Journal of Experimental Psychology:
Learning, Memory, and Cognition
2012, Vol. 38, No. 2, 502–512

0278-7393/12/$12.00 DOI: 10.1037/a0025257

COMMENTARY

The Case for Subphonemic Attenuation in Inner Speech: Comment on Corley, Brocklehurst, and Moat (2011)

Gary M. Oppenheim
University of Illinois at Urbana–Champaign

Corley, Brocklehurst, and Moat (2011) recently demonstrated a phonemic similarity effect for phonological errors in inner speech, claiming that it contradicted Oppenheim and Dell's (2008) characterization of inner speech as lacking subphonemic detail (e.g., features). However, finding *an effect* in both inner and overt speech is not the same as finding *equal effects* in inner and overt speech. In this response, I demonstrate that Corley et al.'s data are entirely consistent with the notion that inner speech lacks subphonemic detail and that each of their experiments exhibits a Similarity × Articulation interaction of about the same size that Oppenheim and Dell (2008, 2010) reported in their work. I further show that the major discrepancy between the labs' data lies primarily in the magnitude of the main effect of phonemic similarity and the overall efficiency of error elicitation, and demonstrate that greater similarity effects are associated with lower error rates. This leads to the conclusion that successful speech error research requires finding a sweet spot between too much randomness and not enough data.

Keywords: inner speech, speech errors, phonemic similarity, phonological encoding, overdetermination

Inner speech is a form of imagery that supports many cognitive activities, including reading, planning (e.g., Baddeley, Thomson, & Buchanan, 1975), and possibly overt speech production monitoring (Levelt, 1983). Its generation is typically thought to involve a subset of the processing required for speaking aloud, with dispute over precisely how far that parallel extends. According to one recent claim, from Oppenheim and Dell (2008), inner speech corresponds to an abstract phonological processing level (e.g., Dell, 1986; Wheeldon & Levelt, 1995) with less robust (i.e., weaker or inconsistent) access to subphonemic information (e.g., featural, phonetic, motoric). It parallels overt production to the point of retrieving and sequencing abstract phonemes, with processing attenuated thereafter. Major empirical support for this *subphonemic attenuation hypothesis* (SAH) comes from comparing inner and overt "slips of the tongue" (e.g., REEF → /lif/). Overt slips tend to involve similarly articulated phonemes (the phonemic similarity effect; e.g., Nooteboom, 1969). For instance, a /r/ to /l/ slip (voiced alveolars, differing in manner of articulation) is more likely than a /r/ to /b/ (both voiced but differing in place and manner). This reliable overt speech effect is often attributed to the influence of subphonemic (featural) details during speech planning (e.g., Dell, 1986), so its size in inner speech, compared to overt, should reflect the relative contribution of subphonemic information. If inner speech tends to involve subphonemic details to the same extent as overt speech, then phonemic similarity should be equally important in determining error patterns, yielding equally strong similarity effects. But, if subphonemic information is less important to inner speech (the SAH claim), then its similarity effect should be weaker.

A weaker phonemic similarity effect was precisely what Oppenheim and Dell (2008) found when comparing tongue-twister-elicited errors in inner speech to those in overt, providing initial support for the SAH. Comparable tendencies for both inner and overt slips to create words (lexical bias) suggested robust engagement at the phoneme level for inner speech, in contrast to the differences in the similarity effects. Oppenheim and Dell (2010) replicated and extended the work, demonstrating that silently mouthing a tongue twister elicited an overtlike similarity effect in inner speech, while the tendency in unarticulated inner speech was again significantly diminished. Therefore, the attenuated similarity effect in unarticulated inner speech could not be due to difficulty "hearing" inner slips. Converging support for the SAH came from observations that the influence of articulatory features in phonological working memory hinges on a task's use of overt articulation (Schweppe, Grice, & Rummer, 2011). Thus, the SAH and associated empirical findings have proven empirically robust and, to judge by recent discussions (e.g., Geva, Bennett, Warburton, & Patterson, 2011; Harley, 2010; Hickok, Houde, & Rong, 2011; Hubbard, 2010; Huettig & Hartsuiker, 2010; Huettig, Rommers, & Meyer, 2011; Laganaro & Zimmermann, 2010; Nooteboom &

Gary M. Oppenheim, Beckman Institute for Advanced Science and Technology and Psychology Department, University of Illinois at Urbana–Champaign.

Preparation of this manuscript was supported by National Institutes of Health Grants DC000191 and HD44458. I thank Gary Dell, Sieb Nooteboom, Kay Bock, Simon Fischer-Baum, Scott Fraundorf, Matt Goldrick, Audrey Kittredge, Tuan Lam, Nazbanou Nozari, and Ying Wang for valuable discussions and other contributions. I also thank Paul Brocklehurst and colleagues for inspiring this response and providing data from their 2011 article.

Correspondence concerning this article should be addressed to Gary M. Oppenheim, Beckman Institute, University of Illinois, 405 North Mathews Avenue, Urbana, IL 61801. E-mail: goppenh2@illinois.edu

Quené, 2008; Nozari & Dell, 2009; O'Seaghdha, Chen, & Chen, 2010; Rahman & Aristei, 2010; Severens, Janssens, Kühn, Brass, & Hartsuiker, 2011; Stemberger, 2009; Vicente & Martinez Manrique, 2011), theoretically useful.

Corley, Brocklehurst, and Moat (2011) recently presented evidence that they interpreted as challenging the SAH. In three experiments modeled on Oppenheim and Dell's (2008) task, they replicated the lexical bias findings but additionally observed simple main effects of phonemic similarity in both inner and overt speech, with only "some small signs that there might be numerical trends" (Corley et al., 2011, p. 169) toward a weaker similarity effect in inner speech. Since their Similarity × Overtness interaction was only marginally significant ($p < .09$ in a 2-*df* model comparison), Corley et al. concluded in favor of the null hypothesis that the similarity effects in inner and overt speech were equal: "Over three experiments, we have shown that overtness does not interact with similarity in predicting the likelihood of an onset substitution" (Corley et al., 2011, p. 169). "Taking data from 18,432 total recitations of four-word tongue twisters 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" (Corley et al., 2011, p. 171). "Phonemic similarity consistently influenced the likelihood of reporting errors to a similar extent in inner speech as [it did] in overt speech" (Corley et al., 2011, p. 171). As Corley et al. noted, "Perhaps most surprisingly, when we replicated our experiments using Oppenheim and Dell's [2010] materials, the results were consistent with our two earlier experiments" (Corley et al., 2011, p. 171). Figure 1 compares results from Oppenheim and Dell's (2008, 2010) two studies with Corley et al.'s experiment using the same stimuli (their Experiment 3). Oppenheim and Dell's experiments show significant crossover interactions, where the phonemic similarity effect was stronger in overtly articulated speech, but Corley et al.'s crossover interaction was not statistically significant. The figure clearly indicates a comparable interaction, though, so it is important to closely examine it and Corley et al.'s other two studies with different but comparable stimuli.

In Corley et al.'s (2011) Experiment 3, the similarity effect in inner speech consisted of a 13-error difference between the similar and dissimilar conditions, yielding an odds ratio of 2.1:1. This difference is asserted to be equal to that in overt speech (29-error difference, an odds ratio of 3.1:1): "participants were once again much more likely (here, by a factor of 2.7) to substitute similar rather than dissimilar phonemes, regardless of whether the speech was overt or not" (Corley et al., 2011, p. 169). More generally, Corley et al. concluded that, "far from being underspecified, people's inner voice sounds much like their overt speech and is produced in much the same way, whether overtly articulated or not" (Corley et al., 2011, p. 172). It is both the specific conclusion that the similarity effects are the same in inner and overt slips and the general one that inner speech is not underspecified that I dispute here.

Like the similarity effect, the size of the Similarity × Articulation interaction can be quantified as an odds ratio. In Corley et al.'s (2011) Experiment 3, presented in the figure, the interaction effect size was 1.45:1. It was even greater for their other two experiments: 1.70:1 and 1.51:1. This means that Corley et al.'s overt speech similarity effects were 45%, 51%, and 71% larger than those in the comparable inner speech conditions. Do these findings justify asserting the null effect? I argue that they do not.

Through a statistical and conceptual reconsideration of the error data from five experiments (i.e., Corley et al.'s, 2011, three experiments and Oppenheim & Dell, 2008, 2010), this analysis first demonstrates that the data clearly and consistently support the SAH prediction that phonemic similarity effects are weaker in unarticulated inner speech. It also shows that Corley et al.'s (2011) central finding—that their data showed no evidence of the expected Similarity × Overtness interaction—is false and, hence, their theoretical interpretation—that inner speech is fully specified for phonological features[1]—is unsupported, and that even a more nuanced version of Corley et al.'s claim—that they found a significantly weaker interaction than Oppenheim and Dell (2008, 2010)—lacks statistical support. Second, since Corley et al. could not explain why they found a simple main effect of phonemic similarity in inner speech whereas Oppenheim and Dell did not ("we are not able to fully account for Oppenheim and Dell's findings"; Corley et al., 2011, p. 171), this response offers a plausible resolution for the discrepant findings—one that draws on a mathematical consequence of the overdetermined nature of speech errors to demonstrate a long-recognized, but underappreciated, aspect of speech error distributions.

The Phonemic Similarity × Articulation Interaction

The first analysis builds on Corley et al.'s (2011) recognition that power issues for individual speech error experiments may be alleviated by combining data across similar experiments. Laboratory studies of speech errors tend to elicit few errors of the desired type, so even seemingly large studies may lack the power to statistically detect a true effect. For instance, Corley et al.'s 18,432 total recitations only provided 340 target errors, spread over 112 subjects, 128 items, and four conditions in three experiments. It is error counts that build the power of a binomial analysis—not the number of trials, subjects, items, or experiments. (Mathematically,[2] increasing the number of observations [trials], while holding constant the number of successes [errors], only reduces one's chance of statistically detecting an effect.) With few errors to examine, even reasonably strong, consistent differences can be difficult to detect. Though each of their experiments replicated Oppenheim and Dell's (2008, 2010) Similarity × Articulation interaction (e.g., in the sense of Killeen's, 2005, p_{rep}), even combining data across experiments only allowed Corley et al. to detect this reasonably large effect with marginal significance.

Of course, the real question is whether the interaction holds overall—not whether it achieves significance in some arbitrary

[1] In Corley et al.'s (2011) words, "inner speech is fully phonologically represented" (p. 171). Corley et al.'s terminology reflects a linguistic tradition where phonemes are considered, if anything, bundles of features (e.g., Browman & Goldstein, 1989), whereas Oppenheim and Dell's (2008, 2010) reflects a psycholinguistic approach where phonemes occupy a level above features (e.g., Dell, 1986; Wheeldon & Levelt, 1995).

[2] For instance, according to formulas such as $SE_p = \sqrt{p(1-p)/n}$. Holding the number of successes constant (where p = successes/n), increasing n only decreases p/SE_p, thus decreasing the power to detect an effect involving p. Since $OR_{p_1p_2} = (p_1/(1-p_1))/(p_2/(1-p_2))$, increasing n decreases $OR_{p_1p_2}$.

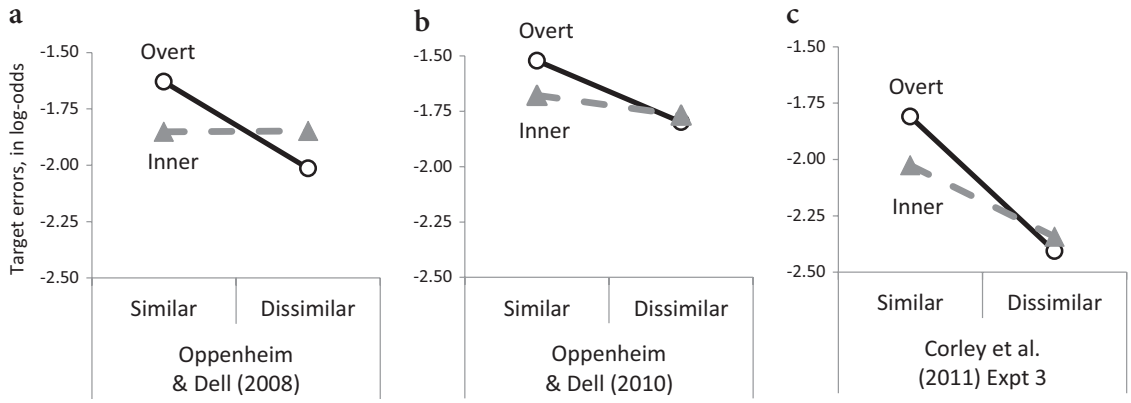

Figure 1. Error distributions for the same stimuli in three experiments testing phonemic similarity effects in overtly articulated and unarticulated (inner) speech. Oppenheim and Dell's (2008, 2010) experiments (Panels a and b) show larger similarity effects in articulated speech compared to unarticulated. The same crossover interaction fails to reach significance in Corley, Brocklehurst, and Moat's (2011) Experiment 3 (Panel c), where, in their words, "participants were once again much more likely (here, by a factor of 2.7) to substitute similar rather than dissimilar phonemes, regardless of whether the speech was overt or not" (p. 169). As Corley et al. noted, the Panel c results "were consistent with [their] two earlier experiments" (Corley et al., 2011, p. 171).

subset. I address this question by extending Corley et al.'s (2011) combined analysis to include Oppenheim and Dell's (2008, 2010) experiments, recoding Oppenheim and Dell's data to match Corley et al.'s methods. These five experiments used comparable materials, tasks, and data coding to evaluate similarity effects and their interaction with articulation. The data are first compiled in a single table, providing the first parametric estimates of the interaction and its interexperiment variability. (Neither Oppenheim & Dell's, 2008, 2010, nonparametric statistics nor Corley et al.'s, 2011, stepwise regression allowed such estimates.) Then per-trial data are used to examine whether similarity effects are characteristically weaker in unarticulated inner speech, as Oppenheim and Dell proposed, and to evaluate evidence for claims that this interaction varies across labs.

Method

Analyses considered Oppenheim and Dell's (2008, 2010) two experiments and Corley et al.'s (2011) three. Each used four-word tongue twisters with ABBA onset phoneme patterns, constructed in matched sets varying featural similarity of the A and B onset consonants (e.g., LEAN REED REEF LEECH vs. BEAN REED REEF BEECH; see Oppenheim & Dell, 2008, for more details). Each participant memorized and then attempted to recite a single variant of each tongue twister four times, in a single articulation condition, reporting their errors aloud.

I obtained the data set from Corley et al.'s (2011) three-experiment meta-analysis (for which I am grateful) and recoded Oppenheim and Dell's (2008, 2010) data with their methods. The main change was that, since Corley et al. counted all errors in a given recitation (up to two target errors per attempt), whereas Oppenheim and Dell (2008, 2010) considered only the first error, I adopted Corley et al.'s scheme. Thus, the new error counts differ slightly from those in Oppenheim and Dell's previous studies. Following Corley et al.'s treatment, each trial was binomially coded as the number of target errors (successes) versus error-free productions (failures). Target errors were reported slips from a B onset to an A onset (e.g., REEF → /lif/) without other deviations (e.g., REEF → /lid/).[3]

Following Corley et al. (2011), analyses used mixed-effects logistic regression via Bates and Maechler's (2010) lme4 package for R (R Development Core Team, 2010), including crossed random effects for subject, nested within experiment, and item, non-nested to reflect some items' reuse across experiments.[4] Cross-lab analyses treated lab as a fixed effect (base level is Oppenheim & Dell, 2008, 2010); within-lab analyses used experiment instead.

Analyses considered three other fixed main effects: phonemic similarity, articulation, and audition. As in the source experiments, phonemic similarity (two levels) classified trials as similar or dissimilar based on the number of feature contrasts between the trial's two onset phonemes (1 vs. >1). Articulation (two levels) coded whether a trial involved overt movements, thus comparing unarticulated inner speech (without overt articulation) to mouthed, noise-masked, and normal overt speech (with overt articulation). Audition (two levels) coded whether a trial involved subject-audible overt speech (the so-called external loop), thus comparing inner speech, silently mouthed speech, and noise-masked overt speech (without external audition) to unmasked overt speech (with external audition). Treating noise-masked and silently mouthed speech as equivalent in this regard draws some justification from Postma and Noordanus (1996), who demonstrated continuity between these conditions across an exhaustive range of self-reported error types.

Fitting began with a full theoretically justified model, including centered binary predictors for phonemic similarity, articulation,

[3] Following Oppenheim and Dell (2008, 2010), I use all B→A target errors to increase the statistical power, but all claimed significant effects in the full analyses still hold if considering only Word 3.

[4] Coding item as a nested random effect yields equivalent statistical outcomes.

audition, and lab, plus two-way interactions between phonemic similarity and the other predictors. To evaluate Corley et al.'s (2011) claim that their Similarity × Articulation interaction differed from Oppenheim and Dell's (2008, 2010), I also included a Phonemic Similarity × Articulation × Lab interaction (and ancillary Articulation × Lab interaction). Nonsignificant components were incrementally removed to reduce collinearity, with the relevant portion of the reduced model then refitted to each lab's individual data. Model coefficients (β) denote log-odds; p values in the text reflect 1-*df* likelihood-ratio tests, evaluable under nondirectional ($\alpha = .05$ for $\beta = 0$) or directional ($\alpha = .10$ for $\beta > 0$ XOR $\beta < 0$) hypotheses as appropriate; tables provide Wald z scores.

Results and Discussion

The five experiments provided 912 self-reported target errors (487 in overtly articulated speech, 435 in unarticulated inner speech), summarized in Table 1, to evaluate differences in phonemic similarity effects between overtly articulated and unarticulated (inner) speech. The table suggests three major patterns. First, similarity effects are clearly stronger in articulated speech than unarticulated inner speech, and this holds for each individual experiment, whether considering counts or odds ratios. Second, although Corley et al.'s (2011) experiments show stronger similarity effects overall, their Similarity × Articulation interaction is remarkably stable and approximately equal in odds size to the estimate from Oppenheim and Dell's (2008, 2010) work. In fact the "small signs that there might be numerical trends in this direction" (Corley et al., 2011, p. 169) are not small at all: In overtly articulated speech, their subjects reported 130 similar errors to 47 dissimilar, an 83-error difference. In inner speech, they reported 105 similar errors to 58 dissimilar, a 47-error difference. Finally, although the two labs contributed comparable numbers of trials (Corley et al., 2011: 18,000 vs. Oppenheim & Dell, 2008, 2010: 16,000), Corley et al.'s experiments elicited far fewer target errors.

Statistical analyses support all of these trends (see Table 2). Consistent with Corley et al.'s (2011) masking analyses, neither audition[5] ($\beta = -.18$, $\chi^2(1) = 2.66$, $p = .10$) nor its interaction with phonemic similarity ($\beta = .04$, $\chi^2(1) = 0.03$, $p = .87$), reached significance. This reinforces Oppenheim and Dell's (2010) conclusion that smaller similarity effects in unarticulated inner speech do not merely reflect the absence of an external loop (cf. Levelt, 1983). Neither the Phonemic Similarity × Articulation × Lab interaction nor its ancillary Articulation × Lab interaction approached significance (each $|\beta| < .10$, $\chi^2(1) < 0.01$, $p > .60$), indicating a lack of support for claims that Corley et al. found a smaller Similarity × Articulation interaction. Incrementally removing these nonsignificant predictors reduced the model to the five fixed effects discussed below (full-reduced comparison: $\chi^2(4) = 2.94 < \text{crit}[\chi^2(1)]_{\alpha = .05}$).

The tendency for target errors to involve similarly articulated phonemes ($\beta = .57$, $\chi^2(1) = 52.09$, $p < .0001$) held for both overtly articulated speech ($\beta = .80$, $\chi^2(1) = 56.18$, $p < .0001$) and unarticulated inner speech ($\beta = .33$, $\chi^2(1) = 6.13$, $p < .02$) separately. This pattern reinforces Corley et al.'s (2011) assertion that inner speech can show similarity effects (contra Oppenheim & Dell's, 2008, data), implying inner speech can incorporate subphonemic information under some conditions (e.g., as Oppenheim & Dell, 2010, suggested).

However, finding *a similarity effect* in inner speech is not the same as finding *equal similarity effects* in inner and overt speech. In fact, the effect in overtly articulated speech is much greater, about 1.6 times the size of that in unarticulated speech (Similarity × Articulation interaction: $\beta = .48$, $\chi^2(1) = 12.04$, $p < .0006$). As Table 1 suggests, this ratio holds quite well for each lab individually (Oppenheim and Dell, 2008, 2010): $\beta = .52$, $\chi^2(1) = 9.02$, $p < .003$; Corley et al., 2011: $\beta = .45$, $\chi^2(1) = 3.56$, $p < .06$), meaning that their data actually agree on the direction and size of the Similarity × Articulation interaction. Note also that the interaction in Corley et al.'s (2011) data would be significant under the directional SAH prediction of a *stronger* similarity effect in articulated speech. Thus, the combined data clearly support the SAH, and each lab and each experiment individually contributes to this support.

If the difference in the labs' conclusions does not reflect differences in the size or direction of the Similarity × Articulation interaction, what does it reflect? Here, the analysis statistically confirms two points suggested by visual inspection of Figure 1. First, Corley et al.'s (2011) experiments elicited target errors at less than half the rate of Oppenheim and Dell's (2008, 2010; main effect of lab: $\beta = -.89$, $\chi^2(1) = 38.47$, $p < .0001$). Second, Corley et al.'s data showed stronger similarity effects overall, not just for inner speech (Similarity × Lab interaction: $\beta = .53$, $\chi^2(1) = 13.02$, $p < .0003$). Together these points appear to resolve the question of why two labs would report contrasting results for the Similarity × Articulation interaction that is the crucial test of the SAH: Although both data sets showed a similar-sized interaction, low error rates could make it harder to statistically detect, and the presence of simple main effects of similarity in both inner and overt speech would make the interaction easier to overlook.

To recap, the patterns from each experiment (see Table 1) and the statistical results from each lab (see Table 2) individually support the prediction of a weaker similarity effect in unarticulated inner speech, indicating that the overt similarity effect is about 60% greater. Corley et al.'s (2011) failure to statistically confirm the Similarity × Articulation interaction does not reflect a reliable or even substantial discrepancy between the two labs' data on that point, as implied by claims that they failed to detect the interaction despite great effort. And their *simple main effect* of similarity in inner speech—erroneously claimed as evidence against the SAH—is better characterized as a difference in the magnitude of the *main effect*.

The Simple Main Effect of Phonemic Similarity in Unarticulated Inner Speech

I have shown that the discrepancy in the presence of a simple main effect of phonemic similarity in unarticulated inner speech, which Corley et al. (2011) could not explain, is merely symptom-

[5] The negative audition coefficient suggests a trend opposite what is typically predicted (Levelt, 1983) and demonstrated (Postma & Noordanus, 1996), so a directional test of the coefficient has not been considered. Retaining the predictor, however, would not noticeably change any claimed results.

Table 1
Aggregated Data for the Five Experiments

Study	*n*	Trials per cell	*p*(target error)	*p*(other error)	Articulated			Unarticulated			Interaction OR
					Sim	Dis	OR	Sim	Dis	OR	
Oppenheim & Dell	128	4,096	.017	.114	196	114	1.72	133	129	1.02	1.69
2008	48	1,536	.016	.098	69	38	1.79	40	52	0.75	2.38
2010	80	2,560	.018	.123	127	76	1.68	93	77	1.20	1.41
Corley, Brocklehurst, & Moat (2011)	112	4,608	.009	.049	130	47	2.88	105	58	1.84	1.57
Experiment 1	32	1,536	.007	.050	34	12	3.01	28	16	1.76	1.70
Experiment 2	32	1,536	.012	.057	53	21	2.62	50	29	1.73	1.51
Experiment 3	48	1,536	.008	.041	43	14	3.13	27	13	2.16	1.45
Total	240	8,704	.013	.064	326	161	2.06	238	187	1.27	1.62

Note. Sim = similar; Dis = dissimilar; OR = odds ratio, estimated by fitting a logistic regression model to the restricted data set.

atic of a difference in the size of the main effect. So, this section proposes a theory-derived explanation that both stands on its own and could account for some variation in similarity effects across experiments. The story, in a nutshell, is that the overdetermined nature of slips of the tongue has the consequence that their specific causes (e.g., dimensions of similarity between interacting representations) should be more evident (as odds ratios) when their more general causes (e.g., stress, time pressure, novelty, priming) contribute less. But with less support from general causes, the resultantly rare events may provide more volatile estimates with less statistical power.

Although speakers may complain that they err too often, researchers complain that they err too rarely. Carefully controlled stimuli often fail to compel unimpaired speakers to produce the kinds of errors that researchers want in the quantities that they need. So, researchers rely on seemingly irrelevant aspects of an experiment to help elicit errors, such as time pressure, priming, and otherwise difficult sequences. The classic SLIP procedure (Baars, Motley, & MacKay, 1975), for instance, elicits spoonerisms in part by priming a particular onset phoneme sequence and then unexpectedly reversing it (ABABABBA). Researchers assume that such manipulations make phonological encoding less deterministic—shifting productions away from near-ceiling accuracy and hence increasing statistical power—but the resulting errors nonetheless reflect the structure and processes of successful speech production. Their assumption reflects what Freud (1901/1958) described as the overdetermined nature of speech errors: Many factors interact to determine if and how production may miss its mark. Thus a slip from BARN DOOR to /dɑrn bɔr/ may simultaneously reflect priming of the /d. . . b. . ./ onset pattern, featural overlap between /d/ and /b/ onsets, lexicality of the resulting utterance, pressure to respond quickly, and latent feelings of bucolic ennui. An underappreciated property of this overdetermination is that, if one factor better supports a slip, the remaining factors become less crucial. For instance, if priming increases the likelihood of substitutions in general, phoneme substitutions will require less support from shared features (i.e., phonemic similarity). Consequently, an error effect like that of phonemic similarity

Table 2
Regression Summaries

Experiment	Odds ratio	Coefficient β	*SE*(β)	*z*
All experiments				
Intercept	0.01	−4.67	0.09	−49.77
Similarity	1.77	0.57	0.08	7.49
Articulation	1.11	0.10	0.07	1.43
Similarity × Articulation	1.62	0.48	0.14	3.44
Lab	0.41	−0.89	0.13	−6.72
Similarity × Lab	1.69	0.53	0.15	3.52
Oppenheim & Dell (2008, 2010)				
Intercept, Experiment 2008	0.01	−4.35	0.16	−27.07
Experiment 2010	1.28	0.25	0.15	1.72
Similarity	1.33	0.28	0.09	3.24
Articulation	1.15	0.14	0.09	1.62
Similarity × Articulation	1.69	0.52	0.18	2.98
Corley, Brocklehurst, & Moat (2011)				
Intercept, Experiment 1	0.00	−5.36	0.20	−26.74
Experiment 2	1.64	0.49	0.27	1.81
Experiment 3	0.98	−0.03	0.28	−0.09
Similarity	2.30	0.83	0.12	6.79
Articulation	1.05	0.05	0.12	0.38
Similarity × Articulation	1.57	0.45	0.25	1.84

should tend to be larger in log-odds terms when fewer factors promote slips and, hence, overall error probabilities are lower.

Error researchers have long recognized that when they are rarer, errors are more likely to exhibit more of the properties that promote them. For example, exchange errors (e.g., BARN DOOR → /dɑrn bɔr/) are relatively rare compared to other substitutions, but it is in exchanges that similarity and familiarity effects are the strongest (e.g., Dell, 1986; Garrett, 1980). Similarly, phoneme substitution errors usually have obvious sources in the surrounding context, but noncontextual slips show proportionally stronger phonemic similarity effects (Stemberger, 1992), suggesting that similarity is more crucial for slips with less contextual support. At the other extreme, aphasic individuals, who make frequent phonological errors, may show attenuated phonemic similarity effects (e.g., Goldrick & Rapp, 2007; Laganaro & Zimmermann, 2010), demonstrating that similarity is less crucial for slips driven by other (e.g., lexical) factors. Consider the following analogy: Students at prestigious institutions often distinguish themselves by possessing certain factors (e.g., intelligence, ambition, industriousness). A student from a rich family, where attending such an institution is more common, may need only one of these factors, but a student from a poor family, who nonetheless makes it in, would be more likely to possess all of them. Making an error is like making it into college. If making it is rare, then those who do will more consistently exhibit the relevant factors.

For a mathematically transparent illustration of this relationship, consider a common mathematical approximation of a stochastic selection process: the Luce-Shepard choice rule (Equation 1, adapted from Luce, 1963; Shepard, 1957). Models often use this equation to translate continuous activations into expected probabilities for discrete outcomes (e.g., Dell, Burger, & Svec, 1997; Gordon & Dell, 2003; Kruschke, 1992; Love, Medin, & Gureckis, 2004; McClelland & Elman, 1986; McClelland & Rumelhart, 1981; Nosofsky, 1986).

$$p(i) = \frac{e^{\mu a_i}}{\sum e^{\mu a}}. \quad (1)$$

Here, a_i is the activation of outcome *i,* and $\mu \geq 0$ makes selection more or less deterministic. To apply the rule to phoneme selection, imagine a target phoneme, /r/, and two competitors, /l/ (similar) and /b/ (dissimilar), with activations, $a_{/r,l,b/} = \{1, 0.5, 0.25\}$, as might arise from a similarity-sensitive retrieval process. From Equation 1, the probability of erroneously choosing /l/, $p(/l/)$, is a function of the activation of /l/, $a_{/l/}$, compared to the summed activation of all potential onsets. With the exponential function, $e^{\mu a}$, μ scales selection to a total error rate: Larger μs yield more "correct" selections, while smaller μs simulate more random selection. Using the equation to derive probabilities of selecting /l/ or /b/, given the target /r/, the similarity effect can be estimated as the odds ratio for /l/ versus /b/ outcomes. Varying μ produces greater similarity effects when errors are less frequent (see Figure 2a).

Dell's (1986) model also instantiates this principle: Reducing activation noise (analogous to $-\mu$ in Equation 1) reduces error rates, yielding stronger similarity effects (see Figure 2b). These simulations also illustrate a downside to very low error rates: Resultant effects are small as counts (see Figure 2c) and variable as odds ratios (see Figure 2b). Noise similarly modulates phonological effects in Dell, Schwartz, Martin, Saffran, and Gagnon's

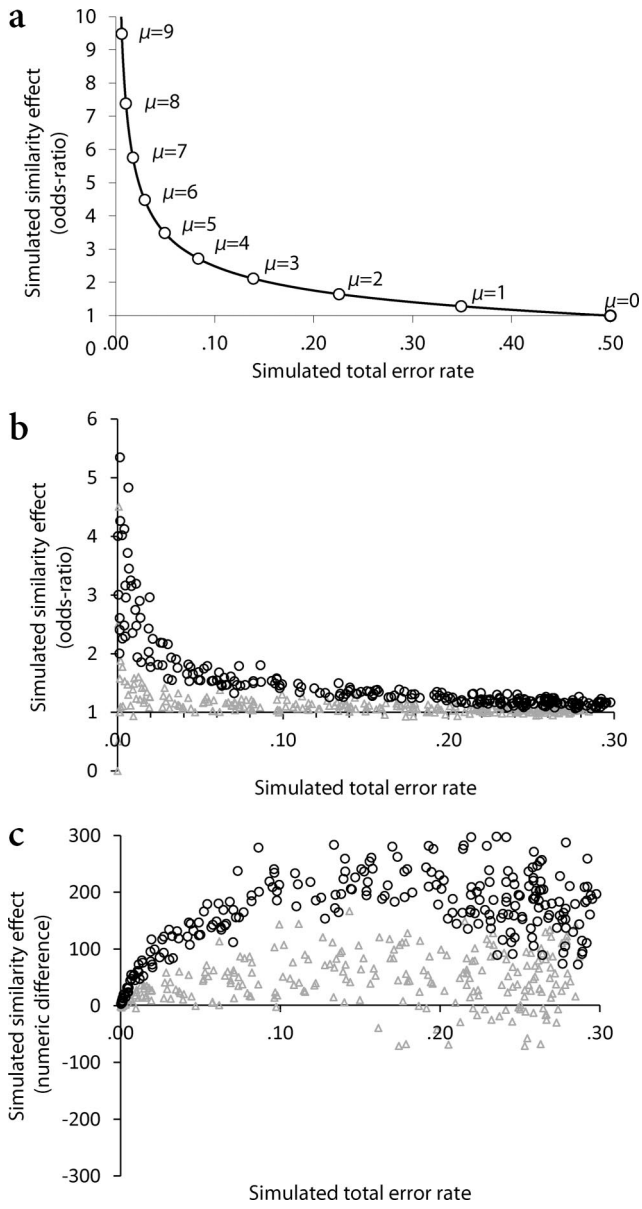

Figure 2. Models predict stronger odds-ratio similarity effects when errors are rare. Panel a: With the Luce-Shepard choice rule, more deterministic selections (larger μ) yield stronger error effects. Following the experiments, similar and dissimilar slip probabilities are calculated on separate trials. Parameters: $a_{\text{Target}} = 1$, $a_{\text{Similar}} = .5$, $a_{\text{Dissimilar}} = .25$, $\mu =$ 0:10. Panels b and c: Modulating activation noise in Dell's (1986) model produces similar behavior, additionally demonstrating that (Panel b) odds-ratio effect estimates grow less precise when one has fewer data with which to estimate them (i.e., at lower error rates) and (Panel c) higher error rates produce numerically larger effects (at least until selection approaches chance accuracy). Modulating phonological priming (not shown) has similar effects. Each point represents a 3,000-trial replication, reflecting the expected variability between similar-sized experiments. Parameters: lexical activation = 1; activation noise = $U(0.2, 1.7)$; competitor phoneme activation = 0.1; decay = 0.4;, connection weights = 0.2. Representing inner speech (triangles) with less robust phoneme–feature connections, weights set stochastically per subject: $0 \leq N(0.1, 0.1) \leq 0.2$, the model shows the same patterns with weaker effects.

Table 3
General Error Rate Regression Summaries

	Odds ratio	Coefficient β	*SE*(β)	*z*
Combined				
Intercept	0.01	−4.68	0.09	−50.60
Similarity	1.76	0.56	0.08	7.38
Articulation	1.11	0.10	0.07	1.44
Similarity × Articulation	1.62	0.48	0.14	3.44
General Error Rate ∗ 100	1.14	0.13	0.02	7.00
Similarity × (General Error Rate ∗ 100)	0.93	−0.07	0.02	−3.09
Articulated				
Intercept	0.01	−4.72	0.11	−44.05
Similarity	2.23	0.80	0.11	7.25
General Error Rate ∗ 100	1.14	0.13	0.02	5.77
Similarity × (General Error Rate ∗ 100)	0.93	−0.07	0.03	−2.15
Unarticulated				
Intercept	0.01	−4.81	0.11	−44.94
Similarity	1.37	0.32	0.11	2.87
General Error Rate ∗ 100	1.14	0.13	0.02	5.15
Similarity × (General Error Rate ∗ 100)	0.94	−0.06	0.03	−2.01

(1997) aphasia model and semantic effects in Oppenheim, Dell, and Schwartz's (2010) lexical retrieval model. In fact, every model of production errors that I know of has the property that lower error rates produce stronger odds-ratio effects. This connection is ubiquitous because it is a direct consequence of the overdetermined nature of speech errors. Factors that make production more deterministic (accurate) produce models where errors are infrequent (thus yielding less precise estimates), but error patterns are dominated by "good" errors (and, hence, greater odds-ratio effects). Factors that make production less deterministic lead to greater numerical error effects (at least until randomness dominates the error profiles) and, hence, more power to statistically detect them. In the current experiments, these influences manifest as larger phonemic similarity effects when slips are rarer but as more power to detect the same-sized effect when slips are more frequent. I claim that both articulated and unarticulated speech show this tendency but that the weaker similarity effect in unarticulated speech is easily overshadowed when production becomes less deterministic. The following statistical analyses demonstrate that the mediating effect of generalized error rate offers a reasonable explanation for the major difference between the labs' findings, providing possibly the first empirical demonstration of this long-recognized property of speech errors.

Method

This analysis used the same data set and basic methods as the previous. Subjects remain nested within experiment (and items remain nonnested), but I replace the nominal lab and experiment predictors with a continuous measure of *general error rate*—analogous to $-\mu$ in Equation 1. Nominal lab and experiment predictors are omitted because, since the analysis offers a theoretical explanation for the discrepancy that these identity-based factors describe, any predictor that successfully explains that discrepancy would necessarily be highly collinear with identity-based descriptions of it. For instance, general error rate hypothesizes a particular ordering and spacing of data across experiments; if it perfectly explained the variation across experiments, it would provide exactly the same estimates as the nominal experiment predictor. Because the previous section addressed the other fixed effects, I restrict discussion to the general error rate predictor and its interaction with phonemic similarity.

General error rate was calculated here as the proportion of Word 2 and Word 3 attempts per experiment that elicited neither target errors nor correct productions, that is, $1 - p(\text{correct}) - p(\text{target error})$. Given the available data, this method provides an estimate that is consistent across labs, reasonably independent of the target error distributions within an experiment,[6] and concordant with ordinal rankings of the published total error rates (Kendall's τ for concordance between general error rate and published per-experiment error rates: overtly articulated speech, $\tau = 1$, $p < .03$; unarticulated inner speech, $\tau = 1$, $p < .03$). However, the results do not hinge on this particular estimation method,[7] and other methods appear to work equally well here (e.g., more focused estimates of general onset slip rates, as in the simulations, may prove more robust for cross-paradigm comparisons). Finally, although Figure 2 suggests similarity effects should relate to target error rate via a power function, I treat general error rate as a linear predictor to avoid overfitting the limited data set and make the resulting coefficients easier to interpret.

[6] The mathematical relation between target error rates, calculated as [target]/([target] + [correct]), and nontarget errors is inconsequential here—calculating target rates as [target]/[total] minimizes the relation while producing essentially the same coefficients, standard errors, *p* values, figures, and so on. Note also that, mathematically, greater nontarget rates would increase odds-ratio effects (per Footnote 2), contra the predicted interaction.

[7] The estimate should, though, collapse across conditions (e.g., similarity). Estimating per experiment, not per subject or item, avoids sparse-matrix artifacts and trivial autocorrelations (e.g., with infrequent target errors, anything predicting more target errors may predict stronger effects merely by reducing sampling error and floor effects).

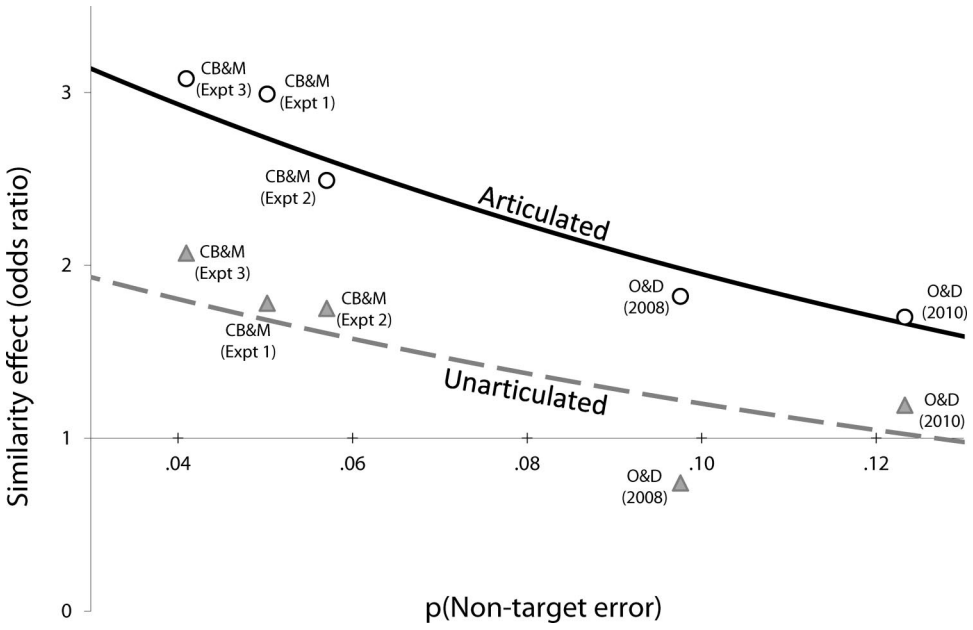

Figure 3. Similarity effects in articulated (circles) and unarticulated (triangles) speech, plotted against general measures of erroneous production. Regression lines depict fixed effects of the fitted model. The main effect of similarity places lines above 1.0 on the *y*-axis. The Similarity × Articulation interaction specifies a superior line for articulated speech. The Similarity × Error Rate interaction specifies the lines' slope, suggesting similarity effects are stronger when production is less error prone. CB&M = Corley, Brocklehurst, and Moat (2011); O&D = Oppenheim and Dell.

Results and Discussion

With 2,206 nontarget errors to estimate general error rates (described in Table 1), more error-prone production generally produced more target errors but weaker phonemic similarity effects (see Table 3). It is perhaps unsurprising that people reported more target errors when reporting more nontarget errors[8] (main effect of general error rate: $\beta = .13$, $\chi^2(1) = 42.43$, $p < .0001$), but this association does support the validity of using nontarget errors to index the general error proneness of a study.

Figure 3 shows that larger general error rates in these experiments were also associated with weaker similarity effects (interaction: $\beta = -.07$, $\chi^2(1) = 9.86$, $p < .002$). This Similarity × Error Rate interaction provides a log-odds slope for the regression lines in the figure, suggesting that a 1% increase in general error rate (e.g., from 5% to 6%) is accompanied by a 7% decrease in the log-odds similarity effect. This pattern holds for both articulated ($\beta = -.07$, $\chi^2(1) = 4.96$, $p < .03$) and unarticulated speech ($\beta = -.06$, $\chi^2(1) = 4.30$, $p < .04$) individually, reinforcing earlier claims that the major quantitative difference between the labs' results lies in the size of the main effect of phonemic similarity, not a simple main effect or interaction with articulation. Visually, the similarity effects in articulated and unarticulated speech form linearly separable sets when paired by experiment because the theoretically motivated inclusion of a mediating effect of overall error rate gives order to the variation between experiments. This overall pattern is easily simulated by a production model where inner speech involves attenuated subphonemic connections (see Figures 2b–2c). (The model also predicts a stronger Similarity × Error Rate interaction for articulated speech—which the data only hint at—but note that in the observed range of error rates, it does not necessarily predict a large log-odds difference, particularly when computing general error rate by collapsing across articulation conditions.)

Might this pattern extend to other data? Figure 4 plots similarity effects for the articulated conditions, plus data from four experiments using the SLIP procedure: Nooteboom (2005), Nooteboom and Quené (2008, two experiments), and Oppenheim and Dell (2008, SLIP task).[9] To compare across procedures with different error profiles, the predictor is now the overall probability of target errors, and following Nooteboom and Quené target errors in the SLIP task are defined as primed-for completed spoonerisms and anticipations.[10] Though methodological differences suggest caution in relating results across

[8] Note this does not mean target error rates scale in fixed proportion to nontarget error rates.

[9] Oppenheim and Dell's (2008) SLIP task used the same stimuli as their tongue-twister task. Data from Nooteboom (2005) and Nooteboom and Quené (2008) were provided by the authors, who recoded their 2005 similarity distinctions to match their 2008 ones. This is every published experiment I know that tests comparable similarity distinctions, uses consonant–vowel–consonant or consonant–vowel–consonant–consonant stimuli, and codes specific primed-for onset substitutions.

[10] Errors for SLIP tasks are experimenter coded. Adding SLIP perseverations would slightly increase target error counts (mostly for Oppenheim & Dell's, 2008, SLIP data) without substantially affecting the plotted effects.

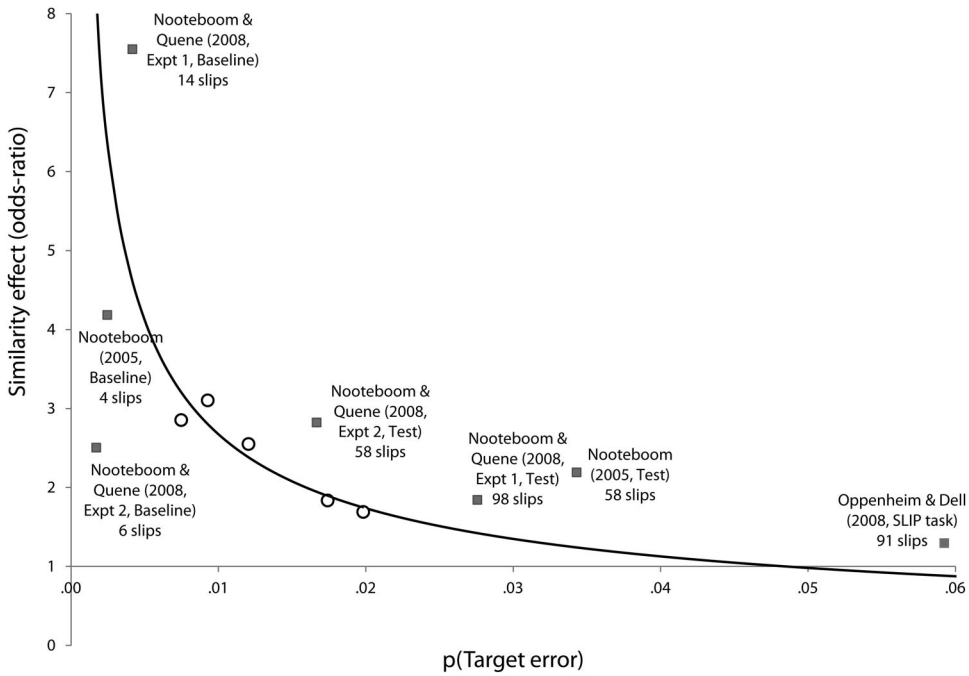

Figure 4. The Similarity × Error Rate interaction generalizes to new data. The regression line, fit only to the articulated tongue-twister data (circles), uses a power function (based on the simulations in Figure 2). Similarity effects from several SLIP tasks (squares) follow the same function.

paradigms, the SLIP data fit the expected pattern surprisingly well. There are three main points to note. First, the SLIP data span a wide range of error rates and actually show the expected Similarity × Error Rate pattern on their own. Second, the unprimed conditions in three experiments (plotted separately) elicited few target errors but particularly strong (and variable) similarity effects—precisely as predicted by simulations with Dell's (1986) model (see, e.g., Figure 2b) and more generally consistent with the notion that similarity more strongly constrains errors that lack support from other sources. Finally, where Nooteboom and Quené removed time pressure and warned participants about the task structure—changes intended to maximize editing, which generally eliminated several irrelevant speech error causes—this reduced overall error while eliciting the strongest similarity effect in their test data, following the prediction that particular error causes should tend to be more evident (in odds ratios) when other error causes contribute less.

To recap, error rates were used here to estimate generalized tendencies to produce target errors unrelated to the manipulations of interest. The influence of general error proneness does not preclude contributions from more specific error causes, but by indexing the noise in the speech production process, it provides a powerful, theoretically motivated resolution for a set of seemingly discrepant findings and has predictive value beyond the current data. This point is separate from the power issues discussed earlier because more accurate production produces fewer errors in the same number of trials, and practically speaking, this makes the estimates more vulnerable to sampling error and less able to support statistically detecting the same-sized effect.

The relationship between error rates and error patterns is a consequence of the fact that speech errors are simultaneously determined by multiple causes. It does not compel a particular feedback explanation and could certainly be described in terms of strategic speech monitoring (e.g., overworked monitors are less effective). Yet, as a domain-general phenomenon, it seems appropriate to posit a domain-general mechanism, and these patterns naturally arise from the kinds of stochastic selection algorithms that have played a crucial role in models of cognition for over half a century.

Conclusion

The first analysis demonstrated a robust Similarity × Articulation interaction, similar-sized in both labs' data. Thus, the SAH is supported by both data sets and generally speaking has strong support. It also identified the major discrepancies between the labs' data: Corley et al. (2011) elicited stronger main effects of similarity and fewer errors in general, making it easy to overlook the Similarity × Articulation interaction.

The second analysis explained these discrepancies together as a consequence of the overdetermined nature of speech errors. Phonemic similarity effects, like other "good" error patterns, are generally greater when error likelihood (indexing support from other error-causing factors) is low. Both inner speech and overt speech show this tendency, but the weaker similarity effect in inner speech is more easily obscured as phoneme selection grows less deterministic. These results are clearly predicted by a model where inner speech involves attenuated phoneme–feature connections and more generally suggest that successful speech error research

requires finding a sweet spot between too much randomness and not enough data.

Corley et al. (2011) questioned Oppenheim and Dell's (2008, 2010) claim that inner speech involves less robust access to subphonemic information. The data however, fully support that claim by showing that phonemic similarity effects are consistently weaker in unarticulated inner speech.

References

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.

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

Bates, D., & Maechler, M. (2010). lme4: Linear mixed-effects models using S4 classes [R Package Version 0.999375–35]. Retrieved from http://CRAN.Rproject.org/package=lme4

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonological units. *Phonology, 6,* 201–251.

Corley, M., Brocklehurst, P. H., & Moat, H. S. (2011). Error biases in inner and overt speech: Evidence from tongue twisters. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 162–175.

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

Dell, G. S., Burger, L. K., & Svec, W. R. (1997). Language production and serial order: A functional analysis and a model. *Psychological Review, 104,* 123–147.

Dell, G. S., Schwartz, M. F., Martin, N., Saffran, E. M., & Gagnon, D. (1997). Lexical access in aphasic and nonaphasic speakers. *Psychological Review, 104,* 801–838.

Freud, S. (1958). *The psychopathology of everyday life.* New York, NY: Macmillan. (Original work published 1901)

Garrett, M. F. (1980). Levels of processing in sentence production. In B. Butterworth (Ed.), *Language production* (pp. 177–220). New York, NY: Academic Press.

Geva, S., Bennett, S., Warburton, E., & Patterson, K. (2011). Discrepancy between inner and overt speech: Implications for post-stroke aphasia and normal language processing. *Aphasiology, 25,* 323–343.

Goldrick, M., & Rapp, B. (2007). Lexical and post-lexical phonological representations in spoken production. *Cognition, 102,* 219–260.

Gordon, J. K., & Dell, G. S. (2003). Learning to divide the labor: An account of deficits in light and heavy verb production. *Cognitive Science, 27,* 1–40.

Harley, T. A. (2010). *Talking the talk: Language, psychology and science.* New York, NY: Psychology Press.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: Computational basis and neural organization. *Neuron, 69,* 407–422.

Hubbard, T. L. (2010). Auditory imagery: Empirical findings. *Psychological Bulletin, 136,* 302–329.

Huettig, F., & Hartsuiker, R. J. (2010). Listening to yourself is like listening to others: External, but not internal, verbal self-monitoring is based on speech perception. *Language and Cognitive Processes, 25,* 347–374.

Huettig, F., Rommers, J., & Meyer, A. S. (2011). Using the visual world paradigm to study language processing: A review and critical evaluation. *Acta Psychologica, 137,* 151–171.

Killeen, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science, 16,* 345–353.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99,* 22–44.

Laganaro, M., & Zimmermann, C. (2010). Origin of phoneme substitution and phoneme movement errors in aphasia. *Language and Cognitive Processes, 25,* 1–37.

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

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111,* 309–332.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 1, pp. 103–189). New York, NY: Wiley.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology, 18,* 1–86.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

Nooteboom, S. G. (1969). The tongue slips into patterns. In A. G. Sciarone, A. J. van Essen, & A. A. van Raad (Eds.), *Nomen: Leyden studies in linguistics and phonetics* (pp. 114–132). The Hague, the Netherlands: Mouton.

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

Nooteboom, S. G., & 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.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115,* 39–57.

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. (2010). Motor movement matters: The flexible abstractness of inner speech. *Memory & Cognition, 38,* 1147–1160.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition, 114,* 227–252.

O'Seaghdha, P. G., Chen, J. Y., & Chen, T. M. (2010). Proximate units in word production: Phonological encoding begins with syllables in Mandarin Chinese but with segments in English. *Cognition, 115,* 282–302.

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.

R Development Core Team. (2010). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Rahman, R. A., & Aristei, S. (2010). Now you see it . . . and now again: Semantic interference reflects lexical competition in speech production with and without articulation. *Psychonomic Bulletin & Review, 17,* 657–661.

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

Severens, E., Janssens, I., Kühn, S., Brass, M., & Hartsuiker, R. J. (2011). When the brain tames the tongue: Covert editing of inappropriate language. *Psychophysiology, 48,* 1252–1257.

Shepard, R. N. (1957). Stimulus and response generalization: A stochastic model relating generalization to distance in psychological space. *Psychometrika, 22,* 325–345.

Stemberger, J. P. (1992). The reliability and replicability of naturalistic speech error data: A comparison with experimentally induced errors. In B. J. Baars (Ed.), *Experimental slips and human error: Exploring*

the architecture of volition (pp. 195–215). New York, NY: Plenum Press.

Stemberger, J. P. (2009). Preventing perseveration in language production. *Language and Cognitive Processes, 24,* 1431–1470.

Vicente, A., & Martinez Manrique, F. (2011). Inner speech: Nature and functions. *Philosophy Compass, 6,* 209–219.

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

Received October 28, 2010
Revision received July 15, 2011
Accepted July 25, 2011 ■