In Proceeedings of the 29th Annual Conference of the Cognitive Science Society (2007), pages 59-64. Mahwah, NJ: Erlbaum

A Behavioral and Computational Integration of Phonological, Short-Term Memory, and Vocabulary Acquisition Processes in Nonword Repetition

Brandon Abbs (brandon-abbs@uiowa.edu)
Prahlad Gupta (prahlad-gupta@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

J. Bruce Tomblin (j-tomblin@uiowa.edu)
Department of Speech Pathology & Audiology, Wendell Johnson Speech & Hearing Center
Iowa City, IA 52242 USA

John Lipinski (john-lipinski@uiowa.edu)
Department of Psychology, E11 Seashore Hall
Iowa City, IA 52242 USA

Abstract

Theories of language acquisition propose an early and important role for perceptual processes, but often present these processes as separate from the short-term sequencing processes and long-term linguistic knowledge that also contribute to the development of language. Using nonword repetition as a measure that reflects the interaction of these processes, we present a computational model that integrates all three of these processes into a single system and examine the similarity between the relationship of these three processes found in the model and the relationship found in adolescent humans.

Keywords: language acquisition; short-term memory; phoneme discrimination; vocabulary; nonword repetition

Processes of Vocabulary Acquisition

While there is an acknowledged role for perceptual processes in language acquisition by models of language acquisition, the precise relationship between these processes and higher-level lexical processes tends to be left unspecified. Specifically, it has been argued that phonological learning involves the organization of sound whereas lexical learning involves the organization of meaning and this shared organizational principle could theoretically be supported by stimulus-general processes (Jusczyk, 2000, p. 1-3). Other theories support the idea that accomplishing the former helps to accomplish the later.

For example, the Competition Model (Bates & MacWhinney, 1987) claims that perceptual units must be isolated from irrelevant and variable contextual factors. Once these perceptual units are isolated and can be reliably recognized, then the mapping between these perceptual units and their function in the language can begin. This process is similar to the discovery of variability and stability in mapping word forms to lexical function, but statistical relationships between these processes are rarely discussed and the bulk of the empirical work related to this model has focused on the ability to map between lexical and syntactic knowledge.

One potential benefit to lexical growth suggested by these perspectives is through a benefit to parsing abilities as a result of enhanced stimulus acuity. The rapid rate of speech transmission and the context-dependent nature of speech perception makes one's perceptual ability in the early stages of language acquisition, before lexical cues can aid processing, an important part in parsing and analyzing the audio-perceptual stream (Jusczyk, 2000, p. 204; Bates & MacWhinney, 1987). Having precise representations of the auditory environment may contribute to the development of an efficient recognition system and aid in the formation of reliable mappings between sound and meaning.

However, acuity may also hinder processing as the lexicon grows by preventing generalization based on the similarity between new words and known words during the process of vocabulary acquisition. Developmental data, such as that presented by Stager and Werker (1997) show changes in phonetic sensitivity in a word-learning task as a function of age (and thus, they argue, experience with learning new words). Certain phonological contrasts were not maintained in word-learning tasks and may become less important as the lexicon grows and known words aid the processing of unknown words. Thus, the system may be jointly influenced by perceptual and memory processes and must ultimately balance forces of acuity and generalization.

The statistical learning literature (e.g., Saffran, Aslin, & Newport, 1996) provides additional information on the function of memory processes by focusing on the acquisition of knowledge about the stimulus environment. Here the focus is not on acuity, but on the ability to learn the statistical structure of an environment. This knowledge contributes to vocabulary acquisition by helping to parse the stream into the appropriate units for learning about a particular aspect of language. Thus, it is the function of a domain-general learning mechanism that contributes to vocabulary growth and the impact of individual differences in phonological processing is felt somewhat indirectly through the effects on memory.

Theoretically, it is clear that both perceptual and memory processes contribute to vocabulary acquisition, but empirically the nature of this relationship remains unclear. Perceptual and memory processes appear to be studied independently though they make joint contributions to vocabulary acquisition and may also interact with one another. A good starting point for bringing together these investigations is to ask how changes in the representation of auditory information, which would be determined largely by the function of perceptual processes, might impact the memory processes responsible for vocabulary acquisition in humans and in a model of vocabulary acquisition.

A recent connectionist model (Gupta & Tisdale, submitted) provides a framework for examining the potential impact of phonological representations on vocabulary acquisition processes. Here it has been demonstrated that short-term memory for phonological information (also known as PSTM) can be conceptualized as a causal factor in determining one's vocabulary and that both PSTM and vocabulary causally determine the generalization and sequencing ability demonstrated by participants in a nonword repetition (NWR) task, a task that has long been thought to measure the functioning of the memory processes responsible for vocabulary acquisition (Gupta & Tisdale, submitted).

Further, this model demonstrates these relationships within a single processing system: a recurrent network. Within this system, PSTM is operationalized as the recurrence in the model and vocabulary knowledge lies in the weights between layers of the network. The goal of the current work is to incorporate the concept of perceptual processing into this model based on the theoretical relationship presented above and the empirical observations presented below.

This new model characterizes NWR as a product of both vocabulary and PSTM. NWR measures how the processing of novel linguistic information is impacted by PSTM processes and vocabulary. Thus, we see NWR as a measure that captures the interaction between the processes responsible for vocabulary. For this reason, we focus on NWR as our dependent measure instead of a measure of vocabulary acquisition. The question we attempt to answer is whether or not perceptual processing (measured by phoneme discrimination) predicts the efficacy of these processes based on their shared reliance on both acuity and generalization.

Behavioral Integration

In order to have a better understanding of the relationship between perceptual processes, PSTM, and vocabulary, we performed a multiple regression analysis on a set of previously collected data. These data are part of a larger longitudinal study of children with Specific Language Impairment (SLI; a disorder of language in children with no neurological cause) and their developmental cohorts (Tomblin, Zhang, Buckwalter, & O'Brien, 2003). As such, measures were taken at different time points, but were completed by the same experimenters in the same testing locations.

Data from 57 adolescents were included in the analysis based on the availability of data on the tasks of interest. These participants were between 15 and 17 years old at the latest testing session. No attention was paid to the participants language status as the interest in the group lies in the inclusion of participants with a wide range of language abilities, rather than the inclusion of participants from a specific population.

Perceptual Processing

Perceptual processing was assessed during the latest testing session using a phoneme discrimination task that had participants discriminate between tokens of English words whose initial consonant had a varying voice onset time (VOT) resulting in a b/p continuum. On the "b" end of this continuum, was a /bol/ token ("bowl") recorded by a human speaker with a VOT of 0 ms. This initial token was spliced to create a range of additional stimuli with VOTs of 10 – 50 ms in 10 ms increments. Stimuli were presented through headphones with experimental parameters that limited the memory components of the task. The participants' task was to discriminate between tokens; reporting whether the tokens were the "same" or "different" (stimulus creation and experimental methods can be found in Coady, Kluender, & Evans, 2005).

The relevant measure for this task is proportion correct on "different" trials involving two within-category tokens (i.e., VOT 0 vs. VOT 20 and VOT 30 vs. VOT 50) and two between-category tokens (i.e., VOT 10 vs. VOT 30 and VOT 20 vs. VOT 40). Limitations of the data set prevented the calculation of a useful *d'* as will be done in the modeling data.

Short-Term Memory Processing

For assessing STM ability using novel linguistic stimuli, a NWR task was given. For assessing ability using known linguistic stimuli, a digit span (DS) task was given. These tasks were completed in the same testing session, one year prior to the discrimination task.

Nonword Repetition In the NWR task, participants were presented with blocks of 20 nonwords from an established corpus of novel linguistic stimuli (Gupta et al., 2004). Within each block, participants were presented stimuli of a single syllable length, beginning with 2-syllable nonwords and ending with 7-syllable nonwords. Each syllable had a consonant-vowel (CV) structure, except for the final syllable, which had a CVC structure. Stimuli were presented through headphones and the participants' task was to simply repeat each word immediately upon hearing it. The measure for this task was percentage of whole words correctly repeated at each syllable length.

Digit Span In the DS task, participants were presented with lists of digits in blocks of 8 lists. Each list in a block was presented through headphones at a rate of 1 digit per second. Lists were initially 2 items in length and grew by 1 item with each subsequent block. Participants continued to a new block if they accurately recalled more than 50% of the lists in a given block. The measure for this task is the longest list length at which the 50% performance criterion was exceeded.

Vocabulary

Participants completed the expressive subtest of the Comprehensive Receptive and Expressive Vocabulary Test (CREVT; Wallace & Hammill, 1994) as a measure of expressive vocabulary. Participants completed the Peabody Picture Vocabulary Test-III (PPVT-III; Dunn & Dunn, 1997) as a measure of receptive vocabulary. Scores on these two components were then used to create a composite z-score for vocabulary. These assessments were completed during the same session as NWR and DS.

Results

The independent predictors of the regression model consisted of the two measures of within category discrimination, the two measures of between category discrimination, the DS measure, and the vocabulary measure. These predictors were applied to the NWR scores of the group at each syllable length, for a total of six analyses. Correcting for these multiple comparisons, the regression model was significant (p < .008) for accuracy on 3-, 4,-, 5-, & 6-syllable nonwords. The variance explained by these models ranged from 29.8 - 37.5%. For the purpose of in-depth analysis, we chose to focus on the strongest model, which was for 4-syllable nonwords (see Table 1). Incidentally, this length was the mode digit span score for the group, so it represents performance on stimuli that are near the boundary between capacity and supra-capacity for these participants.

Table 1: Standardized regression model coefficients for accuracy on 4-syllable nonwords. Significant predictors are in bold.

	Statistical Values	
Variable	Std. β	p
Intercept	-.256	.251
Discrimination - Within		
00 vs. 20	-.087	.481
30 vs. 50	.039	.768
Discrimination – Between		
10 vs. 30	**.283**	**.029**
20 vs. 40	**-.320**	**.011**
Other Cognitive Measures		
Digit Span	**.574**	**<.001**
Vocabulary	.160	.172

In the case of 4-syllable nonwords, only measures of between category discrimination are a significant predictor of NWR in addition to DS. The coefficients for these two measures have opposite signs, suggesting individual differences in the category boundaries of our participants. For some participants, categorical discrimination occurs in the 10 v. 30 contrasts occur, but not the 20 v. 40 and vice versa. For some other participants, categorical discrimination may occur in both. Future analysis will be needed to entangle the nature of this relationship, however it is clear that such a relationship exists.

Success on between-category discrimination indicates a generalization of stimulus properties that leads to strong VOT categories. This generalization affects NWR positively as it can help the system process novel stimuli that align with the phonological structure of its vocabulary. Failure on between-category discrimination indicates a representation of specific acoustic detail rather than general categories, however these data suggest that one can still succeed at NWR under these conditions. Though the locus of these effects are not evident in these data, it is most likely the sequencing processes associated with both NWR and DS (the later of which is also a positive contributor to NWR) that are supporting NWR.

Computational Integration

The interpretation of the behavioral data just presented assumes an integrated system of vocabulary acquisition that incorporates phonological processing, STM, and vocabulary knowledge. Such a system has previously integrated STM and vocabulary knowledge (Gupta and Tisdale, submitted), but can it integrate phonological processing and show the same effects of specification and generalization observed in the behavioral data?

Model Architecture

The model is a modified version of a connectionist model developed by Botvinick and Plaut (2006) to model immediate serial recall within a recurrent network (for details on the psychological and technical motivations for architectural details see Botvinick & Plaut; Gupta & Tisdale, submitted). The model consists of an input layer that represents the individual syllables of a word. This layer has 21 units that can represent the constituent phonemes of a syllable within a CCVCC phoneme frame. 5 units are dedicated to representing the phonemes that may legally occupy the initial C slot of the frame (according to the phonological constraints of English), 3 units are dedicated to the second C slot, 5 units are dedicated to the V slot, 3 are dedicated to the penultimate C slot, and 5 are dedicated to the ultimate C slot (see Figure 1). The input layer also has 1 control unit that cues the network to recall a sequence.

The input layer is fully connected to a 200-unit hidden layer that is fully recurrent. That is, the hidden layer connects to itself through a set of modifiable weights that can transform the current activation of the hidden layer and affect processing at this layer for sequences of information. The hidden layer is also fully-connected, in an interactive fashion, to the output layer, which has the same representational scheme as the input. The only difference in the output layer is that the single control unit is replaced with a single unit that lets the network indicate the end of its recall of a sequence.

Training

The model was trained by having it shadow and recall a set of 4,386 English words that were between 2 and 4 syllables in length. These words were selected from a larger set of 130,000 phonologically distinct words that accompany the Festival speech synthesis software (Black & Taylor, 1997).

In each training epoch, the model was presented with each of these 4,386 words in a random order. For each word, the model was presented with one syllable at a time at the input layer and the target of the output was the syllable that was being presented. Following the last syllable of each word the "Recall" unit of the input layer was activated. This cued the network to enter the recall phase of training in which the model had to repeat each of the syllables that had just been presented, in the same order that they had been presented. No input was provided at the input layer and the target in this phase was the relevant syllable that was to be recalled. For the final syllable, the model was trained to activate the "Stop" unit of the output layer in addition to the phonological representation of the final output. Weight changes were made at the end of each word (including the recall phase) using recurrent back-propagation through time (Rumelhart, Hinton, & Williams, 1986). The context layer was reset between each word. The model was trained for a total of 100 epochs with a learning rate of 0.005 and no momentum term.

Vocabulary and Short-Term Memory

Gupta and Tisdale (submitted) showed that a network trained as described above could learn a substantial vocabulary and engage in NWR (i.e., repeat back words it had never processed before). Further, they (along with Botvinick and Plaut's [2006] initial demonstration in the domain of list recall) showed that STM can be thought of as the recurrence in this system and vocabulary knowledge can be thought of as the weights in the system. In this way, the model encodes not just the specific words it has seen, but information about the phonological structure of the language it is learning.

Vocabulary was measured in the current model by presenting each of the 4,386 words from training and measuring the accuracy of the output for each syllable during recall. If the mean squared error for the output vector of a particular syllable was less than .1, then that syllable was marked as correct. If all of the syllables in the word were correct, then that word was marked as correct. No weight changes were made during this, or any other, testing procedure.

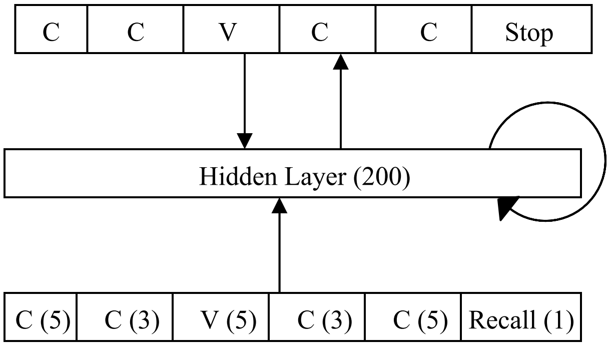

Figure 1: Architecture of the model, including the syllable structure that is represented at input and output. For the input and hidden layers, the number of units is given in parentheses.

NWR was measured in the model by presenting 2-, 3-, and 4-syllable nonwords (2 at each length) to shadow and recall. These nonwords were selected from the same corpus as those that were presented to the humans. Accuracy was measured in the same way as when assessing vocabulary. Reported values for both vocabulary and NWR are proportion correct.

Discrimination

A key insight provided by Botvinick and Plaut (2006) was that this type of network can eventually configure its weights so that upon recall, an entire list (remember, their task was list recall) is represented by the activation vector at the hidden layer when the recall cue is presented. This situation arises because this vector is the only information the model has to carry out its recall of the sequence it has just heard. For recall, the model systematically transforms this initial vector using the weight matrix that supports recurrence to produce each of the individual elements of the list.

Gupta and Tisdale (submitted), have subsequently shown that the same representation arises with whole words when the task is word recall. The whole word is initially represented at the hidden layer and is transformed to produce each of the individual syllables within the word. In this way, the activation of the hidden layer represents the similarity between different words and this information can theoretically be used to discriminate one word from another.

In the current model, we tested this idea by presenting the phonological representations of seven different categorical consonant contrasts and their intermediate representations. For example, in the implementation of the b/p continuum, the model was presented with "bowl" (/bol/) and "pole" (/pol/) as well as gradient representations of intermediate consonants. Specifically, the representation of "b" or VOT 0 was [0 0 1 0 0], the representation of "p" or VOT 50 was [0 0 0 0 1] and the intermediate representations reflected a shift between these two tokens by capturing the negative relationship between the two units that encode these consonants. VOT 10 was represented by [0 0 .8 0 0 .2], VOT 20 by [0 0 .6 0 0 .4], and so on. For each categorical contrast, we constructed the same number of intermediate representations by varying the value of the negatively related inputs. Each of these representations were then presented to the model during test and the activation vector at the hidden layer was recorded. The other six categories that were use were t/d, k/g, p/g, p/d, and r/l. The first five were consonants were presented along with /ol/ and r/l was presented along with / k/. While these are representations of English words, they were not words that were a part of the training corpus.

The measure of discrimination was derived from the normalized dot product (or similarity) of vectors associated with the representations of the contrasts between tokens that the humans made. From these dot products, we averaged across within- and between-category token discriminations to get a single measure of within- and between-category discrimination. We then calculated an overall *d'* for the model to capture how well it was distinguishing within- and between-category contrasts across each of the seven consonant contrasts.

Gain

In order to model individual variation in phonological representation and implement the differences in acuity and generalization that we believe to be evident in the human data, we manipulated the gain variable of the logistic function used to calculate activation values in the model. It is important to point out that this parameter is not one that has been developed specifically for the current application. Rather, it is an intrinsic and task-general parameter used in the implementation of many connectionist models. Further, we are making no claims that a neural gain equivalent leads to the individual differences in human perception in which we are interested. Gain is simply a way of changing the acuity of the representation, though other implementations could be used to the same effect.

Using gain instead of temperature in a traditional logistic activation function (cf. Rumelhart & McClelland, 1988, p. 71) results in the equation:

$$activation = \frac{1}{1 + e^{-net_i * gain}}$$

where net_i is the net activation received by a given unit. As gain increases in this equation, the resulting activation that arises from the same net activation sharpens (see Figure 2).

In the current framework, the effect of gain is to make words easier to discriminate at the hidden layer of the model. We trained and tested 310 different models using the method described above. For each gain value (ranging from .05 to 3.15 in .05 increments) 5 different models (i.e., different initial random weights) were trained and tested.

Averaging across each of the 5 simulations for a given gain value and then treating each gain value as an individual 'participant' allows us to perform a regression model using the models' overall vocabulary and *d'* to predict performance on NWR.

Results

NWR, vocabulary, and *d'* all vary as a function of gain (see Figure 3). The model performed best on these three measures when gain values were set between .5 and 1.25. Below .5, performance on all three measures dropped in a relatively linear fashion. Above these values, performance also dropped in a relatively linear way. However, at values above 2.35 vocabulary acquisition and NWR performance was exceptionally poor, while discrimination ability began to improve. We believe this pattern is likely due to the limits of representational flexibility as a function of the number of hidden layer units, and the number of connection weights with extremely high gain values. At higher gain values, the model may be forced into an unstable weight space that causes a form of catastrophic interference and prevents the acquisition of a large vocabulary, which also affects NWR. We are currently investigating this hypothesis with models that have 400 hidden layer units.

The regression model explains 93.7% of the variance in NWR performance. This is a significant amount, $F(2,60) = 458.53$, $p < .05$, and each of the predictors are also significant. The strongest predictor is vocabulary with a standardized β of .932, $p < .05$, followed by the intercept, β = -.142, $p < .05$, and *d'*, β = .097, $p < .05$. Thus, for both the human participants and the model, a measure of discrimination shows a significant relationship with NWR that is independent of a measure of vocabulary.

QuickTime™ and a
TIFF (Uncompressed) decompressor
are needed to see this picture.

Figure 2: The effect of changes in gain on the logistic activation function given in equation 1. Here gain changes from .1 to 1.5 and the activation function gets progressively steeper.

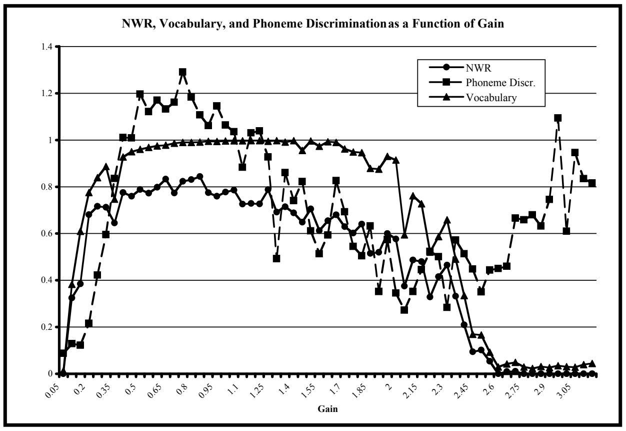

Figure 3: Three measures of performance in the model as a function of the gain parameter. NWR & vocabulary is given as proportion correct, whereas phoneme discrimination is *d*'.

Discussion

In both the human data and the computational model, the exact relationship between phoneme discrimination, NWR, and vocabulary is far from clear, but it has been demonstrated that a relationship exists and that it is a useful field of inquiry for the purposes of further explicating models of language acquisition. In the humans, only between-group categorization was predictive of NWR and the nature of this relationship varied depending on the exact stimulus tokens that were being used. In the model, manipulations of gain, or the sharpness of the activation function, had an effect on vocabulary, NWR, and phoneme discrimination. Further, a regression analysis on the modeling results showed a predictive relationship of phoneme discrimination and vocabulary on NWR. While the modeling and human measures are not equivocal, they are analogous and the existence of a relationship in both is a theoretically important contribution that deserves a more precise comparison and further investigation.

Overall, phonological processing, STM, and vocabulary acquisition can be thought of as a single integrated system whose overall functioning is reflected in the measure of NWR, but their exact relationship will take further untangling. The model presented here provides a specific computational framework for integrating the three processes discussed in this paper into a single account and investigating the relationship that they have with one another.

Acknowledgments

We would like to thank Matt Botvinick and Dave Plaut in their assistance with adapting their model, Marlea O'Brien, Connie Ferguson, Juanita Limas, Amy Schminke, and Marcia St. Clair for their assistance in collecting the human data, and Jeff Coady for providing the stimuli used in the phoneme discrimination task. We would also like to thank members of the Modeling Discussion Group at the University of Iowa and three anonymous reviewers for helpful discussion of this paper. This research was supported in part by grant NIH NIDCD R01 DC006499 to Prahlad Gupta.

References

Bates, E., & MacWhinney, B. (1987). Competition, variation and language learning. In B. MacWhinney (Ed.), *Mechanisms of Language Acquisition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Black, A. W., & Taylor, P. A. (1997). *The Festival Speech Synthesis System: System documentation* (Tech. Rep. HCRC/TR-83). Scotland, UK: Human Communication Research Centre, University of Edinburgh.

Botvinick, M. M., & Plaut, D. C. (2006). Short-term memory for serial order: A recurrent neural network model. *Psychological Review, 113*, 201-233.

Dunn, L., & Dunn, L. (1997). *Peabody Picture Vocabulary Test III.* Circle Pines, MN: American Guidance Service.

Gupta, P., Lipinski, J., Abbs, B., Lin, P.-H., Aktunc, E. M., & Ludden, D., et al. (2004). Space aliens and nonwords: Stimuli for investigating the learning of novel word-meaning pairs. *Behavioral Research Methods, Instruments, and Computers, 36*, 699-703.

Gupta, P., & MacWhinney, B. (1997). Vocabulary acquisition and verbal short-term memory: Computational and neural bases. *Brain and Language, 59*, 267-333.

Gupta, P., & Tisdale, J. (submitted). *Does phonological short-term memory causally determine vocabulary growth? Toward a computational resolution of the debate.* Unpublished manuscript.

Jusczyk, P. W. (2000). *The discovery of spoken language.* Cambridge, MA: The MIT Press.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart, & J. McClelland (Eds.), *Parallel distributed processing,* Vol. 1: Foundations. Cambridge, MA: MIT Press.

Saffran, J.R., Aslin, R.N., & Newport, E.L. (1996). Statistical Learning by 8-Month-Old Infants. *Science, 274,* 1926-1928.

Stager, C.L., & Werker, J.F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388,* 381-382.

Tomblin, B., Zhang, X., Buckwalter, P., & O'Brien, M. (2003). The stability of primary language disorder: Four years after kindergarten diagnosis. *Journal of Speech, Language, and Hearing Research, 46*, 1283-1296.

Wallace, G., & Hammill, D. (1994). *Comprehensive Receptive and Expressive Vocabulary Test.* Austin, TX: Pro-Ed.