JSLHR

Research Article

Large-Corpus Phoneme and Word Recognition and the Generality of Lexical Context in CVC Word Perception

Jessica T. Gelfand,[a] Robert E. Christie,[a] and Stanley A. Gelfand[b,c]

Purpose: Speech recognition may be analyzed in terms of recognition probabilities for perceptual wholes (e.g., words) and parts (e.g., phonemes), where *j* or the j-factor reveals the number of independent perceptual units required for recognition of the whole (Boothroyd, 1968b; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990). For consonant–vowel–consonant (CVC) nonsense syllables, $j \sim 3$ because all 3 phonemes are needed to identify the syllable, but $j \sim 2.5$ for real-word CVCs (revealing ~2.5 independent perceptual units) because higher level contributions such as lexical knowledge enable word recognition even if less than 3 phonemes are accurately received. These findings were almost exclusively determined with the 120-word corpus of the isophonemic word lists (Boothroyd, 1968a; Boothroyd & Nittrouer, 1988), presented one word at a time. It is therefore possible that its generality or applicability may be limited. This study thus determined *j* by using a much larger and less restricted corpus of real-word CVCs presented in 3-word groups as well as whether *j* is influenced by test size.
Method: The j-factor for real-word CVCs was derived from the recognition performance of 223 individuals with a broad range of hearing sensitivity by using the Tri-Word Test (Gelfand, 1998), which involves 50 three-word presentations and a corpus of 450 words. The influence of test size was determined from a subsample of 96 participants with separate scores for the first 10, 20, and 25 (and all 50) presentation sets of the full test.
Results: The mean value of *j* was 2.48 with a 95% confidence interval of 2.44–2.53, which is in good agreement with values obtained with isophonemic word lists, although its value varies among individuals. A significant correlation was found between percent-correct scores and *j*, but it was small and accounted for only 12.4% of the variance in *j* for phoneme scores ≥60%. Mean j-factors for the 10-, 20-, 25-, and 50-set test sizes were between 2.49 and 2.53 and were not significantly different from one another.
Conclusions: The j-factor based on a 450-word corpus and tri-word testing confirms and expands on findings from single-word presentations of isophonemic lists and a 120-word corpus. This enhances the generality (external validity) of the notions that $j \sim 2.5$ for real-word CVCs, and lexical knowledge enables CVC word recognition based on ~2.5 independent perceptual units. The robust nature of isophonemic word test outcomes is confirmed by close agreement with those provided by the high-reliability Tri-Word Test. Percent-correct performance was correlated with *j* but appeared to account for less than 13% of j-factor variance for most scores likely to be encountered in practice. Variability in the size of *j* suggests individual differences in the ability to take advantage of lexical knowledge in word recognition. The j-factor may be useful to inform rehabilitation needs, intervention content, and outcome assessment, as well as for other clinical applications.

Key Words: j-factor, lexical context, lexical knowledge, phonemic scoring, speech recognition, tri-word testing, word scoring

The assessment of speech recognition performance is a fundamental and ubiquitous aspect of audiological practice, with applications in both routine evaluation and differential diagnosis, as well as throughout the intervention process, from determining candidacy through outcome assessment. Moreover, speech recognition performance is at least implicitly involved in many of the assessments made by other professionals, such as speech-language pathologists, psychologists, and special educators. In spite of the wide range of applications, most measures involve some variation of testing recognition performance for monosyllabic real words, although nonsense syllable and sentence tests are also available.

Although measurement details can vary considerably, a fundamental characteristic of assessing speech recognition

[a]Touro College, New York, NY
[b]Queens College, Flushing, NY
[c]The Graduate Center, City University of New York
Correspondence to Stanley A. Gelfand: stanley.gelfand@qc.cuny.edu

Editor: Craig Champlin
Associate Editor: Marjorie Leek
Received June 14, 2012
Revision received December 7, 2012
Accepted May 23, 2013
DOI: 10.1044/1092-4388(2013/12-0183)

Disclosure: The third author, Stanley A. Gelfand, developed the Tri-Word Test.

performance typically involves the right–wrong scoring of whole words. However, the scoring of individual phonemes within test words has been increasingly used and encouraged (Boothroyd, 1968a, 1968b, 1993, 2006, 2008; Boothroyd & Nittrouer, 1988; Gelfand, 1998, 2003; Groen & Helleman, 1960; Guthrie & Mackersie, 2009; Keidser, 1993; Mackersie, Boothroyd, & Minniear, 2001; Markides, 1978; McCreery et al., 2010; Nittrouer & Boothroyd, 1990; Olsen, Van Tasell, & Speaks, 1997; Woods, Yund, & Herron, 2010), and the relationship between word and phoneme scores can provide insight into the listener's ability to take advantage of top–down perceptual processes (e.g., use of lexical context) in speech recognition (e.g., Boothroyd, 1968b; Boothroyd & Nittrouer, 1988; Bronkhorst, Bosman, & Smoorenburg, 1993; Nittrouer & Boothroyd, 1990; Olsen et al., 1997).

A theory-driven approach for examining the nature of the relationship between word recognition and phoneme recognition was introduced and elucidated by Boothroyd (1968b), Boothroyd and Nittrouer (1988), and Nittrouer and Boothroyd (1990). In the context of word recognition, word scores may be viewed as perceptual wholes and phoneme scores as perceptual parts. Here, the scores are expressed in the form of recognition probabilities from 0 to 1.0 instead of percentages. The recognition probability for words (wholes) is given as p_w, and the one for phonemes (parts) is given as p_p. The relationship between the recognition probabilities for perceptual parts and wholes is as follows:

$$p_w = p_p^j, \quad \text{where} \quad j = \frac{\log(p_w)}{\log(p_p)}.$$

The quantity j, commonly referred to as the j-factor, may be interpreted as representing the number of independent perceptual units (e.g., phonemic units) involved in the perception of the whole (e.g., words; Boothroyd, 1968b; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990). Hence, the j-factor for consonant–vowel–consonant (CVC) nonsense syllables is ~3.0 because syllable (whole) recognition requires the accurate reception of all three component phonemes (parts). In contrast, j is <3.0 for real-word CVCs because syllable (whole) recognition is still possible with the accurate reception of less than all three phonemes (parts), because of top–down processing contributions in speech perception, such as lexical context (Benki, 2003; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990; Olsen et al., 1997).

Typically reported mean values for the j-factor approximate 2.5 for real-word CVCs—such as 2.45 or 2.46 for young adults, 2.51 for children, 2.4 for older adults (Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990), and 2.53 and 2.54 for adults with normal hearing and sensorineural hearing losses, respectively (Olsen et al., 1997). Benki (2003) found a slightly lower mean of $j = 2.34$ among young adults and also found that the j-factor depended on the frequency of occurrence of the words, at 2.46 for low-frequency words and only 2.25 for high-frequency words. Overall, these investigators, among others, have interpreted these findings to indicate that the recognition of real-word CVCs involves the perception of roughly 2.5 independent phonemic units. Of course, different values are expected when other types of words are used. For example, Eisenberg, Shannon, Martinez, Wygonski, and Boothroyd (2000) found mean j-factors of ~2.2 for adults and ~2.0 for 5- to 12-year-olds when using phonetically balanced kindergarten word lists (Haskins, 1949), which are composed of CV, VC, and CVC words. Moreover, Felty (2007) found a mean j of 3.34 for German disyllabic words (compared with ~5.05 for nonwords) of the form CVCCVC.

The real-word CVC studies just described all used the same isophonemic word lists (Boothroyd, 1968a; Boothroyd & Nittrouer, 1988), which are composed of a small total word corpus of only less than 120 words;[1] these are further restricted by the requirement that all of the individual test lists be composed of words based on the same group of phonemes. As a result, the preponderance of j-factor values reported in the literature are based on a total of less than 120 test words for which word selection was restricted to achieve the isophonemic criterion. (One might consider in this context that j is influenced by word frequency [Benki, 2003]; and that smaller j values have been reported for CVCs that were originally spoken as nonsense syllables but that constituted real words [Woods et al., 2010].) As a result, the existing j-factor data are based on a word corpus that may well be too small and restricted to be representative of the CVC words in common use.

In addition, the existing j-factor findings were derived from the results of isophonemic word test lists containing a small number of scorable items (just 10 words and 30 phonemes, except for 20 words and 60 phonemes in the study by Olsen et al. [1997], which used double lists). This issue is worthy of consideration because test scores become less reliable as the test size decreases (Boothroyd, 1968b; Carney & Schlauch, 2007; Gelfand, 1998; Hagerman, 1976; Raffin & Schafer, 1980; Raffin & Thornton, 1980; Thornton & Raffin, 1978). Moreover, all but one study (Olsen et al., 1997) of the existing j-factor values used samples with 40 or fewer participants per group.

In addition, all of these studies were based on data from words presented one at a time, rather than in the context of other words. Yet single-word utterances make up a very small part of real-world conversational speech, and multiple-word testing has been proposed for clinical purposes (Gelfand, 1998, 2003; Gelfand & Gelfand, 2012; Haagen, 1945; Harris, 1980; Olsen, 1983; Sergeant, Atkinson, & LaCroix, 1979; Watson & Knudsen, 1940; Williams, Mosko, & Greene, 1976; Wilson, Burks, & Weakley, 2005; Wilson & Weakley, 2004).

Considerations like these make it desirable to determine the nature of the j-factor with a larger and less restricted word corpus, as well as with multiple- rather than single-word

[1]Although that test corpus has 12 ten-word lists, eight of the words are repeated across lists, so the number of different words is actually 112. A more recent version of the isophonemic materials includes 200 words in 20 ten-word lists (Boothroyd, 2006; Mackersie, Boothroyd, & Minniear, 2001).

presentations, to establish the generality (external validity) of the value $j \sim 2.5$ and to inform its eventual clinical applicability. The only clinical use of the j-factor we are aware of to date involves serving as a means for converting between phoneme- and word-recognition scores (Boothroyd, 2008). However, others readily come to mind because a patient's j-factor value might be used to reveal difficulties with top–down aspects in speech perception. For example, it would be especially relevant for ascertaining the need for, as well as the content and outcomes of, aural rehabilitation services. It should also find other clinical applications where it is desirable to consider top–down processing, such as when assessing nonnative speakers of the language, aging individuals, and those with possible auditory processing disorders.

Assessing the generality of a j-factor approximating 2.5 is readily achievable with the Tri-Word Test (originally described as the Computer Assisted Speech Recognition Assessment or CASRA test), which involves a large corpus of 450 real-word CVCs and three-word stimulus presentations (Gelfand, 1998, 2003). The Tri-Word Test was developed as an approach to reach a workable compromise solution for clinicians and researchers who need speech recognition tests that are both highly reliable and as short as possible (Gelfand, 1998, 2003). Each administration of the Tri-Word Test includes 150 test words presented in 50 three-word groups that have been prescreened to exclude semantic or syntactic cues and rhyming words. With phonemic scoring, the test produces recognition scores based on 50 Sets × 3 Words × 3 Phonemes = 450 scorable elements, which have been shown to optimize reliability.[2] This is accomplished with the same number of stimulus presentations used by traditional speech recognition tests, such as the Central Institute for the Deaf W-22 test (Hirsh et al., 1952) and Northwestern University Auditory Test Number 6 (Tillman & Carhart, 1966).

This study was undertaken to assess the external validity of the reported j-factor value (~2.5) that has been derived from a relatively small and restricted (typically ≤120) word corpus (Benki, 2003; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990; Olsen et al., 1997; Woods et al., 2010), by comparing it to the j-factor obtained with a large and less restricted corpus of 450 CVC words. Specifically, j-factors were derived from word scores and phoneme scores obtained from a diverse sample of participants using the Tri-Word Test. This test includes a corpus of 450 real-word CVCs. Each administration involves a unique randomization of 150 test words presented in 50 three-word groups (which are prescreened to exclude semantic or syntactic cues and rhyming words). Finding j values of ~2.5 under these conditions would (a) enhance the external validity of j-factor measurements, (b) reinforce the principle that the perception of CVC words involves ~2.5 independent units because of lexical redundancy, and (c) support the validity of the Tri-Word Test itself.

Because shortened versions of the Tri-Word Test consisting of 10, 20, or 25 three-word presentations may be appropriate in a variety of situations (Gelfand, 2003; Gelfand & Gelfand, 2012), the second part of this study addressed whether the j-factor is influenced by shortened versions of the Tri-Word Test.

Method

Participants

There were 223 participants in our study sample, which was composed of 151 individuals with normal hearing with pure-tone thresholds of ≤10 dB HL (American National Standards Institute [ANSI], 2004) from 250 to 8000 Hz and 72 persons with a wide range of sensorineural hearing losses. The participants with normal hearing included 107 women and 44 men, ranging in age from 18 to 40 years (M = 21.9 years). Those with sensorineural hearing losses included 24 women and 48 men, ranging in age from 31 to 68 years (M = 57.4); pure-tone thresholds for this group are summarized in Table 1. The participants were unpaid volunteers recruited from the academic and surrounding communities. Informed consent was obtained in accordance with a protocol approved by the Queens College institutional review board.

All participants were native speakers of American English with no history or signs of neurological or learning disorders and whose first-administration word and phoneme scores on the Tri-Word Test were between 5% and 95%. Only first test administrations were included to reduce the potential influence of experience and/or practice with the paradigm and to be as representative as possible of the unpracticed conditions of clinical testing. Performance scores were limited to the 5%–95% range because j is sensitive to measurement errors for extreme scores (Boothroyd & Nittrouer, 1988).

A broad continuum of speech recognition performance was achieved by including listeners with a wide range of auditory status from normal hearing through severe sensorineural hearing loss. Those with normal hearing had pure-tone thresholds of ≤10 dB HL (ANSI, 2004) from 250 to

Table 1. Pure-tone thresholds in dB HL (American National Standards Institute, 2004) in the tested ears of the participants with sensorineural hearing losses.

Frequency	*M* (dB HL)	*SD* (dB)	Range (dB HL)
250 Hz	16.18	10.73	0–40
500 Hz	23.47	14.23	0–55
1000 Hz	33.13	16.32	10–75
2000 Hz	47.78	16.01	25–85
4000 Hz	64.03	17.25	40–105
8000 Hz	65.76	19.37	35–95[a]

[a]No response for 8000 Hz pure tones presented at 90 dB HL was coded as 95 dB HL.

[2]The marginal reliability improvement that results from adding items gets progressively smaller as n rises and becomes negligible when n reaches ~450, so practical reliability is optimized with ~450 scorable items (Gelfand, 1998).

8000 Hz and no evidence of otological disorders. Those with sensorineural hearing loss were heterogeneous in terms of the degrees and configurations of their hearing losses and routine speech recognition scores. Analogous approaches have been used in other studies (e.g., Benichow, Cox, Tun, & Wingfield, 2012; Boothroyd, 1968b; Gelfand, 2003; Markides, 1978; Olsen et al., 1997; Raffin & Schafer, 1980).

Instrumentation and Materials

The instrumentation has been described in detail previously (Gelfand, 1998). Briefly, digitally stored speech materials were directed from a personal computer via an input–output board and a 10000 Hz low-pass (anti-aliasing) filter to an audiometer (Grason-Stadler Model 16) and were presented to the participants through standard supra-aural audiometric earphones (Telephonics Model TDH-50P). All tests were done in a sound chamber meeting noise level standards for audiometric rooms (ANSI, 2003).

The test materials consisted of the Tri-Word Test (i.e., CASRA; Gelfand, 1998), in which each administration involves a unique randomization of 150 words selected from a test corpus composed of 450 real-word CVCs. The test items were presented without a carrier phrase in the form of 50 groups of three words each and were scored on the basis of correct identification for both individual phonemes and whole words. The three-word groups of unrelated words in the Tri-Word Test may be considered analogous to the four-word zero-probability "sentences" used by Boothroyd and Nittrouer (1988).

The test words were spoken by a male talker with equal vocal effort and were digitally stored, as was a 1000 Hz tone scaled to the overall root-mean-square (RMS) level of the word pool for use as a calibration signal. All of the test items have real-word alternatives if one or both of their consonants are misheard (e.g., *pat, cap,* or *bad* for *cat*).

The test lists for each administration were generated in advance with software that divided the 450-word pool quasi-randomly into three nonoverlapping groups of 150 words each. These were in turn arranged into 50 presentation sets of three words each. The test generation program also minimized the occurrence of rhyming, semantic relationships (e.g., *dog–cat–pet*) and syntactic cues (e.g., *man–goes–home*) among the words in each presentation group. As a result, each administration of the test involved a unique randomization of the words.

Procedure

The testing procedures were similar to those described by Gelfand (2003). Each participant had an initial interview and routine audiological measures to establish auditory status, followed by administration of the Tri-Word Test.

Participants were instructed to listen to all three words and then to repeat them as they were heard. Consistent with Markides's (1978) recommendations for phonemically scored tests, they also were told that partial credit would be given for every part of a word that was repeated correctly.

The words were presented monaurally to the ear preferred by the listener. Listeners with sensorineural hearing loss heard the materials at 40 dB sensation level (SL) in relation to speech recognition threshold or at the highest comfortable level when 40 dB SL was too loud. To achieve a wide range of performance scores for participants with normal hearing, a presentation level between 0 and 40 dB HL, or at 40 dB HL with a signal-to-noise ratio (SNR) between –15 and 20 dB, was selected at random for each listener.[3] A given listener would then receive all of the test words at that level or SNR.

For each tri-word trial, the three test words were displayed on the tester's monitor and then presented to the listener. The test items were then repeated by the listener, after which the response was scored by the tester on a phoneme-by-phoneme basis. Incorrectly repeated phonemes were recorded by pressing their corresponding numbers (1–9) on a keyboard and were displayed as *X*s on the monitor. Once the tester was satisfied with the proper recording of the responses, the next trial would be administered. After all 50 trials were completed, the software tallied the results, calculated phoneme and word scores, and stored the data. Listening errors by the tester were avoided by having the participant and tester face each other through the booth window so that responses were seen and heard and by asking for clarification whenever the tester was not completely sure of the intended response.

The resulting data files were screened prior to analysis to ensure that the sample included only the results of each participant's first administration of the test and that both the phoneme and words scores were between 5% and 95%. First administrations were used to be maximally representative of the unpracticed conditions of clinical testing as well as to avoid the influences of such factors as experience, practice, and fatigue. Performance scores were limited to 5%–95% because the j-factor is sensitive to measurement errors for extreme scores. The resulting sample of 223 participants was involved in the first part of the study. The second part of the study was based on the findings of a subsample of 96 participants whose results could be reanalyzed to produce separate scores for the first 10, the first 20, and the first 25 sets, as well as all 50 of them. Although this does not produce independent scores for each test size, it does allow one to identify what the results would be if the test size had been based on 10 sets, 20 sets, and so forth. Analogous approaches have been used in other studies (Gelfand, 2003; Gelfand & Gelfand, 2012; Raffin & Schafer, 1980).

Results

Part 1: J-Factor With a 450-Word Corpus and Tri-Word Presentations

Performance scores. Measured percent-correct speech recognition performance ranged from 8.2% to 94.9% for

[3]These levels and SNRs encompass the full range of performance on the Tri-Word Test (Gelfand, 1998; Gelfand & Gelfand, 2012).

phoneme scores and from 5.3% to 88.7% for word scores, as illustrated by the individual points on the bivariate graph in Figure 1. The scattergram shows that a relatively small number of the scores were lower than about 40% for words and 20% for phonemes, which was expected because extremely low speech recognition scores are rare under typical testing conditions.[4] Overall, these findings suggest that the intended goal of including a wide range of performance was achieved, particularly with respect to the breadth of scores commonly encountered in clinical practice.

As shown in Table 2, mean recognition performance was 80.38% (SD = 17.55%) for phonemes and 63.3% (SD = 21.98%) for words. The mean difference between phoneme and word scores was 17.1% (SD = 7.21%), which was consistent with the values typically encountered in the literature (Boothroyd, 1968a, 1968b; Gelfand, 1998, 2003; Keidser, 1993; Markides, 1978; Olsen et al., 1997).

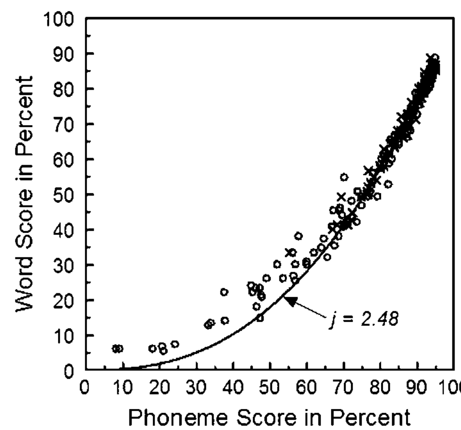
J-factor. The value of the j-factor was calculated for each participant on the basis of his or her phoneme and word scores expressed as proportions. Figure 2 illustrates the distributions of individual values of j collapsed across all levels of performance (percent-correct scores) for all participants combined and separately for those with normal hearing and sensorineural hearing loss. Overall, these values ranged widely, from 1.1249 to 3.2679. Inspection of the data revealed that the wide range largely reflected the influence of the cases in the lower left quadrant of Figure 1. These cases had low percent-correct scores and were obtained from normal-hearing listeners who were tested at low presentation levels or SNRs.

Table 2 shows that the mean value of the j-factor in this study was 2.48, with a standard deviation of 0.34, a standard error of 0.02, and a 95% confidence interval from 2.44 to 2.53. As illustrated in Figure 3, similar values have been reported by others (Benki, 2003; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990; Olsen et al., 1997). As expected, the j-factors of the participants with normal hearing (M = 2.47, SD = 0.37) and sensorineural hearing losses (M = 2.51, SD = 0.26) were not significantly different, $t(221) = 0.97$, $p > .05$, Cohen's $d = 0.15$, which is consistent with the findings of Olsen et al. (1997). The smooth curve in Figure 1 is based on the average value of $j = 2.48$ and provides a close fit to most of the data, with an RMS error of only 3.44. However, it is noteworthy that the data points from the relatively small number of cases with very low phoneme scores clearly deviated from the curve.

The relationship between the size of the j-factor and percent-correct speech recognition performance is illustrated in Figure 4, which suggests that there is a positive relationship between the size of the j-factor and percent-correct speech recognition performance. The correlations between individual j-factors and recognition performance were moderate though highly significant for both phoneme scores

[4]Therefore, the performance range was extended down to ~5% by including scores obtained at low presentation levels or SNRs from normal-hearing participants.

Figure 1. Individual phoneme and word scores in percent on bivariate coordinates. Data from participants with normal hearing are shown as circles, and those from participants with hearing loss are shown as *X*s. The smooth line shows predicted word scores as a function of phoneme scores based on a j-factor of 2.48.



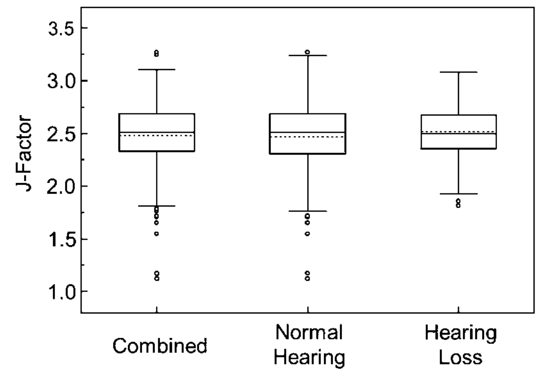
($r = 0.699$, $p < .001$) and word scores ($r = 0.584$, $p < .001$). This relationship was explored further because both Boothroyd and Nittrouer (1988) and Benki (2003) did not find a correlation between phoneme scores and the j-factor. Because the j-factor is based on the relationship between phoneme and word scores, we considered it worthwhile to examine the relationship in the context of percent-correct performance. (Findings are reported only for phoneme scores because the outcome based on word scores was similar and redundant.)

Figure 1 shows that phoneme and word scores converge as performance decreases toward 0% and increases toward 100%, as expected. Consequently, the phoneme–word difference decreases when phoneme scores approach either 0% or 100%, as revealed in Figure 5a. Figure 5b shows the j-factor in the context of this relationship, which follows a different pattern for the 27 cases with phoneme scores less than 60% (filled circles) compared with the preponderance of cases with scores of 60% and above (crosses). On the basis of these patterns, the relationships between the j-factor and both the phoneme–word difference and phoneme scores were analyzed separately for scores above and below 60%, as shown in Figure 6. The correlation between the j-factor and the phoneme–word difference was significant for scores less than 60% ($r = 0.958$, $p < .001$) but not for scores of 60% and above ($r = –0.037$, $p > .05$). There was a significant correlation between the j-factor and phoneme scores of less than

Table 2. Means, standard deviations, and 95% confidence intervals for phoneme scores, word scores, difference between phoneme and word scores, and the j-factor.

Variable	*M*	*SD*	95% CI
Phoneme scores (%)	80.38	17.55	[78.06, 82.69]
Word scores (%)	63.28	21.98	[60.38, 66.18]
Difference (%)	17.10	7.21	[16.15, 18.05]
j-factor	2.48	0.34	[2.44, 2.53]

Figure 2. Box-and-whisker plots with means (dashed lines) and outliers (individual data points) illustrating the distributions of j-factor values collapsed across all percent-correct scores for the entire sample (left) and separately for participants with normal hearing (center) and those with hearing loss (right).



60% ($r = 0.702$, $p < .001$). Although significant, the correlation with phoneme scores of 60% and higher was low ($r = 0.358$, $p < .001$) and accounted for less than 13% of j-factor variance ($R^2 = 0.124$).

Part 2: J-Factor for Shortened Versions of the Tri-Word Test

Percent-correct speech recognition performance scores for the four test size conditions are summarized in Table 3.

Figure 3. The mean and 95% confidence interval of the j-factor for real-word CVCs obtained in the current study along with corresponding findings from studies using isophonemic word lists (Benki, 2003; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990; Olsen et al., 1997) and Dutch CVC words (Bronkhorst et al., 1993). Confidence intervals for other studies are shown if they were reported or could be estimated from reported standard errors. BN = Boothroyd and Nittrouer (1988); NB = Nittrouer and Boothroyd (1990) for children (ch), younger adults (ya), and older adults (oa); O = Olsen et al. (1997) for those with normal hearing (nh) and sensorineural hearing loss (sn); Be = Benki (2003) for all words (all), high frequency of occurrence words (hf), and low frequency of occurrence words (lf); Br = Bronkhorst et al. (1993) findings for words presented in quiet (q) and in noise (n).

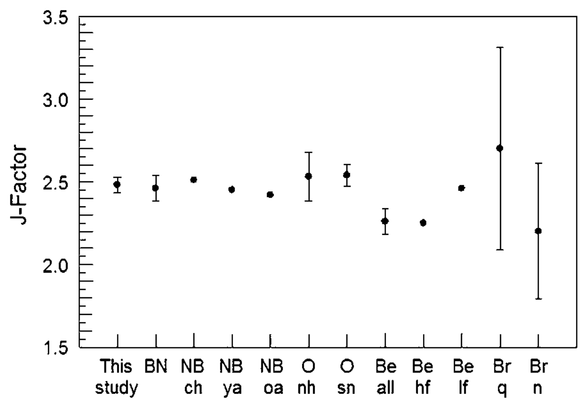
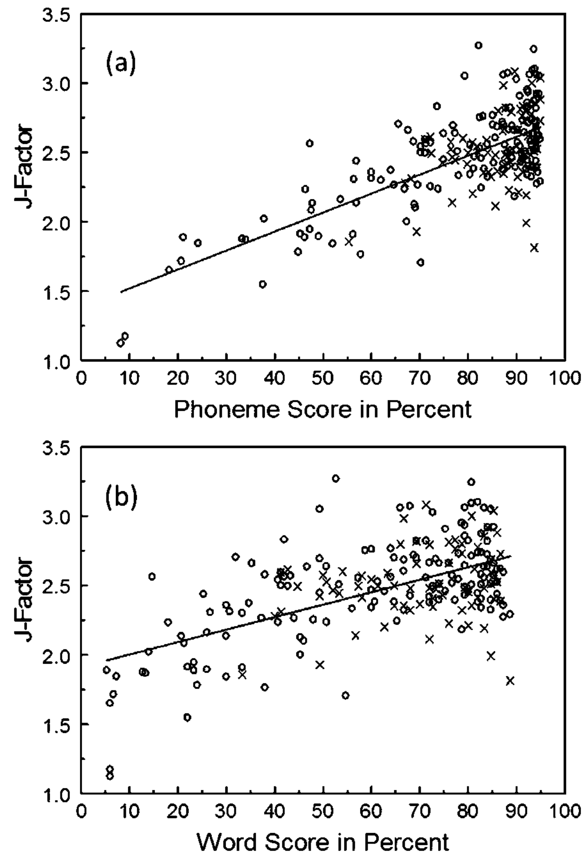


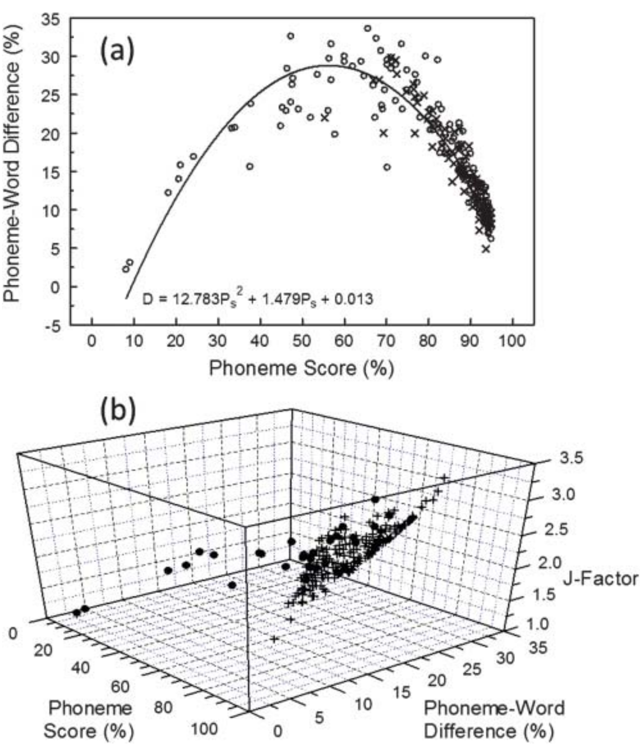
Figure 4. Scattergrams and best-fit linear regression lines showing the relationship between the sizes of the j-factor and the correct percent of speech recognition performance expressed in terms of (a) phoneme scores and (b) word scores. The regression lines are of the form $j = 0.014 \times$ Word Score + 1.383 and $j = 0.009 \times$ Phoneme Score + 1.913. Circles represent participants with normal hearing, and *X*s represent those with hearing loss.



The narrower distributions of these scores compared with those found in the first part of the study were expected because of the additional requirement that phoneme and word scores be between 5% and 95% for all four of the test size conditions. It was considered instructive to provide at least some quantitative evaluation of the test size scores obtained with the different test sizes even though the outcome must be viewed very conservatively because the measures were not independent. Scores obtained across test sizes were therefore examined using one-way repeated measures analyses of variance (ANOVAs). As anticipated, there were no significant differences across test sizes for both phoneme scores, $F(3, 285) = 1.869$, $p > .05$, $\eta^2 = 0.019$; and word scores, $F(3, 285) = 0.691$, $p > .05$, $\eta^2 = 0.007$.

The participants' j-factors for the four test size conditions were calculated from their respective phoneme and word scores. Table 4 displays the ranges, means, standard deviations, standard errors, and 95% confidence intervals for the j-factors calculated for each of the four presentation set

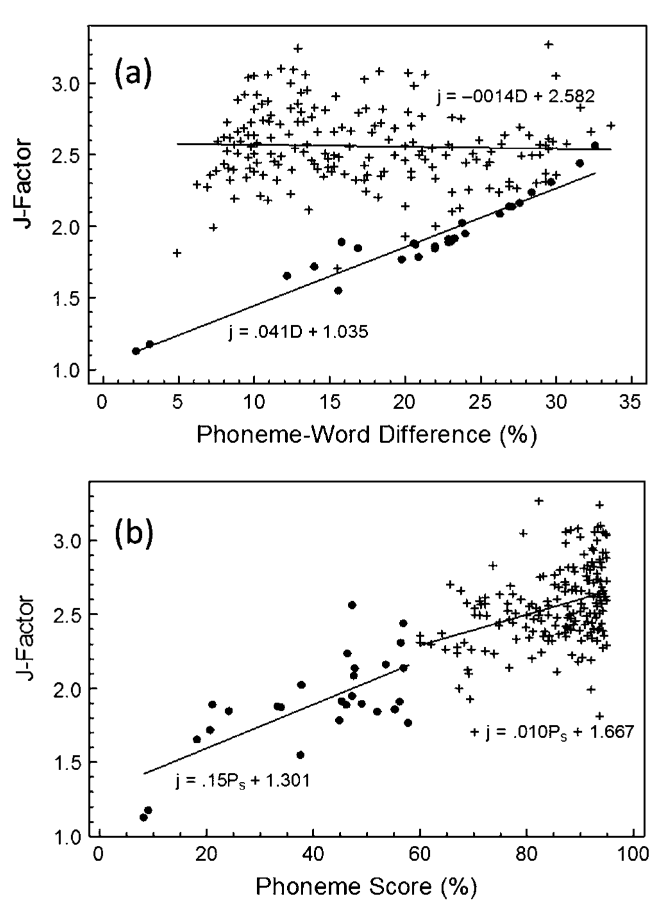
Figure 5. (a) Scattergram of the phoneme–word difference (D) as a function of phoneme score (P_s), with the relationship highlighted by a second-order polynomial regression line. Circles represent data from participants with normal hearing, and *X*s represent data from those with hearing loss. (b) Three-dimensional scattergram illustrating the relationship between the j-factor, phoneme score, and the word–phoneme difference. Different symbols highlight the patterns of findings for the cases with phoneme scores <60% (filled circles) versus those with phoneme scores ≥60% (crosses).



conditions. Corresponding findings from Part 1 are also shown for reference. As shown in the table, the ranges of individual j-factors narrowed with increasing test size, from 1.107–3.385 for 10 sets (30 words) to 1.477–3.179 for 50 sets (150 words). These ranges of *j* values were consistent with the distributions found for the full sample, described above. The j-factors based on 50, 25, and 20 presentation sets provided fits to the data that were comparable to that of the full sample, with RMS errors of 3.20, 3.97, and 3.87, respectively, but the RMS error increased to 5.08 for the 10-set condition.

Similar to the findings for the full sample, Table 5 reveals that the j-factors based on the abbreviated versions of the Tri-Word Test were significantly correlated with both phoneme and word scores for all conditions except for the 10-set word scores (which approached statistical significance). However, there was a clear decrease in the strength of these correlations going from the 50-set findings from Part 1 of the study to those obtained with the more restricted sample involved in Part 2, with another clear drop in the strength of the correlations for the 10-set test size.

Figure 6. Scattergrams and regression lines for the cases with phoneme scores <60% (filled circles) versus ≥60% (crosses) for the j-factor as a function of (a) phoneme–word differences (D) and (b) phoneme scores (P_s).



The correlations among the j-factors yielded by the four set sizes are provided in Table 6. As expected, all of the j-factors were significantly correlated with each other ($p < .0001$), and a one-way repeated measures ANOVA revealed that there

Table 3. Phoneme and word scores (%) for 10-, 20-, 25-, and 50-set tri-word presentations.

Test condition	Range	*M*	*SD*
Phoneme scores			
10 sets	25.56–94.44	82.29	12.63
20 sets	24.44–94.44	82.48	11.76
25 sets	25.33–94.22	82.65	11.74
50 sets	30.89–94.67	82.942	11.46
Word scores			
10 sets	6.67–86.67	65.00	17.12
20 sets	10.00–85.00	64.36	15.95
25 sets	10.67–85.33	64.76	15.98
50 sets	14.00–86.67	64.97	16.11

Table 4. The j-factors obtained with the four set-size conditions (subsample).

Test size condition	Range	*M*	*SD*	*SE*	95% CI
Full sample (50 sets; *n* = 223)	1.13–3.27	2.48	0.34	0.02	[2.44, 2.53]
10-set subsample (*n* = 96)	1.11–3.39	2.49	0.48	0.05	[2.39, 2.58]
20-set subsample (*n* = 96)	1.23–3.31	2.53	0.40	0.04	[2.45, 2.61]
25-set subsample (*n* = 96)	1.29–3.30	2.52	0.40	0.04	[2.43, 2.60]
50-set subsample (*n* = 96)	1.48–3.18	2.53	0.37	0.03	[2.46, 2.60]

Note. Corresponding data from the first part of the study (full sample) are shown in the first row for comparison.

were no significant differences between the j-factors for the four test size conditions, $F(3, 285) = 1.137, p > .05, \eta^2 = 0.012$.

Discussion

Most studies have shown that the average j-factor is roughly 2.5 for real-word CVCs, suggesting that the recognition of these words involves the accurate reception of ~2.5 independent perceptual units (e.g., Benki, 2003; Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990; Olsen et al., 1997). In contrast, the recognition of CVC nonsense syllables relies on the correct reception of all three component phonemes (i.e., $j \sim 3.0$, indicating that three independent perceptual units are involved). This advantage for the perception of real words is generally accepted as revealing the contributions of higher-level perceptual processes, particularly those relating to lexical knowledge. However, the external validity (generality) of these findings may be limited because they were based on the relatively small total corpus of 120 words or less in 12 ten-word test lists containing the same phonemes. In addition, all of the findings to date were based on test words that were presented one at a time, so it was not clear whether $j \sim 2.5$ for CVC words in groups. The current study was undertaken to address these issues by determining j-factor values for a large corpus of 450 real-word CVCs presented in groups of three words each.

The principal finding of the current study was that the mean value of the j-factor for a corpus of 450 real-word CVCs presented three at a time is 2.48, with a 95% confidence interval between 2.44 and 2.53. Consequently, the current study provides evidence supporting the generality of the notion that the recognition of real-word CVCs relies on the correct identification of ~2.5 independent perceptual elements, which is interpreted to reflect the contributions of lexical knowledge in spoken word perception.

That the efficient use of three-word presentations produces the same j-factor commonly found with single-word presentations provides support for the use of the tri-word testing approach in clinical speech recognition assessment. Moreover, the robust nature of isophonemic word test outcomes is highlighted by close agreement between the results of the studies using the shorter isophonemic lists and those reported here for the Tri-Word Test, which provides optimized reliability (Gelfand, 1998).

As shown in Figure 3, the current findings are quite comparable to those reported for various groups of participants with normal hearing (Boothroyd & Nittrouer, 1988; Nittrouer & Boothroyd, 1990; Olsen et al., 1997) as well as those with a wide range of sensorineural hearing loss (Olsen et al., 1997). A possible exception is the mean j-factor of 2.42 for older adults (Nittrouer & Boothroyd, 1990), which fell slightly below our 95% confidence interval. There is also good agreement with the findings reported by Bronkhorst, Brand, and Wagener (1993) for Dutch CVC words presented in quiet (but not in noise). Slightly lower j-factor estimates obtained with isophonemic word lists were reported by Benki (2003), who found an overall mean of $j = 2.34$ for young adults. More interesting, however, was his demonstration that the j-factor is influenced by the frequency of occurrence of the test words, which highlights the importance of lexical factors in word recognition.

Table 5. Correlation coefficients and significance values between j-factors with phoneme and word scores for the four set-size conditions (subsample).

	j-factor with phoneme scores		j-factor with word scores	
Test size condition	*R*	*p*	*R*	*p*
10-set subsample (*n* = 96)	.441	<.0001	.199	.053
20-set subsample (*n* = 96)	.529	<.0001	.327	<.001
25-set subsample (*n* = 96)	.500	<.0001	.286	<.005
50-set subsample (*n* = 96)	.539	<.0001	.358	<.0001
Full sample (50 sets; *n* = 223)	.699	<.0001	.584	<.0001

Note. Corresponding data from the first part of the study (full sample) are shown in the last row for comparison.

Table 6. Correlation coefficients (and significance) among the j-factors for 10-, 20-, 25-, and 50-set Tri-Word Test sizes.

Test size	10 sets	20 sets	25 sets	50 sets
10 sets	—	.824 *	.764*	.703*
20 sets		—	.953*	.856*
25 sets			—	.912*
50 sets				—

*$p < .0001$.

An interesting exception to these findings was found by Woods et al. (2010), who reported the perceptual outcomes for a very large sample of CVCs composed of 20 initial and final consonants in the context of the vowels /i/, /ɑ/, and /u/. They found mean j-factors of only 1.70 for CVCs that were originally spoken as nonsense syllables but corresponded to real words and 1.93 for 797 CVCs that actually were meaningless. Woods et al. suggested that these values (i.e., $j < 2$) may reflect independent processing of the initial and final consonants (effectively as CV and VC instead of CVC) involved in their materials. It is not clear whether this outcome may have been influenced by such considerations as an expansion of the stimuli beyond the stricter restrictions involved in using only CVC words, how the syllables were spoken, the existence of only three vowel alternatives (which participants knew in advance as a result of familiarization with a list of possible consonants and vowels as well as practice with the materials), or other factors. However, what is noteworthy in the context of the current discussion is that the perceptual impact of these factors was readily revealed by the values of the j-factor.

In addition to confirming the overall magnitude of the j-factor, another notable finding pertains to the distribution of j-factor among individual listeners, illustrated in various contexts in Figures 2, 4, 5, and 6. The sizable range of individual j-factor values obtained in this study suggests that people vary widely in their ability to take advantage of lexical knowledge in speech perception. This variability points to potentially useful applications of the j-factor in clinical practice. For example, calculating the j-factor for a patient on the basis of combined phoneme and word scoring of a test might help identify difficulties with the top–down aspects of speech recognition. This would be especially useful in the realm of audiological management. Here, the j-factor may serve as a means of identifying patients who would benefit from intervention, inform the nature and content of communication skills counseling and/or direct practice, and also serve as a means for assessing the outcome of these interventions. Other potential clinical applications of the j-factor may include the assessment of auditory processing disorders, age-related issues, and the impact of diverse linguistic backgrounds among the hearing impaired, among others. Further research appears to be warranted in areas such as these.

A significant relationship was found between the magnitude of the j-factor and speech recognition performance in terms of both phoneme scores and word scores. This finding was not expected because Boothroyd and Nittrouer (1988) and Benki (2003) did not find correlations between the j-factor and phoneme scores for isophonemic word lists. (However, Boothroyd and Nittrouer [1988] and Bronkhorst et al. [2002] did find significant correlations between the j-factor and performance scores for words in sentences.) It is noteworthy that the size of this correlation decreased substantially from ~0.70 for all 223 cases to ~0.36 for the 196 cases with phoneme scores of ≥60%, which although significant accounted for less than 13% of the variance in the j-factor. Phoneme scores in this range roughly correspond to word scores over ~40% and certainly encompass the overwhelming majority of speech recognition scores typically encountered in clinical practice. As such, we do not believe the correlation observed here would, if confirmed by others, have any substantive practical impact on the use of the j-factor.

It is not clear why the correlation was found here but not in other studies, but several possibilities might be considered. It is possible that variations between the studies involving such factors as the distribution of scores, test size, and sample size may be plausible contributors. For example, the current study included performance scores as low as 8.2% for phonemes and 5.3% for words, whereas the lowest real-word phoneme scores were ~43% in Boothroyd and Nittrouer (1988; based on Figure 4B, p. 105) and ~42% in Benki (2003; based on the upper left panel of Figure 2, p. 1694). This line of reasoning is consistent with the small (albeit significant) correlation found here when phoneme scores below 60% were excluded. It is also noteworthy in this regard that the coordinates of the lowest phoneme and word scores in Figure 1 deviated from the values based on $j = 2.48$. In addition, although all but one of the correlations between j and performance scores reported here were statistically significant, the strength of the correlations for the 50-set data declined with decreasing sample size from Part 1 ($n = 223$) to Part 2 ($n = 96$) and dropped further going from the larger test size (50, 25, and 20 set) conditions to the smallest one (10 sets). Moreover, the studies that did not find significant correlations had samples sizes of 32 (Boothroyd & Nittrouer, 1988) and 37 (Benki, 2003),[5] compared with 223 participants in Part 1 of the current study and 96 in Part 2.

The second part of the study examined j-factor outcomes associated with 10-set (30 word), 20-set (60 word), 25-set (75 word), and 50-set (150 word) Tri-Word Test sizes. The j-factors obtained with these four test sizes had mean values between 2.49 and 2.53 and were highly correlated with one another, and there were no significant differences between them. These results appear to support the notion that the j-factors yielded by the shortened versions of the Tri-Word Test are similar to those obtained with the full-length test and previous reports with isophonemic word lists. However, the RMS error of 5.08 for the 10-set condition was wider than those found for the longer tests. It is interesting that this value, which was based on scores for 30 words and 90 phonemes, was comparable to the RMS errors obtained by Olsen et al. (1997) for scores based on 20 words and 60 phonemes (5.4 for their normal-hearing group and 6.4 for their hearing loss group). Thus, the fit of the average j-factor to individual values appears to be influenced by the number of items scores, at least for shorter tests.

Conclusion

This study revealed that the j-factor obtained with tri-word presentations based on a 450-word corpus is similar to the values typically obtained with single-word presentations

[5]The j-factor was based on 37 of the 43 participants involved in the study by Benki (2003).

of short isophonemic lists with a 120-word corpus. It may be concluded that these findings confirm and expand on existing reports and enhance the generality of the notions that the j-factor approximates 2.5 for CVC real words and that lexical knowledge enables the accurate recognition of CVC words based on ~2.5 independent perceptual units.

In contrast to prior findings, the j-factor was found to be correlated with percent-correct performance, although this correlation accounted for less than 13% of the variance in the j-factor for the range of scores likely to be encountered in clinical practice.

The size of the j-factor varies among individuals, implying that people differ in their ability to take advantage of lexical knowledge in word recognition. The j-factor may be useful to inform rehabilitation needs, intervention content, and outcome assessment, as well as in other clinical applications.

Estimates of the j-factor provided by tri-word testing with a variety of test sizes are consistent with those reported in the literature for participants with a wide range of hearing sensitivity. Close agreement between the results obtained with the short isophonemic word and the high-reliability Tri-Word Test highlights the robust nature of isophonemic word list outcomes.

Acknowledgments

This work was supported in part by City University of New York PSC-BHE Grants 663451, 665535, and 666535. Parts of the article were included in presentations at the annual conventions of the American Speech Language and Hearing Association (2012, Atlanta, GA) and the American Academy of Audiology (2013, Anaheim, CA).

References

American National Standards Institute. (2003). *Maximum permissible ambient noise levels for audiometric test rooms* (ANSI S3.1-1999 R2003). New York, NY: Author.

American National Standards Institute. (2004). *Specifications for audiometers* (ANSI S3.6-2004). New York, NY: Author.

Benichow, J., Cox, L. C., Tun, P. A., & Wingfield, A. (2012). Word recognition within a linguistic context: Effects of age, hearing acuity, verbal ability, and cognitive function. *Ear and Hearing, 33,* 250–256.

Benki, J. R. (2003). Quantitative evaluation of context effects in spoken word recognition. *The Journal of the Acoustical Society of America, 113,* 1689–1705.

Boothroyd, A. (1968a). Developments in speech audiometry. *Sound, 2,* 3–10.

Boothroyd, A. (1968b). Statistical theory of the speech discrimination score. *The Journal of the Acoustical Society of America, 43,* 362–367.

Boothroyd, A. (1993). Speech perception, sensorineural hearing loss, and hearing aids. In G. A. Studebaker & I. Hochberg (Eds.), *Acoustical factors affecting hearing aid performance* (pp. 277–299). Boston, MA: Allyn & Bacon.

Boothroyd, A. (2006). Computer-Assisted Speech Perception Assessment (Version 5.0) [Computer software]. San Diego, CA: Author.

Boothroyd, A. (2008). The performance/intensity function: An underused resource. *Ear and Hearing, 29,* 479–491.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America, 84,* 101–114.

Bronkhorst, A. W., Bosman, A. J., & Smoorenburg, G. F. (1993). A model for context effects in speech recognition. *The Journal of the Acoustical Society of America, 93,* 499–509.

Bronkhorst, A. W., Brand, T., & Wagener, K. (2002). Evaluation of context effects in sentence recognition. *The Journal of the Acoustical Society of America, 111,* 2874–2886.

Carney, E., & Schlauch, R. S. (2007). Critical difference table for word recognition testing derived using computer simulation. *Journal of Speech, Language, and Hearing Research, 50,* 1203–1209.

Eisenberg, L. S., Shannon, R. V., Martinez, A. S., Wygonski, J., & Boothroyd, A. (2000). Speech recognition with reduced spectral cues as a function of age. *The Journal of the Acoustical Society of America, 107,* 2704–2710.

Felty, R. (2007). *Confusion patterns and response bias in spoken word recognition of German disyllabic words and nonwords.* Proceedings of the 16th International Congress of the Phonetic Sciences, 1957–1960. Available from http://www.icphs2007.de/conference/Papers/1294/1294.pdf

Gelfand, S. A. (1998). Optimizing the reliability of speech recognition scores. *Journal of Speech, Language, and Hearing Research, 41,* 1088–1102.

Gelfand, S. A. (2003). Tri-word presentations with phonemic scoring for practical high-reliability speech recognition assessment. *Journal of Speech, Language, and Hearing Research, 46,* 405–412.

Gelfand, S. A., & Gelfand, J. T. (2012). Psychometric functions for shortened administrations of a speech recognition approach using tri-word presentations and phonemic scoring. *Journal of Speech, Language, and Hearing Research, 55,* 879–891.

Groen, J. J., & Helleman, A. C. (1960). Binaural speech audiometry. *Acta Otolaryngologica, 52,* 397–414.

Guthrie, L. A., & Mackersie, C. L. (2009). A comparison of presentation levels to maximize word recognition scores. *Journal of the American Academy of Audiology, 20,* 381–390.

Haagen, C. H. (1945). *Intelligibility measurement: Twenty-four multiple-choice tests* (Report No. 5567, U.S. Dept. of Commerce PB 12050). Washington, DC: U.S. Department of Commerce.

Hagerman, B. (1976). Reliability in the determination of speech discrimination. *Scandinavian Audiology, 5,* 219–228.

Harris, J. D. (1980). On the use of a 3-words-per-item format in tests for the hearing of speech. *The Journal of the Acoustical Society of America, 67,* 345–347.

Haskins, H. A. (1949). *A phonetically balanced test of speech discrimination for children* (Unpublished master's thesis). Northwestern University, Evanston, IL.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *Journal of Speech and Hearing Disorders, 17,* 321–337.

Keidser, G. (1993). Normative data in quiet and in noise for "DANTALE"—A Danish speech material. *Scandinavian Audiology, 22,* 231–236.

Mackersie, C. L., Boothroyd, A., & Minniear, D. (2001). Evaluation of the Computer-Assisted Speech Perception Assessment Test (CASPA). *Journal of the American Academy of Audiology, 12,* 390–396.

Markides, A. (1978). Whole-word scoring versus phoneme scoring in speech audiometry. *British Journal of Audiology, 12,* 40–46.

McCreery, R., Ito, R., Spratford, M., Lewis, D., Hoover, B., & Stelmachowicz, P. G. (2010). Performance-intensity functions for normal-hearing adults and children using computer-aided speech perception assessment. *Ear and Hearing, 31,* 95–101.

Nittrouer, S., & Boothroyd, A. (1990). Context effects in phoneme and word recognition of young children and older adults. *The Journal of the Acoustical Society of America, 87,* 2705–2715.

Olsen, W. O. (1983). Research trends and clinical needs in speech audiometry. In D. F. Konkle & W. F. Rintlemann (Eds.), *Principles of speech audiometry* (pp. 375–395). Baltimore, MD: University Park Press.

Olsen, W. O., Van Tasell, D. J., & Speaks, C. E. (1997). Phoneme and word recognition for words in isolation and in sentences. *Ear and Hearing, 18,* 175–188.

Raffin, M. J. M., & Schafer, D. (1980). Application of a probability model based on the binomial distribution to speech-discrimination scores. *Journal of Speech and Hearing Research, 23,* 570–575.

Raffin, M. J. M., & Thornton, A. (1980). Confidence levels for differences between speech discrimination scores. *Journal of Speech and Hearing Research, 23,* 5–18.

Sergeant, L., Atkinson, J. E., & LaCroix, P. G. (1979). The NSMRL Tri-Word Test of Intelligibility. *The Journal of the Acoustical Society of America, 65,* 218–222.

Thornton, A., & Raffin, M. J. M. (1978). Speech discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research, 21,* 507–518.

Tillman, T. W., & Carhart, R. (1966). *An expanded test for speech discrimination utilizing CVC monosyllabic words* (Northwestern University Auditory Test No. 6, Tech Report SAM-TR-66-55). San Antonio, TX: United States Air Force, School of Aerospace Medicine.

Watson, N. A., & Knudsen, V. O. (1940). Selective amplification in hearing aids. *The Journal of the Acoustical Society of America, 11,* 406–419.

Williams, C. E., Mosko, J. D., & Greene, J. W. (1976). Evaluating the ability of aircrew personnel to hear speech in their operational environments. *Aviation, Space and Environmental Medicine, 47,* 154–158.

Wilson, H., Burks, C. A., & Weakley, D. G. (2005). A comparison of word-recognition abilities assessed with digit pairs and digit triplets in multitalker babble. *Journal of Rehabilitation Research and Development, 42,* 499–510.

Wilson, H., & Weakley, D. G. (2004). The use of digit triplets to evaluate word-recognition abilities in multitalker babble. *Seminars in Hearing, 25,* 93–111.

Woods, D. L., Yund, E. W., & Herron, T. J. (2010). Measuring consonant identification in nonsense syllables, words, and sentences. *Journal of Rehabilitation Research and Development, 47,* 243–260.