Available online at www.sciencedirect.com
SCIENCE DIRECT®

Journal of Memory and Language 54 (2006) 145–160

Journal of Memory and Language

www.elsevier.com/locate/jml

Simulating consistency effects and individual differences in nonword naming: A comparison of current models

Jason D. Zevin [a,*], Mark S. Seidenberg [b]

[a] Sackler Institute for Developmental Psychobiology, Weill Medical College of Cornell University, 1300 York Ave., Box 140, New York, NY 10021, USA
[b] University of Wisconsin-Madison, USA

Received 10 May 2004; revision received 29 August 2005

Abstract

The mechanisms underlying nonword pronunciation have a been a focus of debates over dual-route and connectionist models of reading aloud. The present study examined two aspects of nonword naming: spelling-sound consistency effects and variability in the pronunciations assigned to ambiguous nonwords such as MOUP. Performance of a parallel distributed processing model was assessed over multiple runs, representing multiple subjects with varying reading experience. The model provided a good account of behavioral data concerning these phenomena. In contrast, the Dual Route Cascaded model does not produce consistency effects and does not account for the alternative pronunciations that subjects produce. The results highlight the importance of considering multiple aspects of a phenomenon such as nonword naming in assessing computational models.

Keywords: Reading; Statistical learning; Computational models; Nonwords; Individual differences

Word and nonword reading are among the most extensively studied areas in cognitive science and neuroscience (see Posner, Abdullaev, McCandliss, & Sereno, 1999; Rayner, Foorman, Perfetti, Pesetsky, & Seidenberg, 2001 for overviews). Although several word reading models have been proposed, considerable attention has focused on the contrast between dual-route (Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) and connectionist (see Harm & Seidenberg, 1999, 2004; Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989 for discussion) approaches, both of which have evolved over many years. In the present article we consider how recent versions of these models fare with respect to the task of reading nonwords aloud, a task that has long been used to assess their adequacy (Besner, Twilley, McCann, & Seergobin, 1990; Seidenberg, Plaut, Petersen, McClelland, & McRae, 1994).

The dual-route model of reading aloud (e.g., Coltheart et al., 2001) holds that pronouncing letter strings (words and nonwords) involves a lexical route consisting of knowledge of individual words, and a nonlexical route consisting of rules for translating spellings to sounds. Words whose pronunciations violate the rules ("exceptions" such as PINT) can only be pronounced correctly via the lexical route. Nonwords (such as NINT) can only be pronounced using the rules. The central dogma of the dual-route approach (Seidenberg,

* Corresponding author.
E-mail address: jdz2001@med.cornell.edu (J.D. Zevin).

0749-596X/$ - see front matter
doi:10.1016/j.jml.2005.08.002

1995) is that the two routes are required to account for the ability to read these differing types of stimuli.[1]

In the Dual Route Cascaded (DRC) model (Coltheart et al., 2001), the lexical route is construed as an associative network with nodes corresponding to words; it is sensitive to lexical statistics (e.g., the frequencies of words, their orthographic and phonological similarity to one another). The sublexical route operates categorically and deterministically, applying rules that state the valid correspondences between spelling and sound but abstract away from statistical properties such as how often the rules are used across words. The dual-route framework is representative of a general approach to the study of language and other phenomena which holds that distinct mechanisms are involved in the abstraction of categorical, symbolic rules versus the memorization of arbitrary facts (e.g., Pinker, 1999).

In the connectionist (or "triangle") framework, both words and nonwords are pronounced using the same network of weighted connections among units in a parallel distributed processing (PDP) architecture (e.g., Seidenberg & McClelland, 1989). Such models do not incorporate the distinction between a lexical level containing memorized forms of words and a set of rules for decoding novel words. This approach is representative of a broader theoretical stance which asserts the primacy of statistical learning in the acquisition and use of language and other types of knowledge (Kirkham, Slemmer, & Johnson, 2002; Saffran, Newport, Aslin, & Tunick, 1997; Seidenberg, 1997). Thus, the differences between the dual-route and connectionist approaches to spelling-to-sound decoding instantiate contrasting views about the characterization of lexical and other types of knowledge.

Regularity and consistency effects in word reading

Several recent studies have attempted to adjudicate between the competing models by considering the effects of two structural properties of words: "regularity" and "consistency." The concept of regularity is central to the dual-route approach: regular words are ones whose pronunciations are correctly specified by spelling-sound rules. Many words are regular (e.g., MUST and PAVE), as are all nonwords (e.g., NUST and MAVE, which according to this theory can only be pronounced by rule). Typically the rules involve mappings between graphemes and phonemes, but other types of rules are sometimes proposed (see Coltheart et al., 2001, who included multigrapheme and context-sensitive rules). The critical property of regularity is that it is a categorical concept: a word's pronunciation is either correctly specified by the pronunciation rules or not. Words whose pronunciations violate the rules produce regularity effects: longer latencies and/or more errors than for rule-governed words. These arise because the lexical and nonlexical routes produce conflicting pronunciations for exception words (e.g., the lexical route yields the correct pronunciation of PINT and the nonlexical route, the regularization /pɪnt/).

Consistency, in contrast, is a statistical concept central to the triangle approach. The degree of consistency in the mapping between spelling and sound varies continuously. Consistency defined in terms of rimes, sometimes also called word bodies, has been examined most thoroughly in previous research (e.g., Jared, McRae, & Seidenberg, 1990) because this unit happens to be salient given the structure of English monosyllables (see Seidenberg & McClelland, 1989; Treiman, Mullennix, Bijelac-Babic, & Richmond-Welty, 1995 for discussion). For example, the rime -UST is highly consistent because it is always pronounced /ʌst/ in monosyllabic words. A rime such as -AVE is inconsistent because it is usually pronounced as in SAVE, PAVE, and GAVE but differently in HAVE. Other factors being equal, words with less consistent spelling-sound correspondences will be more difficult to read aloud than words with more consistent correspondences. Connectionist models have been highly successful at accounting for the effects of differing degrees of consistency observed in many studies (Cortese & Simpson, 2000; Jared et al., 1990; Jared, 1997).

It is important to note that although rimes exert the greatest influence on naming latencies (and thus have been the primary focus of interest), there are secondary statistical phenomena involving other units, ranging from graphemes and onset-nucleus units to entire words (as in homographs such as WIND). Rimes happen to be most prominent, and the effects of inconsistencies over rimes are of a magnitude that can be readily observed in simple behavioral studies. Effects of other units are weaker but can be picked up in careful experiments (e.g., Treiman, Kessler, & Bick, 2003). We emphasize the point that inconsistency occurs over multiple orthographic grain sizes because it is often obscured in factorial experimental designs. For example, in studies that directly compare consistency and regularity effects

[1] As Harm and Seidenberg (2004) noted, the term "dual-route model" is ambiguous. Sometimes it refers to visual and phonologically mediated processes in the access of *meaning*, and other times to lexical and nonlexical processes for generating *pronunciations* (see also Coltheart, 2000). The claim that reading may involve both visual and phonologically mediated processes is not specific to any one theory; it reflects the fact that in alphabetic writing systems letter strings can be associated with both meanings and pronunciations. However, the claim that two mechanisms are required to pronounce letter strings is specific to the dual-route moded by Coltheart and colleagues. We use the term "dual-route model of reading aloud" in reference to models incorporating this claim.

(e.g., Cortese & Simpson, 2000; Jared, 2002), consistency is typically operationalized at the rime level and regularity at the grapheme level. In fact, inconsistencies exist at multiple levels of orthographic structure, and rules are sometimes defined in terms of units other than graphemes. The important difference between the concepts is that consistency is statistical, whereas regularity is categorical. A statistical learning model such as Seidenberg and McClelland (1989) or Harm and Seidenberg (1999) will pick up on spelling-sound consistencies over many units, to the extent that they occur in the corpus of training examples, constrained by properties of network architecture such as the input (orthographic) and output (phonological) representations, and number of hidden units. In dual-route models, pronunciations are either rule-governed or not, and the procedure by which rules are applied does not take into account statistical properties such as how often they apply across words.

In the triangle framework, regularity effects arise from spelling-sound inconsistencies. Consider, for example, the word ROLL, which is treated as an exception in the DRC model (Coltheart et al., 2001) because the rule governing the grapheme O is associated with the pronunciation that occurs in DOT, TOP, ROCK, and many other words. Hence the DRC model predicts that the exception ROLL should be more difficult than a word such as ROCK that obeys the rules. The triangle model makes a similar prediction but for different reasons. ROLL is more difficult than ROCK because ROLL exhibits spelling-sound inconsistencies at several levels. At the rime level it is consistent with TROLL, POLL, and TOLL but inconsistent with DOLL and MOLL. At the grapheme level, the O in ROLL is consistent with words such as POST and MOLD but inconsistent with every word in which O is pronounced differently (e.g., ROT, ROB, SOT, etc.). The weights governing the orthography–phonology computation encode the statistics of the mapping between spelling and sound. Hence ROLL will be more difficult than ROCK because it exhibits greater spelling-sound inconsistency.

Thus, both theories can in principle account for regularity effects. However, they differ with respect to words that are rule-governed (according to DRC) but inconsistent (according to the triangle model). Such words were designated "regular but inconsistent" by Glushko (1979). PAVE, for example, is rule-governed according to DRC, but inconsistent according to the triangle model because of the irregularly pronounced neighbor HAVE. DRC predicts that PAVE should be as easy to pronounce as PANE, which is rule-governed but also highly consistent because it has no close irregular neighbors. In contrast, the triangle model predicts that the two types of rule-governed words should differ: inconsistent words such as PAVE should be more difficult than consistent words such as PANE, a finding that has been observed in numerous studies (see Jared et al., 1990, who summarized the results of more than a dozen experiments).

Consistency effects have been taken as strong evidence against the dual-route theory. However, according to Coltheart et al. (2001), the consistency effects observed in previous studies were due to confounding factors: the presence of words that DRC treats as "exceptions" among the inconsistent words, and left to right misanalyses of words ("whammies") that they asserted occur more frequently in inconsistent words. They applied this analysis to a single study in the literature (Jared, 1997). However, Jared's data yield a consistency effect even with these factors taken into account (i.e., if the exception and whammy items are removed from the data set). Moreover, the two factors do not account for consistency effects in other studies. For example both Cortese and Simpson (2000) and Jared (2002) conducted naming studies that explicitly compared the regularity and consistency factors, and examined the performance of the Plaut et al. (1996) and Coltheart et al. (2001) models on their stimuli. Whereas human subjects and the Plaut et al. (1996) model produced large consistency and small (for the Cortese & Simpson stimuli) or null (for the Jared stimuli) regularity effects, the Coltheart et al. (2001) model did the opposite, producing very large regularity effects and null (or reversed) consistency effects.

In summary, consistency effects are critically important because the triangle model predicts they should occur whereas the DRC model predicts they should not, in the absence of confounding factors. Existing data indicate that such effects occur for words, presenting a substantial problem for the dual-route approach, including DRC 2001, which does not produce them.

Consistency and variability in nonword naming

Although consistency effects for words strongly favor the triangle approach, it is also important to consider whether similar effects occur for nonwords. In dual-route models, all known words can be pronounced via the lexical mechanism; the grapheme–phoneme correspondence rules that constitute the second route are mainly relevant to generalization (i.e., nonword pronunciation). Because of the serial nature of rule application in the 2001 DRC model and the settings of other parameters, the nonlexical route operates so slowly as to have little effect on the pronunciation of words. In contrast, connectionist models make the strong claim that generalization arises from passing novel items through the same network that encodes knowledge of words. Hence, whether such networks can generate correct nonword pronunciations is important. In fact there has been some question as to whether they can. Seidenberg and McClelland (1989) initially emphasized the fact that their network produced correct output for both rule-governed words and exceptions. It was later noted that the model's performance on nonwords was less accurate than people's (Besner et al., 1990). The Seidenberg and McClelland model correctly generalized to simple nonwords such as NUST, illustrating generalization without rules; however, it made errors on more difficult nonwords such as JINJE. Plaut et al. (1996) and Seidenberg et al. (1994) subsequently showed that with improved phonological representations such models could also pronounce nonwords at levels comparable to human performance and slightly better than the 1993 version of DRC (Coltheart et al., 1993). However, more recent work discussed below (e.g., Treiman et al., 2003) raised further questions about both types of models' capacities to account for nonword performance.

The present research focused on two aspects of nonword naming. The first is nonword consistency effects. We know that performance on a "rule-governed" word such as PAVE is affected by an inconsistent neighbor such as HAVE. The same effect has been reported for nonwords: MAVE is also harder to pronounce than NUST (Glushko, 1979). As with words, the DRC model predicts that such effects should not occur; nonwords are pronounced by the rule component with little if any input from the lexical route. Nonword consistency effects also seem more compatible with the PDP approach; the effects arise from the same source as for words, shaping of the weights by exposure to words with conflicting spelling-sound correspondences. Thus weight settings that produce an inconsistency effect for a word such as PAVE also produce one for a nonword such as MAVE. In the dual-route approach, both MAVE and NUST are pronounced using nonlexical pronunciation rules and hence should behave alike. Consistency effects for nonwords provide a particularly strong basis for deciding between the theories, insofar as they challenge the claim that the generalization must be achieved by rule, which is the primary motivation for including rules in the DRC model.

The second issue is variability in people's pronunciations of nonwords. In previous work, the performance of both dual-route and connectionist models was assessed with respect to whether they produced plausible nonword pronunciations. Judged by this criterion, both types of models do approximately equally well. Many nonwords, however, are pronounced differently by different subjects (Andrews & Scarratt, 1998; Seidenberg et al., 1994). Variability in pronunciations is a fact about performance that models of word and nonword pronunciation need to explain. One potential explanation is that people have somewhat different representations of spelling-sound knowledge because their reading experience differs (e.g., with respect to type and amount of reading). Individual differences of this sort are easily accommodated by an approach in which probabilistic spelling-sound mappings are acquired via a learning mechanism that is sensitive to statistical properties of the words to which the reader (or model) is exposed. Accounting for this variability provides an additional, more stringent criterion for evaluating models. One can consider not merely whether the model produces a plausible pronunciation (e.g., one of the pronunciations produced by people) but whether it produces the alternative pronunciations that people produce.

In principle, variability in experience could lead individuals to formulate different pronunciation rules in a DRC-type model. This proposal is difficult to evaluate because there is no current proposal about how a set of rules is acquired in the DRC framework, or how different readers could acquire different rules. An early version of the model Coltheart et al. (1993) utilized a rule-learning algorithm, but this was later discarded because it lacked psychological plausibility (e.g., in the way it searched the space of possible rules) and because the rules it generated did not work sufficiently well (Seidenberg, Petersen, MacDonald, & Plaut, 1996). Norris (1994) also used an algorithmic procedure to learn pronunciation rules, but the "rules" consisted of weights on connections among units representing letters, graphemes, and larger units, trained using a connectionist learning procedure (the delta rule). This notion of "rule" deviates considerably from Coltheart et al.'s: In the Norris model, the mappings are applied probabilistically, and they are based on sublexical statistics over multiple levels of representation that are built into the architecture of the model. Thus they are much more similar to the kinds of probabilistic constraints learned by PDP models.

Within the connectionist approach, individual differences have tended to be ignored because they require multiple runs of a model, something that until recently was too computationally time consuming to be feasible. The research described below is the first attempt to model behavioral data concerning variability in nonword pronunciation. Individual differences among readers could be due to several factors that can be simulated within the triangle framework, including constitutional factors related to learning efficiency or capacity; differences in the level of detail with which phonological information is represented; and different amounts or types of reading experience (Harm & Seidenberg, 1999). Although other factors may eventually prove to be involved, we began by determining how well differences in experience alone could account for existing data. Thus we examined whether individual differences in the pronunciations of nonwords arise from minor differences in the sample of words to which individuals are exposed. We trained a single model multiple times, using the probabilistic, frequency weighted sampling procedure used in earlier models. This procedure results in each model being exposed to a different set of words. We predicted that this seemingly minor manipulation would give rise to different pronunciations for some nonwords, as in people. Specifically, we expected the sampling differences to have little impact on nonwords containing common, high-frequency spelling-sound correspondences, for which people agree on a single pronunciation; however, they would be expected to affect the pronunciations of inconsistent nonwords such as MOUP, for which people provide different pronunciations.

We assessed performance of two models—the DRC (Coltheart et al., 2001) and multiple runs of a version of the Harm and Seidenberg (1999) model—with respect to the simulation of three studies of nonword reading: Glushko (1979); Andrews and Scarratt (1998), and Treiman et al. (2003). These studies provide different types of evidence about nonword reading, and included stimuli that were suitable for testing the models. Three kinds of data were examined: response latency from the Glushko (1979) and Andrews and Scarratt (1998) studies; effects of spelling-sound consistency on the pronunciations of ambiguous nonwords (Andrews & Scarratt, 1998; Treiman et al., 2003); and individual differences in nonword pronunciation related to spelling-sound consistency (Andrews & Scarratt, 1998).

To anticipate the results, the modeling provides two serious strikes against the dual-route approach. First, DRC does not produce the consistency effects that were observed in behavioral studies conducted with different materials in different labs. The effects are real and they cannot be explained by the alternative factors invoked by Coltheart et al. (2001). Second, DRC does not account for individual differences in the pronunciations assigned to nonwords; the pronunciation rules generate one pronunciation for each letter pattern. Although the connectionist modeling does not capture all aspects of the variability observed across individuals, it clearly establishes the feasibility of the approach.

Methods

PDP model architecture

For these simulations we used a slightly modified version of the Harm and Seidenberg (1999) model. The model had 133 orthographic units, 200 phonological units and 100 hidden units. Twenty cleanup units mediated connections from each unit of the phonological layer to itself and every other unit on the layer. The phonological representation consisted of 8 slots and 25 phonological features for each slot (see also Harm, 1998; Harm & Seidenberg, 2004). Each phoneme was coded as a binary vector, with each "on" bit representing the presence of a phonological feature. This representation differs somewhat from the Harm and Seidenberg (1999) model. However, our experience with different output phonological representations is that the precise choice of coding scheme has little effect, as long as the representation effectively encodes the similarity space of the phonetic inventory. This is true even when, as in Keidel, Zevin, Kluender, and Seidenberg (2003), the model directly addresses how such representations are learned from acoustic input.

The pronunciations on which the model was trained were based on the dialect of American English prevalent in Southern California. The most distinctive aspect of this dialect vis à vis other versions of American English is the lack of a contrast between /a/ and /ɔ/ (as in *doll* and *ball*, respectively). The behavioral data sets include a range of dialects (including Australian and Midwestern English) which creates minor discrepancies between model and human data. A nonword such as YALD, for example, is inconsistent in the Midwestern dialect spoken by the subjects in the Treiman et al. (2003) study, but not in the model's dialect. On our view, the noise that these items add to the assessment of the model was outweighed by the advantage of simulating data from a range of studies using a single architecture and training set. Clearly, it would be possible to train models in different accents and achieve better fits to particular data sets.

Model training and testing

A list of 5870 monosyllabic words was used as the training set in all simulations. During training, the probability of using any word on a given trial was proportional to the square root of its frequency (taken from Marcus, Santorini, & Marcinkiewicz, 1993), with raw frequencies were capped at 10,000. This transformation and capping ensured that low-frequency words would be selected a reasonable number of times out of the 1,000,000 training trials used for each run. For example, an item with a nominal frequency of one per million occurs 40 times on average during the training phase in the current simulations. This gives such items a chance of being acquired within a reasonable amount of training time. Similar results would obtain without this transformation, but at the cost of multiplying the amount of computational time per simulation run to an impractical degree. This is a major consideration in the present study, which involves multiple runs of the model.

All runs of the model started with the same set of initial weights set to small, random values. We kept the initial weights constant in order to be able to isolate effects of training corpus variability. On each training trial, the orthographic units were set to the values for a given target for 10 time ticks. After 12 ticks, the pattern on the phonological layer was compared to the desired output and an error signal was propagated back through the network using a variant of the continuous recurrent backpropagation algorithm. The learning rate was .01.

The model was tested by setting the orthographic values to represent a nonword for 10 ticks and observing the output on the phonological layer two ticks after the orthographic input was removed. A winner-take-all scoring system was used to determine the model's output: for each slot on the output layer, we determined which phoneme was closest to the pattern on the output at the final time tick and reported this as the model's pronunciation. Pronunciations were then scored as "regular" (according to the rules defined by Andrews & Scarratt, 1998) or "critical" (according to the definition in Treiman et al., 2003). All items from the Andrews and Scarratt (1998) study were included, although a small number included spellings or spelling-to-sound mappings not common in American English. In contrast, one condition (Case 2, onset and body) had to be eliminated from simulations of the Treiman et al. (2003) study because it depended on a contrast between the vowels /a/ and /ɔ/ not present in the model's dialect.

Because these simulations were designed to examine variability in the pronunciations generated to ambiguous nonwords, we did not designate a single pronunciation as the correct one a priori. The model's output was only scored as an error if one or more graphemes were assigned pronunciations that did not occur in any word in the lexicon. Response latencies for nonerror responses were simulated in the model using a settling time measure, which is the time (in processing cycles) required for the model to arrive at its final response. Settling times were determined with respect to the pronunciation generated for each item on each run. Settling time is not an ideal proxy for naming latency, which reflects the time to initiate a response, rather than the time required to fully specify a pronunciation. However, settling times do roughly reflect relative differences in pronunciation difficulty items and are sufficient for many purposes (e.g., simulating data averaged across items of a given type). All data presented below are means from the 10 runs of the model, except for the H statistics (defined below) for which fourteen additional runs were included to make the analyses more compatible with the human data.

Items from the experiments were also submitted to the DRC model using the standard parameter set. The model (downloaded from http://www.maccs.mq.edu.au/~max/DRC/max/DRC/) was presented with the items in "batch naming" mode. Response latencies were recorded in cycles, and the model's pronunciations were scored as above.

Results

Overall performance of the PDP model

After one million training trials, the PDP model correctly pronounced an average of 95% of the training set (5579/5870 words). Errors tended to be so-called strange words, (e.g., ACHE, CHUTE, GAUCHE, VELDT). Because these items are highly unusual both in their spelling patterns and spelling-to-sound mappings, they depend on semantic knowledge to be pronounced correctly (see Harm & Seidenberg, 2004; Plaut et al., 1996; Strain, Patterson, & Seidenberg, 1995 for a discussion of the role of semantics in reading strange words). Another difficulty is the "slot problem" (Plaut et al., 1996): the model's orthographic input consists of vowel-centered slots which can take on a different value for each letter that occurs at that location. This means that what the model knows about, e.g., the letter T in VELDT does not overlap with its representation of the same letter in more common positions (i.e., first or second position after the vowel). This also creates difficulties for some nonwords. However, the fact that the model's performance differs from people's as little as it does suggests that although the slot based representation needs to be replaced with a more realistic representation, it does not greatly interfere with learning spelling-sound correspondences.

The error rate was 5% for the Glushko nonwords, 5% for the Andrews and Scarratt nonwords and 10% for the items in the Treiman et al. (2003) study. These error rates are somewhat larger than the means reported for the subjects in the behavioral studies. Errors tended to be vowel blends. For example, the grapheme EA is most frequently pronounced /i/ or /ɛ/. Because the phonological features of vowels are encoded continuously in the model, it will occasionally settle on the midpoint between /i/ and /ɛ/, which happens to be /ɪ/, yielding, for example /bɪlm/ for BEALM. These responses were scored as incorrect, which may be a more stringent criterion than employed in scoring the human data, given expectancy effects in speech perception (e.g., Ganong, 1980), and the fact that the researcher scoring the responses is expecting one of a limited number of pronunciations.

Settling time data were right skewed in a manner similar to RT data. Specifically, a small number of observations were observed during the last two time ticks, when the output targets were typically provided. Data were trimmed by excluding trials with settling times at these last two time ticks. This resulted in discarding less than 1% of the data for both the Glushko and the Andrews and Scarratt data.

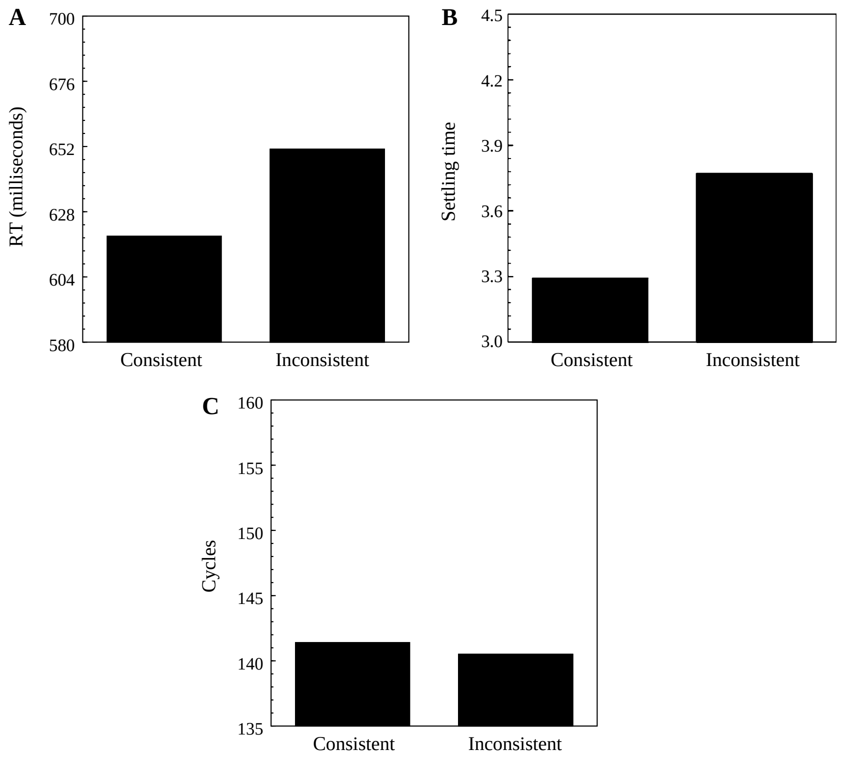

Fig. 1. Response latency data for Glushko's (1979) consistent and inconsistent nonwords: (A) subjects' response latencies in milliseconds; (B) mean settling time from PDP simulations reported in this paper; and (C) response latencies (in cycles) from the DRC.

Consistency effects on response latency

Fig. 1 presents data from the Glushko (1979) study and the current simulation. The consistency effect was highly reliable in the human data, and also in the simulation, $F(1,78) = 18.59$, $p < .001$. The item-wise correlation between settling times and naming latencies was also significant, $r = .22$, $t(84) = 2.09$, $p < .05$. The DRC model did not replicate the consistency effect, and latencies measured in DRC cycles did not correlate significantly with human latencies, $r = .15$, $t(83) = 1.40$.[2]

Coltheart et al. (2001) claimed that consistency effects for words arise from two confounding factors. One, the presence of exception words among the inconsistent stimuli, is irrelevant here because the stimuli are nonwords. The other, "whammy" effects due to the serial application of rules, could apply to nonwords. Thus if behavioral effect were due to more "whammies" in the inconsistent nonwords, DRC would produce the effect. However, it does not, indicating that the behavioral effect is not due to this factor.

Andrews and Scarratt (1998) presented a detailed study of factors that affect nonword pronunciation and data about variability across subjects, using stimuli whose neighborhood properties varied. Four types of stimuli were used (Table 1): Regular, consistent body

Table 1
Categories of Nonwords in the Andrews and Scarratt (1998) Study

Stimulus type	Example	Pronunciation	
		Regular	Analogy
RCB	TUNK	/tʌŋk/	/tʌŋk/
RIB	SULL	/sʌl/	/sʊl/
NRAU	VONTH	/vanθ/	/vʌnθ/
NRAM	WALF	/wælf/	/wæf/

Note. Abbreviations for item types described further in text: RCB, rgular, consistent body; RIB, regular, inconsistent body; NRAU, no regular analogy, unique body; NRAM, no regular analogy, many neighbors body. Regular pronunciations determined according to rules in Andrews and Scarratt (1998); analogies represent an alternative pronunciation based on the most frequent pronunciation of each body, TRUNK, FULL, MONTH and HALF, respectively.

[2] One of Glushko's items, HOVE, was removed from the analysis because it was in DRC's lexicon and had a response latency more than 3 standard deviations faster than the mean for the remaining items. Including this outlier does not improve the correlation between human latencies and DRC data.

(RCB) nonwords contain bodies that are assigned a single, regular pronunciation in all words in which they occur; regular, inconsistent body (RIB) nonwords contain bodies that assigned an irregular pronunciation in many words, and the rule-governed pronunciation in many others; no regular analogy, many neighbors (NRAM) nonwords contain bodies that occur in multiple irregular words; no regular analogy, unique (NRAU) nonwords are those whose bodies occur in only one, irregular word.

Although they were designed to test a distinction between rule and analogy mechanisms, the effect of Andrews and Scarratt's subtyping procedure was to create conditions in which consistency varied in a graded manner. The RCB items are the most consistent, because there the statistics at the rime level and the grapheme levels are consistent with each other. For example, the most frequent pronunciation in the lexicon for U is /ʌ/ and this is the only pronunciation it is assigned in the context of the body -UNK. The RIB items represent a slightly lower degree of consistency, because there is support in the lexicon for both a regular pronunciation (e.g., HULL and GULL for -ULL) and an irregular pronunciation (e.g., FULL and PULL). In the case of the NRAM and NRAU items, statistics at the body level overwhelmingly favor a different pronunciation from the statistics at the grapheme level: By definition, there are no instances in which the most frequent pronunciation for a critical grapheme is assigned in the context of the body in question. Here we assess the models with respect to naming latencies in this study; later we consider data about the types of pronunciations produced.

The latency data shown in Fig. 2 reflect an orderly, graded influence of consistency on response latency. The most consistent items (RCB) were read most quickly. Items that were regular but inconsistent at the word-body level (RIB) were read more slowly. Items with no regular analogy (NRAU and NRAM) are interesting because they exhibit conflicts between statistics at the grapheme and word-body levels. For these items, the number of neighbors had a large effect: items with many neighbors were named significantly more quickly than items with only one. Average settling times for the PDP models exhibited this same difference between the NRAU and NRAM items, $F(1,126) = 14.34$, $p < .001$. In the human data, the advantage for RCB over RIB items was marginally significant; the model data also produce this trend, $F(1,78) = 2.01$. Finally, the difference between NRAM and NRAU items was also significant, $F(1,46) = 5.62$, $p < .05$. The item-wise correlation between the model and human data was significant, $r = .28$, $t(126) = 3.22$, $p < .005$.

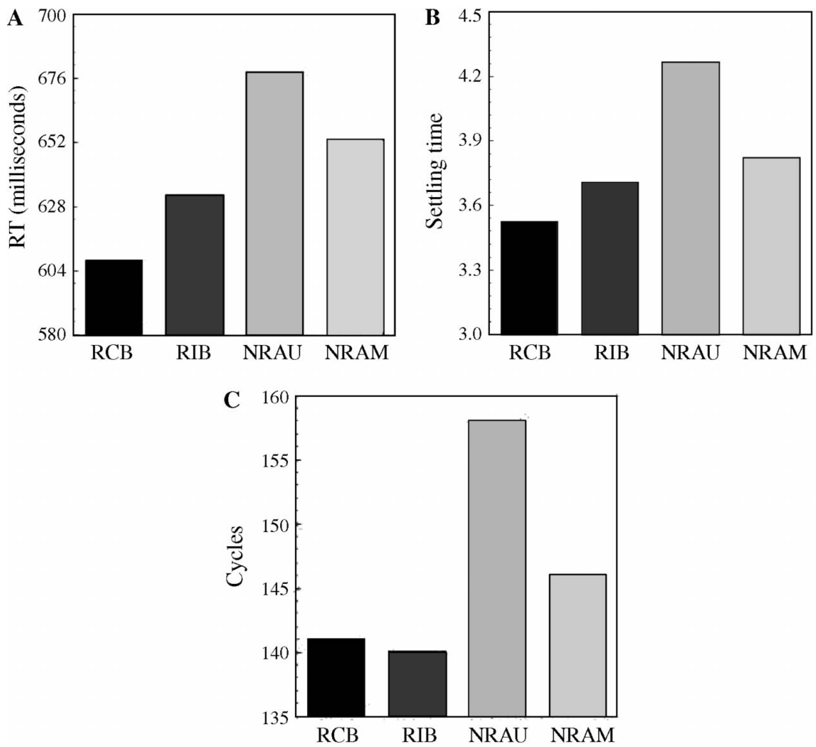

Fig. 2. Response latency data for Andrews and Scarratt's (1998) items: (A) subjects' response latencies in milliseconds; (B) settling time from PDP simulations reported in this paper; and (C) response latency data from the DRC in cycles.

The DRC model captures the overall advantage of items with predominately regular neighbors over items with no regular neighbors, $F(1, 126) = 11.85$, $p < .001$, as well as the difference between NRAM and NRAU items, $F(1, 46) = 7.30$, $p < .01$. Item-wise correlation between the DRC and human data was significant, $r = .36$, $t(122) = 4.27$ $p < .001$.[3] The only major discrepancy between the DRC model and the human data is in regard to the difference between the RCB and RIB conditions, which is marginally significant in the human data but nonsignificant and numerically reversed in the DRC model.

To summarize, both models provide a good account of the latency data from Andrews and Scarratt (1998) study. In both the human and PDP data, there is a trend toward a consistency effect for nonwords with "regular" word bodies, which DRC does not produce. Taken with the results for the Glushko (1979) study, these findings suggest that the PDP model is more sensitive to consistency effects on nonword latencies.

Consistency effects on pronunciation "regularity"

We now consider the effects of consistency on how nonwords are pronounced. An item such as SULL, which has both regular (HULL and DULL) and irregular (PULL and FULL) neighbors, can be pronounced to rhyme with either of these alternatives. Two studies have examined the degree to which skilled readers are sensitive to lexical statistics at this level when assigning pronunciations to nonwords.

Andrews and Scarratt (1998)

Fig. 3A shows the percentage of regular pronunciations produced by subjects in the Andrews and Scarratt (1998) study. Regularity was defined in terms of Andrews and Scarratt's rules. As in the latency data, the effects are graded. The RCB and RIB conditions produced the highest percentages of regular pronunciations, with NRAU producing a much lower percentage and the NRAM condition the fewest. Statistically, the difference between the RCB and RIB conditions was marginal, whereas the RIB-NRAU and NRAU-NRAM differences were significant. The PDP model (Fig. 3B) produced the same ordering of conditions, and the same pattern of significant differences between conditions ($t(64) = 7.95$, $p < .01$ for RIB-NRAU and $t(23) = 7.74$, $p < .01$ NRAU-NRAM) except that the RCB-RIB difference was larger than for humans, $t(78) = 7.82$, $p < .01$.

The results from the DRC model are different. First, the model is at ceiling (100% regular pronunciations) in the RCB and RIB conditions, which was not observed in human subjects or the PDP model. Second, DRC does not reproduce the significant difference between the NRAU and NRAM conditions that is present in both human data and PDP model. Numerically, the difference is in the wrong direction, although it is not statistically reliable. DRC produces a significant main effect of regular body vs. no regular analogy body items (RCB-RIB vs. NRAU–NRAM) rather than the graded effects in the human and PDP data. These results implicate the same problem as the latency simulations: DRC's rule set does not capture people's sensitivity to degrees of spelling-sound consistency.

In Table 2, stimuli are grouped by the proportion of regular responses assigned by subjects in Andrews and Scarratt (1998). The data from the PDP simulation show a pattern similar to the human data, although the model overestimates the proportion of regular pronunciations throughout, particularly in the 20–40 bin. The DRC only produces a large proportion of irregular responses in one bin. It is important to note here that the "irregular" pronunciations produced by the DRC are the result of a discrepancy between how the rules are encoded, and not the result of the involvement of analogical or lexical processing. The rule set adopted by Andrews and Scarratt (1998) represents a minimalist approach, wherein only very small units (graphemes) are coded. On their scheme (and others in a similar vein, see e.g., Venezky, 1970), each grapheme has a rule associated with it, and pronunciations are considered regular if all of the rules apply. By not including context-sensitive rules and multigrapheme rules, this scheme avoids a number of issues that arise for more complex rule sets. For example, if both single- and multigrapheme rules are allowed, a mechanism to adjudicate between them is required. However, restricting the rule set to rules encoding single graphemes necessarily ignores meaningful regularities at larger grain sizes. This is clear from the small number of regular pronunciations produced by human subjects in the NRAU and NRAM conditions.

The rule set adopted in the DRC approach does not have the same restrictions as the Andrews and Scarratt (1998) rules. It contains both context-sensitive and multigrapheme rules and explicitly defines a mechanism by which rules operating at larger unit sizes (e.g., the word body) can override rules at smaller unit sizes. As shown in Fig. 3 and Table 2, these rule sets make slightly different claims about what should count as a regular pronunciation. In particular, the word-body level rules in the DRC cause it to produce irregular pronunciations for 16 of the 24 items that most often elicited irregular pronunciations from human subjects—for all other items, the DRC agrees strongly with the Andrews and Scarratt (1998) rules.

[3] Three items, LANG, RATCH and TOPE were present in the DRC's vocabulary and were removed from the analyses because their RTs were nearly 4 standard deviations faster than the remaining items.

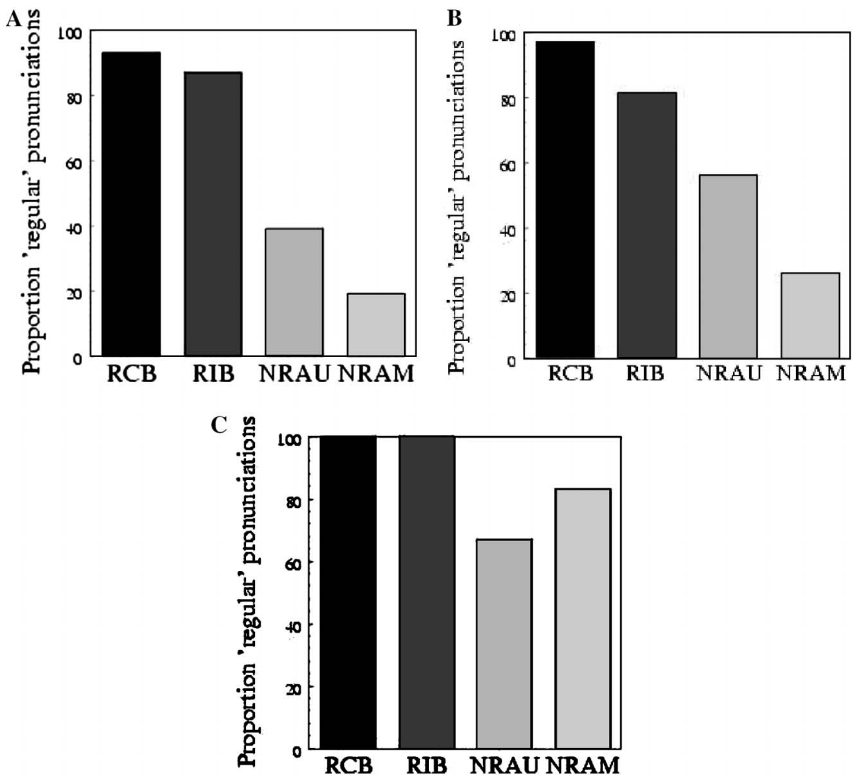

Fig. 3. Percentage of regular pronunciations for Andrews and Scarratt's (1998) items: (A) human data; (B) data from PDP simulations reported in this paper; and (C) data from the DRC.

Table 2
Percentage of regular pronunciations generated by humans and the two models

BIN	0–20	20–40	40–60	60–80	80–100
N	24	15	8	8	73
Percent regular					
AS98	6.74	31.89	46.28	69.93	96.53
PDP	24.13	55.38	50.31	86.25	90.38
DRC	33.33	100.00	100.00	87.50	95.89

Note. Bins represent ranges of percent regular pronunciations and are organized according to human data from the original Andrews and Scarratt (1998) study; *N*, number of items in each bin; AS98, Andrews and Scarratt's (1998) subjects; PDP, current parallel distributed processing model; DRC, Dual Route Cascade model; percent regular, mean percentage of regular pronunciations for nonwords in each bin.

Treiman et al. (2003)

Treiman et al. (2003) examined the pronunciation of specific vowel graphemes in specific onset or coda contexts. Their items were developed on the basis of statistical analyses of a large corpus of English monosyllables (Kessler & Treiman, 2001). Examples of the items are shown in Table 3. They obtained naming data for these items in order to examine the effects of different contexts on the pronunciations of vowels. For each vowel, a "critical pronunciation" was chosen. Critical pronunciations were defined as involving relatively infrequent grapheme-to-phoneme correspondences that are highly conditioned by context. One dependent measure in their study was the difference in the proportion of critical pronunciations for experimental items (which contain onsets or codas that occur in words with the critical pronunciation) and control items (which contain "neutral" onsets or codas). For example, the probability of pronouncing the A in nonwords with the body -ANGE as /eɪ/ was compared to the probability of producing the same vowel in items with the body -ANCE. Treiman et al. (2003) found that connectionist models (Plaut et al., 1996; Harm & Seidenberg, 2004) captured the broad pattern of the data in showing a strong influence of the critical context on vowel pronunciation; however the fit at the smaller grain size of predicting particular pronunciations for particular items was "not impressive" (p. 67).

Data from 10 runs of the current model are given in Table 3 along with the human data and data from two other models that Treiman et al. (2003) used for comparison. Multiple runs of the present model provide a closer approximation of the human data than the results

Table 3
Difference in proportion of critical vowel pronunciations in experimental and control nonwords from the Treiman et al. (2003) study

Context	CV1	VC1	VC3	VC4	VC5	VC6
Critical pronunciation	/a/	/eɪ/	/ɛ/	/aɪ/	/oʊ/	/ʊ/
Example	Squant	Crange	Chead	Crind	Prold	Blook
Human data	0.58	0.55	0.12	0.33	0.83	0.70
Current model	0.39	0.59	0.32	0.83	0.77	0.68
Harm and Seidenberg (2004)	0.56	0.90	0.40	1.00	1.00	0.80
DRC	0.00	0.00	0.00	0.00	0.00	0.00

Note. C, consonant; V, vowel; and DRC, Dual Route Cascade model.

reported by Treiman et al. (2003) in their Experiment 1, which were based on single runs of earlier models. For example, the single run of the Harm and Seidenberg (2004) model tested by Treiman et al. (2003) generated 100% critical responses for the Case 4 and Case 5 items. Among runs of the model tested in the current work, different runs produced different pronunciations for some of these items, leading to lower (and thus more human-like) proportions of critical pronunciations. The only cell in which a large deviation from the human data was observed is the VC4 condition. Interestingly, the word body in this condition was -IND, and the only word in the lexicon for which this body is assigned the regular pronunciation is for one sense of the homograph WIND. Homographs were not included in the training set (their pronunciations are normally disambiguated by context, which the current model lacks). Thus the lexical item that would contribute most to producing the "critical pronunciation" was not included, and so the model "regularized" more than people. Otherwise the simulation and behavioral data are very similar.

Individual differences

Unlike classical box-and-arrow models which are most often treated as deterministically generating predictions (Coltheart, 1999), PDP models implement a set of principles regarding the acquisition and use of linguistic knowledge that are fundamentally probabilistic (Seidenberg, 1997). In a quasiregular system such as English spelling-to-sound, this means that different runs of the same model can arrive at slightly different encodings of the same mapping, making it quite natural to explain individual variability in the human data.

For example, data from the Andrews and Scarratt (1998) items for 10 runs of the model are depicted in Fig. 4. While in all instances the ordinal pattern was the same, the relative proportion of regular pronunciations differed considerably across runs of the model, particularly for the items with "no regular analogy" (NRAU/NRAM). This pattern could be interpreted in terms of differences in "reading style" among runs of the model. A number of studies have attempted to iden-

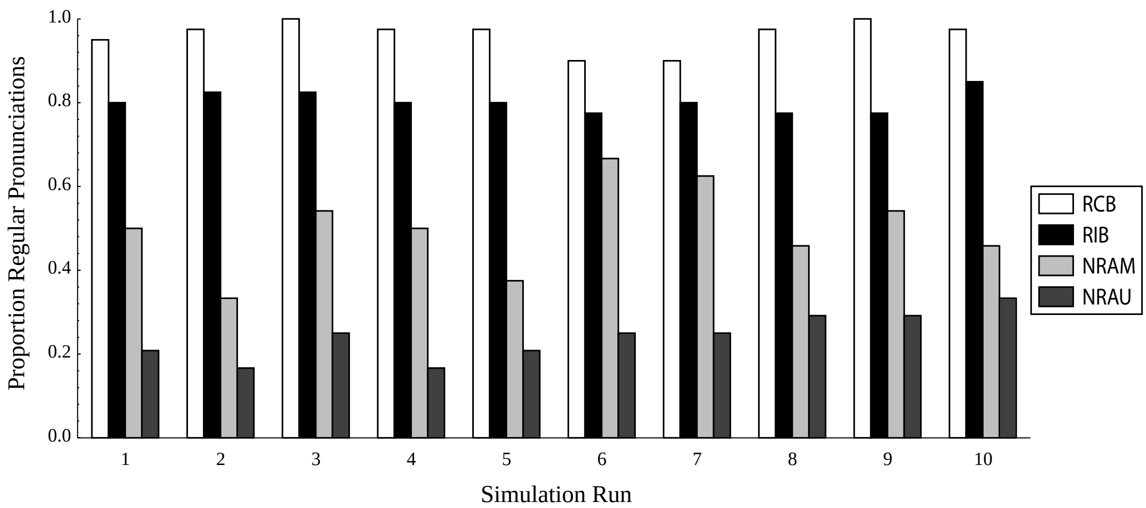

Fig. 4. Percentage of regular pronunciations for Andrews and Scarratt's (1998) items from all 10 runs of the PDP model reported in this paper. RCB (white), regular, consistent body; RIB (black), regular, inconsistent body; NRAM (light gray), no regular analogy, many neighbors; NRAU (dark gray), no regular analogy, unique body.

tify subgroups of skilled (Baron & Strawson, 1976; Brown, Lupker, & Colombo, 1994) and developing (Goswami & East, 2000) readers who vary in the degree to which they appear to apply grapheme-to-phoneme rules or lexical analogies in their reading of nonwords. The current results suggest that at least some of this variability may be the result of fairly subtle differences in experience, and may be explicable in terms of the encoding of spelling-to-sound statistics rather than different reading strategies.

In addition to quantifying the proportion of regular responses given for each type of item, Andrews and Scarratt (1998) also quantified variability among subjects in the pronunciation assigned to particular nonwords. Variability was quantified using the information theoretic H statistic, a measure of entropy which, in this instance, quantifies the heterogeneity of responses to a given nonword. It is computed using the formula

$$\sum[-p_i \times \log_2(p_i)], \quad (1)$$

where p_i is the probability of a given pronunciation. A value near 0 represents a nonword for which a single pronunciation is highly dominant, whereas higher values represent a nonword for which many pronunciations are equiprobable. Because nonwords generally have only two or three possible pronunciations, the maximal value for H is rarely approached. Furthermore, because values for H are related to the number of observations involved, we ran an additional 14 models in order to match the number of subjects (24) in the Andrews and Scarratt study. As shown in Fig. 5, the greatest consensus (thus, lowest H values) was observed in the RCB condition with incrementally less in the RIB condition. In the human data, there was a large increase in H values between the RIB and NRAM that is much larger than that observed in the model.

Similar to the human data, variability was greater for the NRA items than the items with regular analogies in the model $F(1,126) = 19.82$, $p < .001$. Also like the human data, the difference between RIB items and RCB items was not significant, $F(1,78) = 1.34$. Unlike the human data, however, the difference between NRAU items and NRAM items was significant, $F(1,46) = 6.73$, $p < .05$. Overall, the pattern of results from the human subjects and multiple runs of the model are fairly similar, although the model somewhat underestimates the variability in the human data for all categories, and particularly for the NRAM condition.

The lower level of overall variability in the model may be the result of the small range in which frequency was actually manipulated, or the fact that all runs of the model started with the same set of weights. A more serious problem for the model is the large difference in variability between the NRAU and NRAM items, also clearly depicted in Fig. 4. The proportion of regular pronunciations for the NRAM items is tightly clustered around the (very low) mean. This suggests that when there is abundant evidence for a particular body-level mapping between spelling and sound, this tends to override grapheme-level mappings more regularly in the models than in human subjects. There is little evidence that the models *generally* prefer body-level statistics to grapheme-level statistics: The model is as likely to overestimate the proportion of grapheme-level pronunciations as to underestimate it in a given data set (Table 3, cases CV1 and VC1; Fig. 3, NRAU items). However, the NRAM items present a special case in which the body-level statistics are rich enough to support good generalization (because the bodies all occur in multiple words) and consistent enough to be impervious to small differences in the frequencies of particular words (because they are all 100% consistent at the body-level).

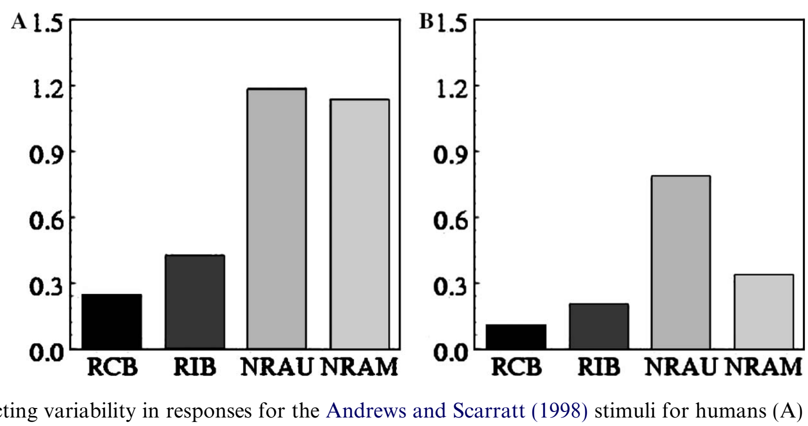

Fig. 5. H values reflecting variability in responses for the Andrews and Scarratt (1998) stimuli for humans (A) and the current model (B). No data are presented for the DRC because it does not produce variable pronunciations.

Discussion

The behavioral studies and modeling discussed here provide evidence bearing on theories of how the systematicity in spelling-sound correspondences is encoded by skilled readers. Three aspects of the results are more consistent with the view that this knowledge consists of probabilistic constraints rather than categorical rules. People and PDP models both show a graded sensitivity to consistency in chronometric measures of nonword reading. Consistency also influences the pronunciations generated by both people and models, suggesting that statistical properties of the lexicon can be used productively. Finally, individual variability among readers and runs of a PDP model in the pronunciation of particular nonwords reflects consistency as well: highly consistent nonwords generate a high degree of agreement among readers, whereas less consistent items generate a greater variety of responses. These phenomena are less successfully simulated by an implemented model (DRC) that relies on categorical rules to translate from spelling to sound, suggesting a role for lexical statistics in the translation of spelling to sound.

Consistency effects on response latency

That consistency influences both word and nonword reading latency is clear from experiments over a many-year period. Studies of word reading that independently manipulated consistency (defined at the word-body level) and regularity (mainly at the grapheme to phoneme level) have yielded large effects of consistency and small or nonexistent effects of regularity (Cortese & Simpson, 2000; Jared, 2002). Nonwords yield similar effects. These results follow naturally from a view in which spelling-sound correspondences are statistical rather than categorical. It is thus not surprising that computational models that incorporate statistical learning can capture these kinds of effects.

The fact that DRC does not account for performance on the Glushko (1979) nonwords is particularly important because Coltheart et al. (2001) cite this result as a motivation for adopting cascaded as opposed to thresholded processing: in a cascaded model, there is an opportunity for partially completed output from the lexical route to influence nonword pronunciation, providing a potential basis for the advantage of consistent nonwords over inconsistent ones in naming latency. Although used to motivate this property of the DRC model, it does not produce the Glushko effect. In principle, the effect could arise in the DRC model from the activation of exception words in the lexical network, producing a conflict between the two routes. For a nonword to activate words to a sufficiently high level, the parameters governing inhibition must be set to low values. Doing so creates a problem, however: the model generates lexicalization errors, particularly for simple wordlike nonwords such as STARN. Thus it is difficult to tune the model's parameter set so that it correctly simulates this aspect of nonword naming.

What is the correct pronunciation of CHEAD or MOUP?

Early discussions of nonword reading in computational models focused on pronunciation accuracy: a model's output was scored as correct if it generated a plausible pronunciation (e.g., one that rhymed with a similarly spelled word). Besner et al. (1990) noted that the Seidenberg and McClelland (1989) model frequently produced pronunciations that differed from people's, suggesting that pronunciation rules might be required. Plaut et al. (1996) traced this behavior to limitations in the way phonological information was represented in such models, rather than the need for rules, and reported human-level accuracy, as did Harm and Seidenberg (1999, 2004).

Later discussion focused on nonword pronunciation at a finer level of detail. Seidenberg et al. (1996) obtained behavioral data concerning the pronunciations of several hundred nonwords. They found that many nonwords are pronounced in more than one way. The two most common pronunciations accounted for more than 90% of subjects' responses. Rather than assessing how often a computational model produced a single, intuitively plausible nonword pronunciation, Seidenberg et al., examined whether the computed pronunciations matched either of the most common ones produced by subjects. In that large-scale study, a connectionist model with an improved phonological representation provided a slightly better fit to the data than the Coltheart et al. (1993) version of the dual-route model. This was important both because of claims that rules were necessary for nonword reading and because the rules in the Coltheart et al., model were specifically created to account for nonword pronunciation.

The current simulations extend these findings to two additional data sets. These studies show that pronunciation of novel forms is influenced by statistical properties of the spelling-sound mapping that arise from similarity relations among words. A central claim of the DRC framework is that nonwords are pronounced by applying rules. The rules state how spellings deterministically map to pronunciations. The defining characteristic of the rules (and the mechanism by which they are applied) is that they are not influenced by statistical properties such as how often particular mappings occur across words. Thus, evidence that such statistical properties affect word and nonword performance counts against the dual-route framework. Statistics at the level of individual graphemes and larger units such as the word body, and even ad hoc units such as the *oncleus* (onset + nucleus) all play a role in determining the pronunciations of novel items.

Modeling individual variability

Extending modeling to account for variability across individuals with respect to nonword pronunciations is a natural direction for research in this area to follow. For the DRC model there are two challenges. One is to develop an account of how grapheme-phoneme conversion rules are learned. At present the model lacks this learning component. The second challenge would then be to account for how different sets of GPCs could be learned, as demanded by the behavioral data.

Previous assessments of connectionist models' performance also ignored individual variability in nonword naming performance: they were conducted by comparing mean latencies or modal pronunciations from experiments in which many subjects were run to the data from a single run of a model. This introduces a mismatch insofar as the model data is actually more comparable to the results for an individual subject rather than group results (see Seidenberg & Plaut, 1998, for discussion). The present research is a step toward more serious evaluations of individual differences in both human and model performance. The data from multiple runs provides a better fit to the human data, especially when the dependent measure is a qualitative one such as the proportion of "regular" or "critical" responses generated. The model also captures some of the more specific data concerning the variability of responses to different types of nonwords observed in the Andrews and Scarratt study. This suggests a number of possible future directions for research.

In the current study, we introduced variability by using slightly different randomizations of the same word frequency list for each run of the model. This hardly captures the much greater variability in the kind and amount of reading people do—even within the relatively homogenous population of university students who participate in psychology experiments. In addition, current computational models (both connectionist and DRC) are limited to monosyllabic words, which may introduce error in their performance. Some of the statistics that are relevant to the pronunciation of monosyllabic words and nonwords arise from exposure to a much larger vocabulary that includes multisyllabic words. This creates obvious discrepancies between the model's experience and the human reader. Furthermore, variability in the training regime is not the only possible source of variability in connectionist models. There are also effects due to differences in the random assignment of initial weights whose effects need to be explored further.

Finally, the method of examining multiple runs of the same model can be used to explore whether the patterns of impairment seen in cases of acquired dyslexia are related to premorbid individual differences in reading. Such patients sometimes exhibit extreme patterns of dissociation (e.g., in reading words vs. nonwords); such patterns are often taken as a basis for identifying isolable processing systems (e.g., routes). However, little information is available about the patients' premorbid reading abilities, how often they read and what types of materials. A given type of brain injury may have different behavioral effects as a function of premorbid individual differences. This possibility is the complement of the situation studied by Plaut (1996), who demonstrated that random damage to a single model can produce highly variable patterns of impairment. Thus, behavioral impariments that are observed in acquired dylsexia depend on individual differences with respect to both premorbid capacities and effects of neuropathology. These factors are likely to produce a wide range of behavioral profiles.

Conclusions

It is a positive reflection of the degree of sophistication of contemporary models of word and nonword reading that strong tests of their basic assumptions can be derived. Several researchers have been able to identify stimuli that contrast the effects of regularity vs. consistency in word and nonword reading (e.g., Andrews & Scarratt, 1998; Cortese & Simpson, 2000; Jared, 1997 and Treiman et al., 2003). Simulations of these behavioral studies provide further data bearing on the adequacy of these competing theories. The speed with which nonwords are pronounced, the pronunciations assigned to them, and the degree of agreement among individuals regarding their pronunciation all depend on the statistical properties of their lexical neighborhoods. Such phenomena are more easily captured by connectionist models which acquire this knowledge via statistical learning mechanisms. Both behavioral and modeling results support the view that generalization results from using a network that encodes this statistical knowledge, rather than application of rules.

Acknowledgments

This work supported by NIMH Grants P50 MH 64445, K02 MH 01188, and NICHD grant R01 MH 29891 (M.S.S.) and NIH fellowship F32 DC 06352 (J.D.Z.). The model used in this paper is based on work by Mike Harm, whom we also thank for the simulation software. We also thank Max Coltheart for making an executable version of the DRC available online.

References

Andrews, S., & Scarratt, D. R. (1998). Rule and analogy mechanisms in reading nonwords: Hough dou peapel rede gnew wirds. *Journal of Experimental Psychology: Human Perception & Performance, 24*, 1052–1088.

Baron, J., & Strawson, C. (1976). Use of orthographic and word-specific knowledge in reading words aloud. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 386–393.

Besner, D., Twilley, L., McCann, R., & Seergobin, K. (1990). On the connection between connectionism and data: Are a few words necessary. *Psychological Review, 97*, 432–446.

Brown, P., Lupker, S. J., & Colombo, L. (1994). Interacting sources of information in word naming: A study of individual differences. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 537–554.

Coltheart, M. (1999). Modularity and cognition. *Trends in Cognitive Sciences, 3*, 115–120.

Coltheart, M. (2000). Dual routes from print to speech and dual routes from pring to meaning: Some theoretical issues. In A. Kennedy, R. Radach, J. Pynte, & D. Heller (Eds.), *Reading as a perceptual process* (pp. 475–492). Oxford: Elsevier.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review, 100*, 589–608.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review, 108*, 204–256.

Cortese, M. J., & Simpson, G. B. (2000). Consistency effects: What are they? *Memory & Cognition, 28*, 1269–1276.

Ganong, W. F. (1980). Phonetic categorization in auditory word perception. *Journal of Experimental Psychology: Human Perception and Performance, 6*, 110–125.

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5*, 674–691.

Goswami, U., & East, M. (2000). Rhyme and analogy in beginning reading: Conceptual and methodological issues. *Applied Psycholinguistics, 21*, 63–93.

Harm, M. W. (1998). Division of labor in a computational model of visual word recognition. Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading, and dyslexia: Insights from connectionist models. *Psychological Review, 163*, 491–528.

Harm, M. W., & & Seidenberg, M. S. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review, 111*, 662–720.

Jared, D. (1997). Spelling-sound consistency affects the naming of high-frequency words. *Journal of Memory and Language, 36*, 505–529.

Jared, D. (2002). Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language, 46*, 723–750.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language, 29*, 687–715.

Keidel, J. L., Zevin, J. D., Kluender, K. R., & Seidenberg, M. S. (2003). Modeling cross-linguistic speech perception. In *Proceedings of 15th annual International Conference of Phonetic Sciences*. Barcelona, Spain.

Kessler, B., & Treiman, R. (2001). Relationships between sounds and letters in English monosyllables. *Journal of Memory and Language, 24*, 592–617.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition, 83*, PB35–PB42.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313–330.

Norris, D. (1994). A quantitative multiple-levels model of reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1212–1232.

Pinker, S. (1999). *Words and rules: The ingredients of language*. New York: Basic Books.

Plaut, D. (1996). Relearning after damage in connectionist networks: Toward a theory of rehabilitation. *Brain and Language, 52*, 25–82.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.

Posner, M. I., Abdullaev, Y. G., McCandliss, B. D., & Sereno, S. (1999). Neuroanatomy, circuitry and plasticity of word reading. *Neuroreport, 10*, R12–R23.

Rayner, K., Foorman, B., Perfetti, C., Pesetsky, D., & Seidenberg, M. S. (2001). How psychological science informs the teaching of reading. In *Psychological science in the public interest monographs* Vol. 2 (pp. 31–74). American Psychological Society.

Saffran, J., Newport, E., Aslin, R., & Tunick, R. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science, 8*, 101–105.

Seidenberg, M. S. (1995). Visual word recognition: An overview. In P. Eimas & J. L. Miller (Eds.), *Handbook of perception and cognition: Language*. New York: Academic Press.

Seidenberg, M. S. (1997). Language acquisition and use: Learning and applying probabilistic constraints. *Science, 275*, 1599–1603.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review, 96*, 523–568.

Seidenberg, M. S., & Plaut, D. C. (1998). Evaluating word reading models at the item level: Matching the grain of theory and data. *Psychological Science, 9*, 234–237.

Seidenberg, M. S., Petersen, A., MacDonald, M. C., & Plaut, D. C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning Memory and Cognition, 22*, 48–62.

Seidenberg, M. S., Plaut, D. C., Petersen, A. S., McClelland, J. L., & McRae, K. (1994). Nonword pronunciation and models of word recognition. *Journal of Experimental Psychology: Human Perception and Performance, 20*, 1177–1196.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory and Cognition, 21*, 1140–1154.

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition, 88*, 49–78.

Treiman, R., Mullennix, J., Bijelac-Babic, R., & Richmond-Welty, E. D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General, 124*, 107–136.

Venezky, R. L. (1970). *The structure of English orthography*. Mouton: The Hague.