Clinical Aphasiology, Vol. 24, 1996, pp. 121–133

The Philadelphia Naming Test: Scoring and Rationale

April Roach, Myrna F. Schwartz, Nadine Martin, Rita S. Grewal, and Adelyn Brecher

A review of the clinical and experimental literature on confrontation naming in aphasia reveals procedural and scoring variation from study to study. Consider the matter of scoring a subject's multiple attempts at the target. While many scoring systems accept the best response, some select the first response (Blanken, 1990; Martin and Saffran, 1992) and others multiple responses per attempt (Kohn, 1984; Le Dorze and Nespoulous, 1989). Surprisingly, investigators give few details about how responses are extracted, although in practice this can be quite difficult to carry out reliably, especially with paraphasic or jargon aphasic speakers. (Nicholas, Brookshire, MacLennan, Schumacher, and Porrazzo, 1989, however, considered this issue in selecting the responses of non–brain-injured subjects performing the Boston Naming Test.)

Selecting a taxonomy for coding errors is also less than straightforward. For example, the term *phonemic paraphasia* may describe an error that bears any phonological resemblance to the target (Lesser, 1978), or that contains at least half the target's phonemes (Goodglass and Kaplan, 1983), or that overlaps the target in just one salient phonological unit (Kohn and Smith, 1994; Martin and Saffran, 1992).

In this paper, we describe a new naming test and a set of detailed procedures that have been developed for selecting and coding responses. This is part of a larger collaborative study that seeks to explain normal and aphasic error patterns in terms of an interactive, spreading activation model of word retrieval (Dell and O'Seaghdha, 1991, 1992; Gagnon, Schwartz, Martin, Dell, and Saffran, in press; Martin, Dell, Saffran, and Schwartz, 1994; Schwartz, Dell, Martin, and Saffran, 1994a). It is widely believed that error types seen in normal and aphasic naming are reflective of the stages of word retrieval. Semantic errors arise at the first stage, which involves retrieval of an abstract semantic-syntactic representation known as the lexical entry, or *lemma*. Phonemic paraphasias arise during

later retrieval or assembly of the phonological form. Other errors appear to reflect the interaction of information across these two stages. One is the so-called mixed error: a substituted word that bears both semantic and phonological resemblance to the target (e.g., "snake" for *snail*). Another is the *formal paraphasia* (malapropism; phonic verbal paraphasia): a word that bears a purely phonological resemblance to the target (e.g., "shaft" for *fish*). On interactive, two-stage models (Dell, 1986; Stemberger, 1985), formal paraphasias reflect the influence of Stage 2 phonological information on Stage 1 lexical retrieval.

Also reflecting interaction across the stages of retrieval is the quantitative relationship between phonological errors that comprise words (formal paraphasias) versus nonwords (phonemic paraphasias). A statistically reliable bias to produce word outcomes is viewed as additional evidence of phonological influences on lexical retrieval (Dell and Reich, 1981; Schwartz, Saffran, Bloch, and Dell, 1994b).

In the psycholinguistics literature, there is controversy concerning whether it is necessary to postulate such interaction or, alternatively, whether a strict, sequential model will suffice to account for the error and experimental data (Levelt, Schriefers, Vorberg, Meyer, Pechmann, and Havinga, 1991, and response in Dell and O'Seaghdha, 1991). Our group is addressing this issue in normal and aphasic subjects through a modeling effort that seeks to simulate actual naming data (proportions of correct responses and various error types) using a computational version of the Dell word retrieval model, developed by Dell and O'Seaghdha (1991). The procedures we have developed for extracting and coding responses make it possible to test directly the predictions of the model. In addition, these procedures make certain clinically relevant analyses easier and more transparent than is the case with other scoring systems. These analyses are illustrated with data from the first nine aphasic subjects that have been tested. First, however, we describe in detail the naming test and the system for extracting and coding responses.

METHOD

Philadelphia Naming Test (PNT)

The PNT consists of 175 high-, medium-, and low-frequency nouns that range in length from 1 to 4 syllables (based on Francis and Kučera, 1982, noun frequencies; low: 1–25; medium: 26–78; high: 80–2110). These items were selected from a larger set of 277 items, based on the consistency with which a group of 30 control subjects responded to the designated target.

Each item on the PNT was named correctly by at least 85% of the control subjects; 136 items were named correctly by all 30 controls.

The items of the PNT are digitized for computerized display on a Macintosh computer. Sessions are tape-recorded for later analysis, and subsequently transcribed by two speech–language pathologists. Trials begin with a brief tone, which coincides with the onset of the picture. Naming latencies are measured from the onset of this tone to the onset of the subject's responses.

Subjects are asked to name the picture as soon as possible after it appears on the computer screen. Subjects are also encouraged to respond with a single word, that is, to avoid descriptions and multiple attempts. Trials end with the subject's response, or after 30 seconds have elapsed. Each trial terminates with the experimenter's saying the correct target in order to discourage perseverative rumination that might interfere with subsequent trials. We are currently investigating how this feedback affects the test-retest reliability.

Scoring System for the PNT

Labeling the Transcript for Scoreable Attempts. Some fluent aphasic subjects have difficulty restricting their naming attempts to a single utterance. Our instructions to the subjects are designed to minimize these multiple attempts, but they do not eliminate them. As noted previously, the decision about which response to score is a problematic one. Selecting only one response when many are produced may involve discarding potentially useful information. Our solution is to identify up to three responses on each trial:

- The initial attempt (I)—minimally a CV or VC utterance, excluding schwa; this may be a complete utterance or a fragment;
- The first complete attempt (C) in cases where I is a fragment (i.e., not complete); and
- The final complete attempt (F).

Let us take as an example the target *basket,* to which a subject responds as in (1):

(1) "/bə/ /bə/ /bɪs-/ /"bɪskət/ basket."

In identifying the initial attempt, we chose to bypass single phonemes and consonants combined with schwa as an aphasic subject prepares to name a word. The initial attempt or "I" is "/bɪs-/," because neither of the first two utterances satisfies the minimal criterion for I. "/bɪs-/" is also

labeled as a fragment or I-f, because upon listening to the tape, it is self-interrupted. Fragments are identified on the basis of auditory cues indicating self-interruption or continuous correction, i.e., segment duration, intonation, or pausing. We also identify as fragments monosyllabic responses to targets of three or more syllables (e.g., "ther-" for "thermometer"). When the initial attempt is a fragment, we label it as such and continue to search for the first complete attempt. Continuing with example (1), "/biskət/" is labeled as the first complete attempt or "C" and "basket" is the final complete attempt or "F."

Consider now the responses to *basket* exemplified in (2) and (3):

(2) /bʌskət/
(3) /bæn/ basket

In (2), the subject has produced only a single attempt at the target, which appears from the auditory evidence to be complete. This response is labeled ICF. In (3), "/bæn/" is both the initial attempt and the first complete attempt, based on the auditory evidence, and is labeled "IC." Whenever the first attempt is complete, it is labeled as both I and C. "Basket" is the final complete attempt and is labeled "F."

Our system also identifies trials on which the subject produced *multiple phonological attempts* at a target, also called "conduites d'approache" (Joanette, Keller, and Lecours, 1980; Kohn, 1984). These trials contain at least three different attempts at a single target, excluding the correct response.

Once I, C, and F attempts are labeled on the typed transcript, they are entered onto the scoresheet for coding (Figure 1).

Coding. One of the unique features of this coding system is that each response is given a two-level code (Table 1).

Level 1 classifies the response at a lexical level, and Level 2 classifies the response at a phonological level. The Level 1 code *target attempt* is defined as a response that bears phonological similarity to the objective target; specifically, both objective target and subject's response share one or more phonemes at corresponding syllable and word positions or two or more phonemes at any position, including stressed vowels. Consider the following two examples of target attempts:

(4) ghost→ /goθ/
(5) fish→ shaft

In example (4) there are two phonemes in common in the same positions as the target (/g/ and /o/). In (5) there are two shared phonemes (/f/ and /ʃ/), but they are not in the same positions.

The next Level 1 code listed in Table 1 is *semantic* error, which is a response that bears a semantic or associated relationship to the target. A *mixed* error is a response that meets criteria for both phonological similarity and semantic

No.	Target	Initial Attempt (I) Response	Codes L. 1	Codes L. 2	First Complete Attempt (C) Response	Codes L. 1	Codes L. 2	Final Attempt (F) Response	Codes L. 1	Codes L. 2	MPA
1	ghost	/goθ/	TA	S/NW							
2	fish	shaft	TA	S/W							
3	pirate	bugganeer	S	S/NW							
4	van	bus	S								
5	snail	snake	M								
6	banana	camp	O								
7	cane	/tis/	O	S/NW							

Figure 1. The Philadelphia Naming Test (PNT) scoresheet with sample subject responses and codes entered.
Note: See Table 1 for Level I (L.1) lexical codes and Level 2 (L.2) phonological codes.

Table 1. PNT Naming Codes

	Level 1 Codes (Lexical Level)	*Level 2 Codes (Phonological Level)*
Correct (√)		Sound deviation with:
Target attempt (TA)	Phonological similarity	Word outcome (S/W) Nonword outcome (S/NW)
Semantic (S)	Associated relationship to target	Indeterminate outcome (S/I–fragments)
Mixed (M)	Phonological similarity and semantic relationship	
Other (O)	Unrelated response	
Blend (B)	Combination of two semantically related items	
Picture part (PP)	Response refers to part of the picture shown	
Perseveration (P)	Duplication of a previous response	
Description (D)	Characterization of target	
No response (NR)	Includes comments, e.g., "I don't know"	

relationship. An example is "snake" for *snail.* The Level 1 code, *other,* refers to a response unrelated to the target. A *blend* is a response combining two identifiable parts of semantically related items (e.g., "/bə-'næ-pəl/" for *pineapple,* combining "banana" and "pineapple").

Level 2 codes are: sound deviation with word outcome, abbreviated **S/W**, and sound deviation with nonword outcome, abbreviated **S/NW**. For fragments containing errors, there is also the Level 2 code, sound deviation with indeterminate outcome, abbreviated as **S/I**. The term *sound deviation* refers to phonemic paraphasia and not to errors secondary to motor speech disorders. Aphasic subjects with motor speech disorders were excluded from the study. In those rare cases where there was some ambiguity between phonemic paraphasia and apractic error, the response was scored correct.

As mentioned earlier, the word-nonword distinction embodied in Level 2 codes is important to the theoretical interpretation of error patterns

(Martin and Saffran, 1992). This distinction is not, however, respected in conventional scoring systems.

In the scoresheet (Figure 1), "/goθ/" produced for *ghost* is coded **TA** at Level 1, because of phonological similarity. The Level 2 code is **S/NW** for nonword outcome. Example 2, "shaft" for *fish,* is also a target attempt, but with a word outcome, thus the code **TA S/W**.

All target attempts receive a Level 2 code. Other Level 1 errors may also receive a Level 2 code in cases where there is phonemic deviation from lexical productions other than the target. In example 3, the response "bugganeer" given to *pirate* is coded at Level 1 as a semantic error (assuming "buccaneer") and at Level 2 as **S/NW**. Example 7, "/tʌs/" for *cane,* is coded **O S/NW**.

Our two-level coding system readily converts to more conventional codes to allow for comparisons with other studies. For example, our target-related nonword (**TA S/NW**) is a target-related neologism or phonemic paraphasia in conventional codes; a target-related word (**TA S/W**) is a formal paraphasia or phonic verbal paraphasia; and *other* with **S/NW** is an abstruse neologism.

Subjects

Controls. Control subjects were 30 non–brain-injured controls ranging in age from 40 to 75 years (M 56.3; SD 11.4).

Experimental Subjects. The experimental subjects were nine right-handed individuals with posterior left hemispheric CVA and fluent aphasia ranging from 28 to 84 years of age (M 54.1; SD 16.8). Based on the Boston Diagnostic Aphasia Examination classification system (Goodglass and Kaplan, 1983), three exhibited conduction aphasia, three Wernicke's aphasia, two anomic aphasia, and one transcortical sensory aphasia. All but one were less than 1 year post onset; seven of the nine were less than 6 months post onset.

RESULTS

Inter-rater Agreement

Inter-rater agreement has been calculated on the complete data from six subjects, first for labeling the responses and then for coding them. For these subjects, labeling and coding were carried out by two or three members of the research team, working independently. For each comparison of

Table 2. Percentage Agreement in Labeling Initial (I), Complete (C), and Final (F) Responses and in Coding the Labeled Responses

	Labeling			*Coding*	
Subject	*Initial (I)*	*Complete (C)*	*Final (F)*	*Level 1*	*Level 2*
G.B.	97	93	93	93	94
N.C.	99	98	99	93	97
J.F.	97	97	93	90	100
L.H.	98	87	98	93	93
G.L.	100	95	97	96	92
B.M.*	98	97	98	86	89

*Entry represents means for 3 coders.

interest the point-to-point percentage agreement among each pair of raters was calculated according to the following formula (Nicholas et al., 1989):

$$\frac{\text{Number of agreements}}{\text{Number of agreements + disagreements}} \times 100$$

Where three raters were used, the percentage agreement for each pair was averaged. Table 2 shows that there is very high agreement on the labeling of I, C, and F attempts, as well as on the assignment of Level 1 and Level 2 codes.

Response Accuracy

Control Subjects. As expected, control subjects were highly accurate. Mean percent correct for Initial, Complete, and Final attempts were, respectively, 96 (SD 7.0), 96 (6.9), and 97 (6.4).

Aphasic Subjects. Table 3 shows the aphasic subjects' performance at I, the initial attempt. Correct scores ranged from 26% to 92%. There were three measures that bear on phonological production: fragments, errors containing Level 2 codes, and trials with multiple phonological attempts or MPAs. As expected, the three measures correlated significantly with one another. For example, the Spearman rank order correlation coefficient for Level 2 codes and fragments was .76 ($p < .05$); and for Level 2 codes and MPAs, it was .84 ($p < .05$). There was also a negative correlation between Level 2 codes and percent correct, which approaches significance (Rho = −.68; $p = .055$). Thus, phonemic deviation, as measured by the percentage of errors containing Level 2 codes, is a predictor of accuracy in our sample. On the other hand, it did not appear to be related to subtype of fluent aphasia.

Table 3. Aphasic Subjects' Performance at Initial Attempt (I)

Subject	*Correct (%*)*	*Fragments (%)*	*Errors with L. 2 Codes (%)*	*Trials with MPAs (%)*	*Clinical Class*
W.B.	92	4	29	2	Wernicke
T.T.	92	2	7	0	Anomic
B.M	82	3	16	0	Transcortical Sensory
N.C.	74	2	62	1	Conduction
J.F.	55	3	8	1	Anomic
L.H.	54	22	73	7	Conduction
H.B.	53	22	66	16	Conduction
G.B.	39	6	31	2	Wernicke
G.L.	26	21	67	26	Wernicke

*Correct percentages calculated on 175 trials; MPAs = multiple phonological attempts.

Subjects' increase in scores from Initial, to Complete, to Final can be taken as a measure of self-correction. Percent correct scores for each of the nine subjects at I, C, and F is shown in Figure 2. Subjects differ in their naming success at I and also in their tendencies to self-correct across a trial. Four of the nine subjects showed greater than 30% improvement from I to F, despite low scores at I.

In a subsequent analysis we sought to determine whether self-correction for these four subjects was associated with particular error types. This analysis focused on the change from C (first complete attempt) to F (final complete attempt) and on the three error types that were most prevalent in these subjects: target-related nonwords (neologisms), target-related words (formal paraphasias), and semantic errors.

The data are shown in Table 4. The null hypotheses that the three error types were equally likely to be corrected at F was tested by Chi Square and rejected for subjects L.H. (10.2; $p < .01$) and G.L. (9.8; $p < .01$). Inspection of their data showed that semantic errors were less likely to be corrected than target-related errors of both types—nonwords and words (neologisms and formals). Subject G.B. also showed a trend in this direction. The pattern did not hold for H.B.

DISCUSSION

The Philadelphia Naming Test (PNT) and its associated scoring system were developed to address theoretical issues concerning the nature of naming errors and what they reveal about the mental processes underlying normal

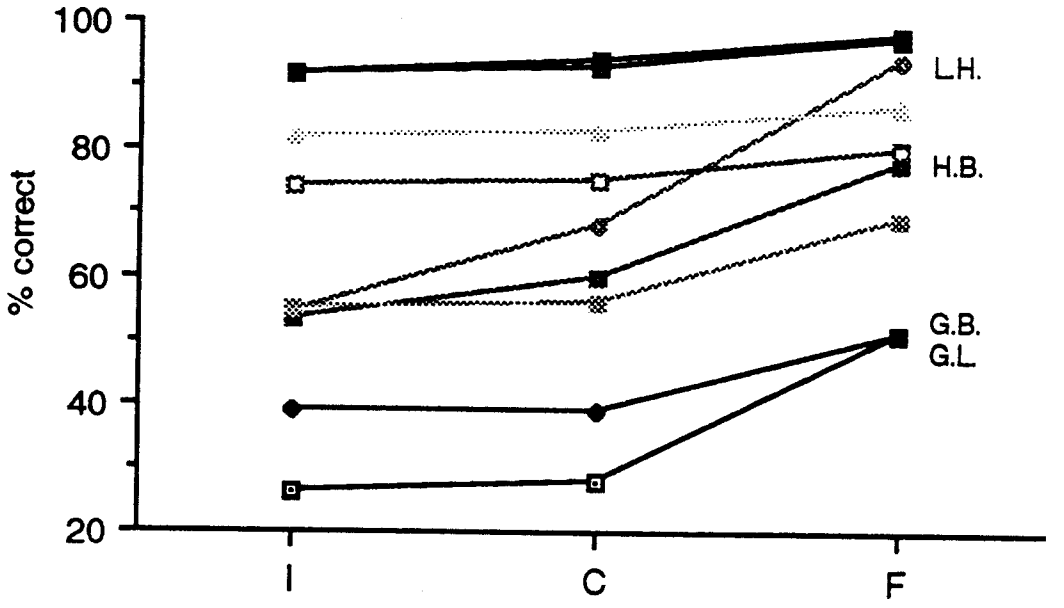

Figure 2. Percent correct scores (out of 175) for aphasic subjects at initial attempt (I), first complete attempt (C), and final attempt (F). The four identified subjects (L.H., H.B., G.B., G.L.) showed greater than 30% improvement from I to F.

and pathological word retrieval. For such purposes, it is important to know what the initial attempts at retrieval yield, as well as whether success is ultimately achieved. The division of utterances into first, first complete, and final complete responses fulfills this requirement. It also makes it quite simple to measure self-correction across a trial and to compare different error types for likelihood of being corrected. For three of the four subjects in our study who showed the greatest tendency to self-correct, errors that bear a phonological relation to the target were more likely to be corrected than semantically related errors. We plan in the future to conduct longitudinal studies to determine whether particular error and self-correction patterns are associated with good recovery on follow-up testing.

One of the motivations for the two-level coding system was to quantify the production of errors that may be said to reflect Stage 2 retrieval operations, that is, selection and assembly of a word's phonological form. The percentage of errors with Level 2 codes is one measure of Stage 2 impairment. Even in our small sample, the ranking of subjects on this measure correlated negatively with their overall naming success. This shows the importance of phonological impairments in the naming performance of fluent aphasic subjects. In our sample, this relationship seems to hold across the various subtypes of fluent aphasia; we are interested in seeing whether this holds true over larger samples as well. At this time the length

Table 4. Errors at First Complete Attempt (C) Subsequently Corrected at Final Attempt (F)

	Target-Related Nonword (Neologism)			*Target-Related Word (Formal Paraphasia)*			*Semantic Error*		
Subject	*No. Corr.*	*Total*	*% Corr.*	*No. Corr.*	*Total*	*% Corr.*	*No. Corr.*	*Total*	*% Corr.*
G.B.	3	8	38	6	16	38	4	18	22
H.B.	11	27	41	11	22	50	9	17	53
L.H.	23	25	92	12	12	100	5	9	56
G.L.	15	43	35	18	36	50	0	12	0

of the PNT precludes its use in a clinical setting. The coding system, however, could be adapted for other naming tests.

As mentioned earlier, mixed errors are an important consideration in the debate concerning whether information at one stage of retrieval affects processing at other stages. The question is whether mixed errors are real; that is, whether the degree of phonological relatedness among semantic errors and their targets occurs at a level greater than chance (the prediction of interactive accounts). In a recent study, our group used data derived from the PNT to argue for greater than chance occurrence of phonological relatedness in the semantic errors generated by nonaphasic and aphasic subjects (Martin, Gagnon, Schwartz, Dell, and Saffran, in press).

Our investigations of error patterns in aphasic and nonaphasic speakers are guided by the interactive, spreading activation model of word retrieval set forth in Dell (1986) and computationally implemented in Dell and O'Seaghdha (1991). The Dell and O'Seaghdha model incorporates a three-level lexical network: semantic nodes connect to lexical (lemma) nodes, which connect to phoneme nodes. All connections are bidirectional; that is, activation flows from higher to lower nodes and back up. Two parameters are important: the rate at which activation flows in the network, reflecting the strength of connections among nodes, and the rate at which activated nodes decay. Varying these parameters affects both the frequency of error and the types of error that the model produces. For example, weak connection strength promotes nonword errors (phonemic paraphasias; target-related neologisms), while a high rate of decay differentially promotes errors that comprise words, especially mixed errors and formal paraphasias. Manipulating these parameters alone and in combination offers a way to account for individual differences in the naming profiles of fluent aphasic speakers (Martin et al., 1994; Schwartz et al., 1994a), in a manner that assimilates the aphasia data within a model of normal production.

REFERENCES

Blanken, G. (1990). Formal paraphasias: A single case study. *Brain and Language, 38,* 534–554.

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

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: A comment on Levelt et al. *Psychological Review, 98,* 604–614.

Dell, G. S., & O'Seaghdha, P. G. (1992). Stages of lexical access in language production. *Cognition, 42,* 287–314.

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.

Francis, W. N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Gagnon, D. A., Schwartz, M. F., Martin, N., Dell, G. S., & Saffran, E. M. (in press). The origins of form-related paraphasias in aphasic naming. *Brain and Language.*

Goodglass, H., & Kaplan, E. (1983). *The assessment of aphasia and related disorders,* (2nd ed.). Philadelphia: Lea & Febiger.

Joanette, Y., Keller, E., and Lecours, A. R. (1980). Sequence of phonemic approximations in aphasia. *Brain and Language, 11,* 30–44.

Kohn, S. E. (1984). The nature of the phonological disorder in conduction aphasia. *Brain and Language, 23,* 97–115.

Kohn, S. E., & Smith, K. L. (1994). Distinctions between phonological output deficits. *Applied Psycholinguistics, 15,* 75–95.

Le Dorze, G., and Nespoulous, J. L. (1989). Anomia in moderate aphasia: Problems accessing the lexical representation. *Brain and Language, 37,* 381–400.

Lesser, R. (1978). *Linguistic investigations of aphasia.* New York: Elsevier.

Levelt, W. J. M., Schriefers, H., Vorberg, D., Meyer, A. S., Pechmann, T., & Havinga, J. (1991). The time course of lexical access in speech production: A study of picture naming. *Psychological Review, 98,* 122–142.

Martin, N., Dell, G. S., Saffran, E. M., & Schwartz, M. F. (1994). Origins of paraphasias in deep dysphasia: Testing the consequences of a decay impairment to an interactive spreading activation model of lexical retrieval. *Brain and Language, 47,* 609–660.

Martin, N., Gagnon, D. A., Schwartz, M. F., Dell, G. S., & Saffran, E. M. (in press). Phonological facilitation of semantic errors in normal and aphasic speakers *Language and Cognitive Processes.*

Martin, N., & Saffran, E. M. (1992). A computational account of deep dysphasia: Evidence from a single case study. *Brain and Language, 43,* 240–274.

Nicholas, L. E., Brookshire, R. H., MacLennan, D. L., Schumacher, J. G., & Porrazzo, S. A. (1989). Revised administration and scoring procedures for the Boston Naming Test and norms for non–brain-damaged adults. *Aphasiology, 3,* 569–580.

Schwartz, M. F. (1987). Patterns of speech production deficit within and across aphasia syndromes: Application of a psycholinguistic model. In M. Coltheart, G. Sartori, & R. Job (Eds.), *The cognitive neuropsychology of language,* (pp. 161–199). London: Lawrence Erlbaum Associates.

Schwartz, M. F., Dell, G. S., Martin, N., & Saffran, E. M. (1994a). Normal and aphasic naming in an interactive spreading model of lexical retrieval. *Brain and Language, 47,* 391–394.

Schwartz, M. F., Saffran, E. M., Bloch, D. E., & Dell, G. S. (1994b). Disordered speech production in aphasic and normal speakers. *Brain and Language, 47,* 52–88.

Stemberger, J. P. (1985). An interactive model of language production. In A. Ellis (Ed.), *Progress in the psychology of language, Vol. 3.* London: Lawrence Erlbaum Associates.