.

Predictor	Value	Coefficient	Std. Error	p(coefficient = 0)
Inner vs. overt speech, word 3				
(Intercept)	Nonword, Dissimilar, $\log_{10}$(Freq) = 0	−5.51	0.26	< .001
Lexicality	Word	0.97	0.18	< .001
Similarity	Similar	0.88	0.18	< .001
$\log_{10}$(Frequency)	+1	−0.36	0.13	.004
Lexical outcomes only, word 3				
(Intercept)	Dissimilar, $\log_{10}$(Freq) = 0	−4.50	0.28	< .001
Similarity	Similar	0.76	0.21	< .001
$\log_{10}$(Frequency)	+1	−0.33	0.17	.056
Inner vs. overt speech, words 2 and 3				
(Intercept)	Dissimilar, Inner	−5.61	0.14	< .001
Similarity	Similar	0.83	0.12	< .001