J Am Acad Audiol 11: 489–493 (2000)

The Question of Phonetic Balance in Word Recognition Testing

Frederick N. Martin*
Craig A. Champlin*
Desirée D. Perez*

Abstract

Twenty subjects with normal hearing and 15 subjects with mild-to-moderate sensorineural hearing losses were tested with eight lists of words using monosyllabic pronunciation to determine word recognition scores. Four of the lists were taken from Northwestern University Test No. 6 and four were simply made up by randomly selecting words from a dictionary. All of the word lists were used to determine performance-intensity functions. No clinically meaningful differences were observed among the lists.

Key Words: Phonemically balanced word list, phonetically balanced word list, word recognition score

Abbreviations: ANOVA = analysis of variance, NU-6 = Northwestern University Auditory Test No. 6, PB = phonemically/phonetically balanced, SRT = speech recognition threshold, WRS = word recognition score

Word recognition testing is an invaluable diagnostic tool. A comprehensive audiologic evaluation is often considered incomplete without the use of speech stimuli. Word recognition scores (WRSs) may provide valuable information regarding auditory performance in real-world situations (Hirsh et al, 1952). Word recognition testing is routinely used by clinical audiologists to aid in the selection and evaluation of appropriate amplification, to determine site of lesion, to assess specific rehabilitative needs, and to assess central auditory function (Bess, 1983).

Following World War II, word recognition lists were developed specifically for assessing speech discrimination ability. Because the creators of the lists believed that a discrimination test should mimic conversational speech, the construction of the word lists was based on the phonetic composition of English (Eldert and Davis, 1951). Each list was intended to contain the phonetic elements in approximately the same proportion that they occur in English (Eldert and Davis, 1951).

Although the idea of developing word lists that mimic conversational English was commendable, researchers have found it impossible to create lists of words that are truly phonetically balanced (PB). Although it is possible to approximate the frequency of occurrence of sounds in "average" speech, the actual distribution of sounds in speech depends on the topic being discussed and who is speaking (Egan, 1948). A speech sound will vary depending on the sounds that surround it; therefore, it is impossible to have a list of words that is truly PB (Lehiste and Peterson, 1959). To make the lists more homogeneous and to attempt phonemic balance, Lehiste and Peterson, among others, developed lists of words that followed the consonant-nucleus-consonant model, where the nucleus is either a vowel or a diphthong. Word lists that aspire to either phonetic or phonemic balance are traditionally referred to as PB word lists.

Despite existing evidence disputing the accuracy of phonetic balance, the most widely used set of materials for word recognition testing continues to be PB word lists (Martin et al, 1998). The heavy reliance on these materials has led to the current study, the primary purpose of which was to ascertain the impact, if any, of

*Department of Communication Sciences and Disorders, The University of Texas at Austin, Austin, Texas

Reprint requests: Frederick N. Martin, Department of Communication Sciences and Disorders, The University of Texas at Austin, Austin, TX 78712

phonetic balance on word recognition scores. The main question asked in this study was whether the scores derived from PB word lists were comparable to scores derived from deliberately non-PB word lists.

METHOD

Subjects

This study was conducted using two subject groups. The first group consisted of 20 normal-hearing persons (3 males, 17 females) 19 to 31 years old. All subjects in this group passed a pure-tone screen at 15 dB HL at frequencies 250 to 8000 Hz and exhibited speech recognition thresholds (SRTs) no poorer than 15 dB HL (ANSI, 1996).

The second group of subjects consisted of 15 adults (7 males, 8 females) ranging in age from 52 to 83 years old with diagnosed sensorineural hearing loss. Audiometric criteria for inclusion in this group included an SRT between 30 and 50 dB HL.

Equipment/Materials

Subjects were tested in a sound-treated chamber using a Madsen Electronics OB822 audiometer. All testing was conducted using Ear Tone ER-3A insert earphones with standard-size EARLink foam eartips. Prior to data collection, all equipment was calibrated using ANSI (1996) specifications. Insert earphones were calibrated using a Bruel & Kjaer (type 2203) sound level meter. Throughout the testing period, calibration checks were randomly administered.

Each subject was tested using two different sets of word recognition lists. The first set was comprised of conventional PB word lists selected from Northwestern University Auditory Test No. 6 (NU-6) (Tillman et al, 1966). The second set of lists was developed for this research and consisted of 200 randomly selected words that lent themselves to monosyllabic utterance. The words were chosen from *Webster's Ninth New Collegiate Dictionary* by chance selection of a page number and choosing the first word on each page that could be pronounced monosyllabically. These randomly selected words were then divided into four 50-word lists. In total, there were eight lists, four new lists and four lists (1A, 2A, 3A, and 4A) from the NU-6 test.

The word lists were recorded on a compact disc by an experienced female speaker. Each word was preceded by the carrier phrase "Say the word. . ." with a 4-sec interval between test items. A VU meter was used to monitor the level of speech such that the last word of the carrier phrase peaked at 0 dB VU.

Procedure

SRTs were determined for each subject using spondaic words delivered through an insert earphone. Each subject was then presented with four conventional and four experimental word lists (Appendix). To help minimize order effects, test ear and list type were counterbalanced and sensation levels (10, 20, 40, and 50 dB) were randomized. All responses were scored as either correct or incorrect.

RESULTS

In Figure 1, the mean WRS is plotted at three presentation levels. The open symbols represent the scores for the normal-hearing listeners, whereas the filled symbols indicate the scores for the hearing-impaired listeners. The triangles and circles denote the scores on NU-6 and experimental word lists, respectively. As expected, the WRS improved with level. Also not surprising is the finding that normal-hearing listeners had higher scores than did listeners with hearing impairment. Although the normal-hearing listeners performed comparably with both lists, the hearing-impaired listeners

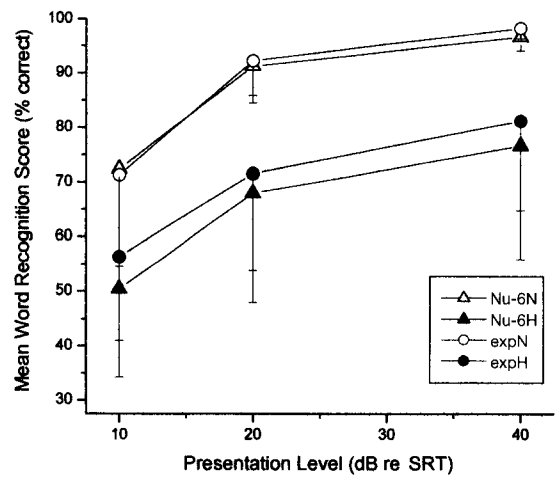

Figure 1 Mean word recognition score by presentation level (Nu-6N = NU-6 list, normal-hearing listeners; Nu-6H = NU-6 list, hearing-impaired listeners; expN = experimental list, normal-hearing listeners; expH = experimental list, hearing-impaired listeners). Error flags show 1 SD.

scored somewhat higher with the experimental words.

Before performing statistical analyses on the data in Figure 1, the scores (percent) were converted to rationalized arcsine units as described by Studebaker (1985). Table 1 is a summary of a 2 × 2 × 3 (Groups × List × Level) mixed-design analysis of variance (ANOVA) with repeated measures. The three main effects are significant at the .01 level. None of the interactions is significant.

Although the ANOVA revealed a significant list effect, an examination of the difference scores reveals that the effect is small. The WRS obtained with the NU-6 list is about 2 percent lower (median difference) than the one measured with the experimental words.

Figure 2 is a scatter plot of the scores achieved with each word list by the 35 listeners. Linear regression analysis shows that the data are well described by a line with a slope close to unity (1.03). The r^2 value of 90 percent indicates that the score for the experimental word list is a good predictor of the NU-6 word score.

Although Ross and Huntington (1962) found that the four Central Institute for the Deaf W-22 lists were not equivalent, a search of the literature failed to reveal any such equivalency studies for the NU-6 word lists. No attempt was made to determine whether the experimental word lists used in the present study were equivalent. However, results of the study indicate that the four experimental word lists yielded similar mean scores.

Normal-hearing subjects demonstrate small test–retest variability on WRSs (Engelberg, 1968). The test–retest variability among patients

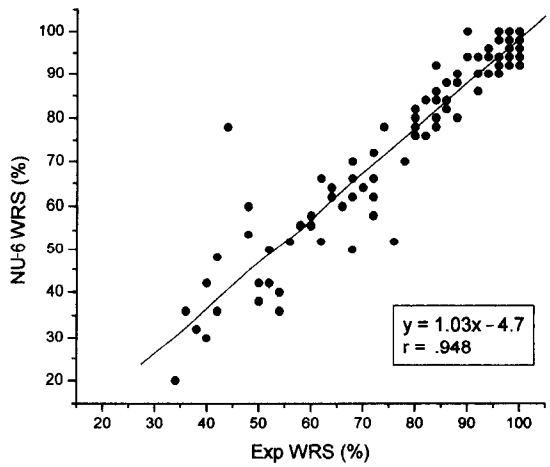

Figure 2 Scatter plot of word recognition scores (WRSs) obtained with NU-6 versus experimental (Exp) words. Inset summarizes linear regression analysis.

with sensorineural hearing loss depends largely on the number of test items and the true score for that test (Thornton and Raffin, 1978). Although an ANOVA did reveal significant differences between the conventional and experimental word lists for the sensorineural hearing-impaired group, test scores showed no significant differences based on the lower and upper limits of the 95 percent critical differences for percentage scores according to Thornton and Raffin (1978).

DISCUSSION

The concept of PB monosyllabic word lists has been integral to audiologic assessment for

Table 1 Analysis of Variance Summary

Source	*df*	*SS*	*MS*	*F*	*p*
sbj	33	37,268.2	1,129.3		
grp	1	29,040.6	29,040.6	25.7	<.0001
lvl	2	36,839.1	18,419.6	140.0	.0001
sbj*lvl	66	8,681.7	131.5		
grp*lvl	2	663.8	331.9	2.5	.0879
lst	1	604.7	604.7	17.7	.0002
sbj*lst	33	1,126.5	34.1		
grp*lst	1	116.8	116.8	3.4	.0733
lvl*lst	2	75.6	37.8	1.1	.3485
sbj*lvl*lst	66	2,329.1	35.3		
grp*lvl*lst	2	101.9	50.9	1.4	.2435
Error	0	0.0			
Total	209	118,905.0			

lvl = presentation level, lst = list type, grp = subject group, sbj = individual subjects.

more than 50 years. Indeed, it has evolved its own mystique. For more than a decade, for example, there was spirited discussion in the literature on the advantages (e.g., Elpern, 1961; Resnick, 1962) versus the hazards (e.g., Grubb, 1963a, b) of using only 25 words per list. Is the saving in time countered by loss of phonetic balance? But the present results show that, even when 50 words are used, the total score based on randomly selected words is not substantially different from the total score based on carefully selected, PB word lists.

Although many factors influence WRSs, including presentation level, word use frequency and familiarity (Pisoni, 1985), dialect of the speaker, and lexical neighborhood (Luce, 1986), phonetic balance does not appear to be one of them. There may be good reasons for using particular PB word lists based on their standardization and known properties, but the notion that the score uniquely reflects recognition of the sounds of conversational speech in their relative frequency of occurrence is, perhaps, illusory.

REFERENCES

American National Standards Institute. (1996). *American National Standard Specification for Audiometers.* (ANSI S3.6-1996). New York: ANSI.

Bess FH. (1983). Clinical assessment of speech recognition. In: Konkle D, Rintelmann W, eds. *Principles of Speech Audiometry* Baltimore: University Park Press, 127–202.

Egan J. (1948). Articulation testing methods. *Laryngoscope* 58:955–991.

Eldert E, Davis H. (1951). The articulation function of patients with conductive deafness. *Laryngoscope* 61:891–909.

Elpern BS. (1961). The relative stability of half-list and full-list discrimination tests. *Laryngoscope* 71:30–35.

Engelberg M. (1968). Test-retest variability in speech discrimination testing. *Laryngoscope* 78:1582–1589.

Grubb P. (1963a). Phoneme analysis of half-list speech discrimination tests. *J Speech Hear Res* 6:271–275.

Grubb P. (1963b). Considerations in the use of half-list speech discrimination tests. *J Speech Hear* Res 6:294–297.

Hirsh I, Davis H, Silverman SR, Reynolds E, Eldert E, Benson R. (1952). Development of materials for speech audiometry. *J Speech Hear Disord* 17:321–337.

Lehiste I, Peterson G. (1959). Linguistic considerations and intelligibility. *J Acoust Soc Am* 31:280–286.

Luce PA. (1986). A computational analysis of uniqueness points in auditory word recognition. *Percept Psychophys* 39:155–159.

Martin FN, Champlin CA, Chambers JA. (1998). Seventh survey of audiometric practices in the United States. *J Am Acad Audiol* 9:95–104.

Pisoni D. (1985). Speech perception. Some new directions in research and theory. *J Acoust Soc Am* 78:381–388.

Resnick D. (1962). Reliability of the twenty-five word phonetically balanced lists. *J Auditory Res* 2:5–12.

Ross M, Huntington D. (1962). Concerning the reliability and equivalency of the CID W-22 auditory tests. *J Auditory Res* 2:220–228.

Studebaker GA. (1985). A "rationalized" arc sine transformation. *J Speech Hear Res* 28:455–462.

Thornton A, Raffin M. (1978). Speech discrimination scores modeled as a binomial variable. *J Speech Hear Res* 21:507–518.

Tillman TW, Carhart R, Wilber L. (1963). *An Expanded Test for Speech Discrimination Utilizing CNC Monosyllabic Words. Northwestern University Auditory Test No. 6.* Technical Report, SAM-TR-66-55. Brooks Air Force Base, TX: USAF School of Aerospace Medicine, Aerospace Medical Division (AFSC).

APPENDIX

Experimental Word Lists

List A		List B		List C		List D	
for	hall	pill	jug	black	gate	ripe	sick
lie	crop	ban	lip	milk	cold	gear	feed
cart	tone	seat	dive	rock	first	shop	look
mint	verse	draw	flesh	toe	club	toll	print
cry	rank	ant	cat	rank	rule	heal	ring
gel	cute	gate	ham	blind	hunt	ink	bread
smoke	ace	plant	heat	fix	lamp	key	year
flag	wind	robe	off	man	phone	cut	skirt
beat	bin	bank	vent	play	keep	bill	work
peg	dump	east	sharp	reach	back	way	boat
hat	price	mail	clock	shoe	dot	shy	trick
last	root	right	kind	tax	bed	head	line
bash	seam	time	worm	will	side	dwell	green
four	wood	vase	zone	wait	box	leave	long
mush	void	slope	ball	flow	run	hat	smooth
boat	ear	old	tell	look	set	point	cap
rest	flap	pear	quick	dorm	worn	sold	stone
orange	here	ask	inch	dust	why	coat	trade
lump	glaze	noise	mode	heart	born	hate	hold
home	hand	poor	shark	grow	light	room	rag
choke	drew	set	crew	note	size	less	red
mind	land	high	best	learn	net	rail	close
pan	match	brain	hack	proud	call	rope	park
roll	bag	waste	ice	hang	wind	top	show
blank	fish	proud	pull	band	seize	sheet	pin