
Development of Materials for the Clinical Assessment of Speech Recognition: The Speech Sound Pattern Discrimination Test

Joseph H. Bochner
Professional Assessment and
Information Services and
National Technical Institute
for the Deaf, Rochester, NY

Wayne M. Garrison
Professional Assessment and
Information Services
Baltimore, MD

Joan E. Sussman
Robert F. Burkard
University at Buffalo
Buffalo, NY

The purpose of this investigation was to evaluate the validity and reliability of materials designed for an assessment procedure capable of making meaningful distinctions in speech recognition ability among individuals having mild-to-moderate hearing losses. Sets of phonetic contrasts were presented within sentence contexts to 53 listeners (22 normal hearing, 31 hearing impaired) in 4 listening conditions (quiet and with background competition at signal-to-noise ratios of +5, 0, and -5 dB). The listeners were asked to discriminate pairs of sentences (e.g., "The man hid the dog" and "The man hit the dog") using *same-different* judgments. Their performances were analyzed in a manner enabling comparisons among items in terms of the classification of phonetic contrasts. Listener performance was also compared to performance on a set of independent variables, including the W-22 and QuickSIN speech tests, high-frequency hearing loss, speech reception threshold, listener age, and others. Results indicated that the new materials distinguished the normal-hearing from the hearing-impaired group and that listener performance (a) declined about 17% for each 5 dB decrement in SNR and (b) was influenced by the phonetic content of items in a manner similar to that reported by G. A. Miller and P. E. Nicely (1955). The performances of the hearing-impaired listeners were much more strongly related to high-frequency hearing loss, listener age, and other variables than were their performances on either the W-22 or QuickSIN tests. These findings are discussed with specific reference to the use of a mathematical model (i.e., the Rasch model for person measurement) for scaling items along a continuum of difficulty. The mathematical model and associated item difficulty values will serve as the basis for construction of a clinically useful computerized, adaptive test of speech recognition ability known as the Speech Sound Pattern Discrimination Test (Bochner, J., Garrison, W., Palmer, L., MacKenzie, D., & Braveman, A., 1997).

KEY WORDS: speech recognition testing, speech audiometry, hearing loss, adaptive testing, software

Current methods, materials, and approaches to the clinical assessment of speech recognition tend to suffer from limitations in reliability and validity (Gelfand, 1998, 2001; Mendel & Danhauer, 1997). The tests most frequently used in clinical practice are phonetically balanced lists of monosyllabic words (e.g., PB-50 tests such as Central Institute for the Deaf (CID) Auditory Test W-22 and NU-6, the origins of which date back approximately 50 years; Martin, Armstrong, & Champlin,

1994). The limitations of PB-50 tests are well known, especially with respect to their reliability and measurement error (Thornton & Raffin, 1978). Because diminished measurement precision can have an adverse impact on the quality of audiological services (e.g., for differentiating among hearing aids or for determining the degree of communication impairment), a strong case can be made that current practices used for the clinical assessment of speech recognition are in need of improvement.

Given the history and limitations of speech perception tests currently available (see Mendel & Danhauer, 1997, for a review), new approaches adapted to the characteristics of individuals receiving audiological services may offer significant opportunities for improvement in the measurement of speech recognition abilities (see Tyler, 1994). In this regard, the introduction of computer-controlled adaptive test-delivery systems into the field of audiology is likely to be advantageous. Adaptive-testing procedures are not new to the field. For example, adaptive procedures have been used in auditory research (e.g., Elliott, Busse, Partridge, Rupert, & DeGraaff, 1986; Levitt, 1971) and have also been proposed as clinical tools for determining speech reception thresholds (Laitakari, 1996; Plomp & Mimpen, 1979). Similarly, an adaptive procedure is sometimes used in administering the Speech in Noise (SPIN) test (e.g., Frisina & Frisina, 1997). Although some adaptive procedures are familiar to audiologists, a different kind of adaptive test is envisioned in this study. Unlike adaptive procedures that adjust the magnitude of a stimulus attribute along a physical dimension (e.g., intensity), the adaptive procedure envisioned in this study entails adjusting stimulus content along a dimension of human performance that we associate with speech processing ability. As such, this study focuses on the development and scaling of a set of stimuli that will form the basis for a computerized, adaptive speech recognition test.

Modern approaches to adaptive testing have developed within a mathematical framework known as item response theory. Item response theory models are mathematical abstractions that can be applied to the analysis of data and used in the scaling and selection of items in adaptive testing. The Rasch model of person measurement (Rasch, 1960/1980) is the simplest item response theory model, and it is used in the present study. Although the use of item response theory models is now fairly common in educational and psychological measurement, its use in this study represents one of the first applications of item response theory methods in the communication sciences and audiology.

Item response theory generally refers to three probabilistic measurement models of increasing complexity. They are the one-parameter (Rasch, 1960/1980), the two-parameter, and the three-parameter models, named by

the number of item parameters estimated by each (i.e., item difficulty, discrimination, guessing). On first evaluation, this would seem to imply that the Rasch model is merely a scaled-down version of more complex models that must be superior because they account for more of the presumed reality of test-taking behavior.

Test items do differ in their ability to differentiate the performance of individual test takers, as well as in their ability to encourage guessing behavior. However, it can be demonstrated algebraically that these parameters must diverge (i.e., making them inestimable) unless they are constrained in an arbitrary manner, as occurs in practice. The Rasch model is not intended to fit data, as is the case with more complicated models. Rather, the Rasch model is a definition of measurement. When data are found to fit the model (not model found to fit the data), the measurement of persons and the calibration of items enable us to place persons and items on a common scale that functions according to the rules of arithmetic (Shaw, 1991). It is important to emphasize that item difficulty is an intrinsic parameter of item response theory models and that variations in item difficulty are necessary for the creation of item pools used in adaptive testing.

The purpose of this investigation was to evaluate the validity and reliability of materials designed for an assessment procedure capable of making meaningful distinctions in speech processing ability among individuals having mild-to-moderate hearing losses. Such distinctions have relevance in clinical practice. A combination of statistical methods was used to study validity and reliability from multiple perspectives. The study was intended to provide the foundation for the construction of a computerized, adaptive-testing system for the clinical assessment of speech recognition. Specifically, we sought to enhance the difficulty and efficiency of the Speech Sound Pattern Discrimination Test (SSPDT; Bochner, Garrison, Palmer, MacKenzie, & Braveman, 1997), and demonstrate its capabilities for use with listeners having mild-to-moderate sensorineural hearing losses. Together with the results of preliminary studies conducted on listeners having severe-to-profound hearing losses (Bochner et al., 1997), a demonstration of the efficacy of the procedure with listeners having mild-to-moderate hearing losses would constitute strong evidence that the SSPDT has the potential to become an effective and practical means of assessing speech recognition across the full range of hearing loss seen in clinical practice. Such a demonstration would support the additional work needed to develop a clinical instrument for speech recognition measurement utilizing computerized, adaptive-testing technology to obtain reliable and valid results with as few as 15 items.

Method

Participants

Fifty-three participants (22 male, 31 female) took part in this study. Their average age was 45.1 years ($SD = 22.2$). Twenty-two participants were individuals with normal hearing (predominately college students) and 31 were individuals with sensorineural hearing losses. Each participant was given an audiological evaluation at the Speech and Hearing Clinic at the University at Buffalo. Air-conduction thresholds were obtained for the octave frequencies from 0.25 to 8 kHz. Half-octave frequency and bone-conduction thresholds were assessed as needed clinically, and tympanograms were obtained for each ear. Speech reception thresholds were measured via live voice using CID W-1 and W-2 spondaic word lists, and word recognition was assessed with recorded W-22 word lists (Auditec Revised Auditory Tests CD). Finally, a measure of speech recognition in noise and a self-report scale of hearing handicap were administered. Specifically, the QuickSIN Speech-in-Noise Test (Etymotic Research, 2001) was administered by presenting the stimuli at 70 dB HL, and participants responded to Form A of the Hearing Handicap Scale (High, Fairbanks, & Glorig, 1964).

All participants showed a peak in the tympanogram between -200 and $+200$ daPa of ambient pressure in the test ear. No tympanogram amplitude criterion was used. Normal hearing sensitivity was defined by air-conduction thresholds ≤ 20 dB HL at all test frequencies. Average hearing losses falling in the 25–40 dB HL range for pure tones at 0.5, 1, and 2 kHz were considered mild, while those falling in the 41–55 dB HL range were considered moderate. Most participants showed a flat loss or a more severe loss in the higher frequencies. One participant showed normal hearing sensitivity at all frequencies except 500 and 1000 Hz in the test ear, and another showed normal hearing thresholds except at 1000 Hz in the test ear. With one exception, all hearing-impaired participants had hearing losses in the mild-to-moderate range. This one participant had a pure-tone average (PTA) of 63 dB HL in the right (test) ear, with the hearing loss extending into the moderate-to-severe range.

Stimuli

Each item comprised one standard and two comparison sentences. Examinees were asked to determine whether each of the comparison sentences was the *same* or *different* from the standard. All *same* trials involved repetition of the standard stimulus utterance token. The stimuli consisted of 274 items. Of the 274 items, 250 contained phonetic contrasts intended to elicit meaningful

information concerning listeners' speech processing abilities (e.g., information concerning the ability to process phonetic features of place, manner, and voicing as described below). These items, by definition, included at least one *different* trial because they each contained a phonetic contrast. The remaining 24 items, however, were foils comprising two *same* trials. The foils were intended to demonstrate to listeners that instances of two *same* trials actually were included in the stimulus array so that they had tangible evidence that all possible combinations of *same* and *different* trials were being presented. In other words, the inclusion of foils was intended to avert or at least minimize the potential for response bias. Foils were not included in scoring the SSPDT.

The 250 items containing contrasts differed in phonetic properties. Specifically, consonant contrasts were classified according to phonetic features of place, manner, and voicing, while vowel contrasts were classified according to features of tongue position/advancement, tongue height, and tenseness/length. In addition, some items contained prosodic contrasts consisting primarily of differences in utterance length. Four categories of items were developed based on these distinctions. Previous research (Bochner et al., 1997) suggested that these categories of items, in general, comprised a hierarchy of difficulty similar to the hierarchy demonstrated by Miller and Nicely (1955; also see Peterson & Barney, 1952). Sample items from each category appear in Table 1, with *same* trials (i.e., repetitions of the standard stimulus) indicated in italics.

The category of items hypothesized to be most difficult, Category Z, comprised contrasts in place of articulation for consonants and tongue position/advancement for vowels (see 1–3 in Table 1). The category of items hypothesized to be the second most difficult, Category Y, comprised contrasts in manner of articulation for consonants and tongue height for vowels (see 4–6 in Table 1). The category of items hypothesized to be the third most difficult, Category X, comprised contrasts in consonant voicing and vowel length/tenseness (see 7–8 in Table 1). The category of items hypothesized to be the easiest, Category W, included time-intensity variations in the envelope of the speech waveform and multiple segmental contrasts that could extend across syllable and word boundaries (see 9–10 in Table 1).

The distribution of items according to category membership and phonetic properties is displayed in Table 2. Inspection of the table reveals that the distribution of items is not balanced across or within categories. The test materials were purposely constructed in this manner to enhance the difficulty of the listening task. Because the purpose of the present study was to develop an assessment procedure capable of making meaningful

Table 1. Sample items from Categories Z, Y, X, and W.

Category Z

1. Free books are available. (standard)
 - a. *Free books are available.* (comparison)
 - b. Three books are available. (comparison)
2. The soldiers spared the children. (standard)
 - a. The soldiers scared the children. (comparison)
 - b. The soldiers scared the children. (comparison)
3. The committee is working to revise the proposal. (standard)
 - a. The committee is working to revive the proposal. (comparison)
 - b. The committee is working to revive the proposal. (comparison)

Category Y

4. Many tourists visit the old fort every summer. (standard)
 - a. Many tourists visit the old part every summer. (comparison)
 - b. Many tourists visit the old port every summer. (comparison)
5. Did you see those new shows on TV? (standard)
 - a. *Did you see those new shows on TV?* (comparison)
 - b. Did you see those new shoes on TV? (comparison)
6. Don't pass the bread! (standard)
 - a. *Don't pass the bread!* (comparison)
 - b. Don't pat the bread! (comparison)

Category X

7. The workers painted a wide stripe on the street. (standard)
 - a. The workers painted a white stripe on the street. (comparison)
 - b. The workers painted a white stripe on the street. (comparison)
8. Tyson beat his opponent. (standard)
 - a. Tyson bit his opponent. (comparison)
 - b. Tyson bit his opponent. (comparison)

Category W

9. The schools closed yesterday. (standard)
 - a. The schools closed early. (comparison)
 - b. *The schools closed yesterday.* (comparison)
10. The accident was caused by poor visibility. (standard)
 - a. *The accident was caused by poor visibility.* (comparison)
 - b. The serious accident was caused by poor visibility. (comparison)

distinctions in speech processing ability among individuals having mild-to-moderate hearing losses, a balanced distribution of items would likely have been too easy (see Bochner et al., 1997). Accordingly, the materials contain many more contrasts involving consonants than vowels, many more contrasts involving place and manner of articulation than voicing and/or prosody, and many more contrasts involving voiceless consonants than voiced consonants.

The location of contrasts according to their serial position within words and sentences is shown in Table 3. Two explanations are in order with respect to this table. First, items in Category W are not included among the within-word locations, because these items are characterized by multiple segmental contrasts involving entire syllables and words. As such, contrasts in these items

Table 2. Distribution of items by category membership and phonetic properties. Bold entries denote item categories, and *italic* entries denote categories of phonetic properties.

Classification	Voiced	Voiceless	Total
Category Z			110
<i>Place</i>			92
Fricative	9	44	53
Stop	9	30	39
<i>Position/advancement</i>			18
Category Y			71
<i>Manner</i>			60
Labial	6	23	29
Alveolar	5	26	31
<i>Height</i>			11
Category X			39
<i>Voicing</i>			26
Fricative			13
Stop			13
<i>Tenseness/length</i>			13
Category W			30
<i>One syllable</i>			6
<i>Two or more syllables</i>			24
Total			250

cannot be located within words. Second, while the terms *initial*, *medial*, and *final* clearly denote locations within words, these terms require definition when referring to locations within sentences. Specifically, we define sentence-initial position as the first phrase in the sentence and sentence-final position as the last phrase in the sentence. Sentence-medial position, then, is defined as all other locations within the sentence. Finally, the full item set comprised 196 declarative sentences, 30 interrogatives (19 yes-no questions and 11 Wh- questions) and 24 imperatives.

An adult male having professional experience as an actor and radio announcer produced between three and nine tokens of each stimulus sentence. The stimuli were recorded in a double-walled IAC booth. Digital signal

Table 3. Distribution of items according to the location of contrasts as determined by the serial position of contrasts within words and sentences. Bold entries denote the linguistic domain of the contrasts.

	Initial	Medial	Final	Total
Within word	65	89	66	220
Category Z	27	53	30	110
Category Y	28	21	22	71
Category X	10	15	14	39
Within sentence	51	131	68	250

processing software (Milenkovic, 2000) was used for all recording and editing tasks. The utterances were digitized at a sampling frequency of 11.025 kHz and passed through an anti-aliasing (low-pass) filter with a cutoff frequency of 4906 Hz. The stimuli were presented via a Creative Labs Sound Blaster (SB Live!) card, housed in a Dell Optiplex GX100 Celeron computer, to an Etymotic ER-3A insert earphone. Stimuli were calibrated with a Larson Davis 800B sound level meter using a Bruel and Kjaer 1 in. condenser microphone coupled to the insert earphone via a 2-cc coupler. A representative token, usually one appearing in the middle of each set of utterances, was selected for editing. The selection and editing of utterances involved listening to the signals, visually inspecting waveforms and spectrograms, and considering measurements of signal duration and root mean square (RMS) amplitude. The waveforms and spectrograms were particularly useful in determining optimal points at which to excise sentences. After the sentences were excised, overall RMS voltages (and hence sound pressure levels) were equalized across utterances using the signal processing software. Specifically, the RMS voltage for each sentence was measured and then multiplied by a factor to achieve a constant RMS voltage of 0.5 V. In this way, each stimulus was presented at an approximately equivalent overall level. The maximum level of the sentences ranged from 68 to 74 dBC, in the fast meter mode. Although the overall level of each sentence was equivalent, the RMS voltage within sentences varied with oscillations in the amplitude of the waveform.

Stimuli were either presented in quiet or in background noise at three signal-to-noise ratios (SNRs) to make the perceptual task more challenging. Specifically, test items were presented at SNRs of +5, 0, and -5 dB, with multitalker babble serving as background competition. The background noise used in this study was dubbed from the master recording of 20-talker babble described by Frank and Craig (1984); a commercial version of this recording is available from Auditec.

The test materials consisted of sentences. Sentences were used in the construction of test materials because they are generally considered a closer reflection of actual communicative events than isolated words or phrases. For each item, the standard sentence was presented first and followed in succession by the two comparison sentences. The interstimulus interval following the standard stimulus was 1.5 s, and the response interval following each comparison stimulus was 5.0 s. An interval of 1.5 s occurred between items.

Procedure

Participants were tested in a double-walled IAC booth over the course of two or three sessions, each of

which typically lasted less than two hr. As described above, an additional session was devoted to the audio-logical evaluation. The SSPDT was presented to each listener's right ear to capitalize on any benefits that might be associated with the right ear advantage for speech processing tasks (Studdert-Kennedy & Shankweiler, 1970). For all participants, signals were presented at 70 dB SPL. A written representation of the standard stimulus appeared on a computer monitor throughout the presentation of the comparison stimuli to minimize the memory load associated with the discrimination task. A set of eight practice items was administered at the beginning of each testing session, and three practice items were administered when participants returned from a short break to begin a new listening condition. Participants indicated *same* and *different* responses by pressing a button within the response interval. If no response was produced within the designated interval, then the trial was scored incorrect and the next item was presented.

The experimental design called for the 250 test stimuli to be divided into two lists (A and B), each consisting of 125 representative items and 12 foils. The participant sample was also divided into two representative groups of listeners (Groups 1 and 2). Both lists (250 items and 24 foils) were presented to all participants in the quiet listening condition. The presentation of items in the other listening conditions was counterbalanced across participants such that Listener Group 1 was presented with Item List A in the +5 dB and -5 dB SNR listening conditions and List B in the 0 dB SNR condition. Likewise, Listener Group 2 was presented with Item List B in the +5 dB and -5 dB SNR listening conditions and List A in the 0 dB condition.

Data Analysis

An item response theory model was used to conduct detailed item analyses of the stimuli, and reliability and validity were evaluated from multiple perspectives in terms of established criteria. Specifically, SSPDT data were subjected to Rasch (1960/1980) scaling analysis. Underlying our analysis of these data is the supposition that items and respondents can be positioned simultaneously, in an orderly manner, along a dimension representing the variable(s) that they share.

The Rasch measurement model is one in a family of item response theory models. Item response theory models are mathematical abstractions that describe the probability of an observed response to a test item in terms of an individual's trait or ability level in combination with a set of constants associated with the test item itself. The Rasch model for dichotomously scored data conceptualizes the Person \times Item interaction in terms of only two parameters—respondent ability and item difficulty.

In the present study, the respondent parameter is the ability to process speech. This variable is assessed by the speech stimuli that define the SSPDT scale. Associated with each test item is a calibration value indicating the item's difficulty (i.e., the role that the item plays in the measurement process).

The responses to the SSPDT items yield item calibrations and person measures. Parameter estimation is accomplished through the method of unconditional maximum likelihood. Specifically, the estimated values of person ability and item difficulty yield the pattern of observed responses to the test stimuli that is the most likely. When a matrix of data conforms to the expectations of the Rasch model, there is indication that (a) the variable measured is unidimensional (i.e., a common factor explains the response to each item in a set) and (b) items and persons do not differ substantially with respect to response factors that are not represented in the model, such as a respondent's propensity for guessing. Unidimensionality does not necessarily imply that the construct assessed consists of only one component. Invariance of the item calibrations and person measures can also be demonstrated when the parameters are estimated with subsamples of data drawn from the same pool of items and population of respondents.

SSPDT scale data were analyzed for fit to the Rasch model using the program BIGSTEPS (Version 2.0) developed by Wright and Linacre (1991). Person measures and item calibrations, together with their standard errors, are expressed in logits (i.e., log-odds units). Fit statistics indicate the extent to which individual items conform with the core variable being assessed, as well as the extent to which responses by individuals to items are internally consistent. BIGSTEPS provides a person separation reliability index that indicates how well a set of items separates the persons tested. BIGSTEPS also provides an item separation reliability index that indicates how well the persons tested separate the items.

Results

Test Scoring and Analysis

The Rasch measurement model was applied to data collected in the quiet listening condition only. The analysis of fit of data to the model begins with the construction of a Persons \times Items response matrix. There were 53 persons (22 normal-hearing, 31 hearing-impaired respondents) and 250 items. The matrix consists of 0s and 1s, interval scores computed from responses to the test stimuli presented in quiet. The SSPDT protocol involved the presentation of a standard sentence followed in succession by two comparison sentences (trials). In our work, items were defined as blocks of utterances (a standard sentence and two trials), requiring a total of

two discrimination judgments. Each respondent's performance on individual items was scored "1" if both discrimination judgments were correct. Otherwise, performance on individual items was scored "0" (incorrect).

The usefulness of any psychophysical testing device can be gauged, in part, by its success in differentiating, or separating, the performances of the persons tested. In Rasch measurement practice, a person separation index has been developed to evaluate the efficiency of a set of items to separate respondents (Wright & Stone, 1991). Person separation reliability is comparable to the familiar Kuder-Richardson (K-R 20) measure of internal consistency, with values ranging from 0.0 to 1.0. The higher the value of the person separation reliability index, the better the differentiation among the persons tested and the clearer the internal structure of an item set. The person separation reliability index provides initial insight into the extent to which the items comprising a test are members of the same conceptual domain. The person separation reliability for the 250-item SSPDT was found to be .95, indicating that the items separated the 53 respondents very well, and that these items measure a single, dominant variable. On the basis of previous research (Bochner et al., 1997), we interpret this variable to represent speech processing ability.

An item separation reliability index (Wright & Stone, 1991) is also used in Rasch measurement practice to evaluate how well the persons tested differentiate items. This index, algebraically similar to the person separation reliability index, also ranges in value from 0.0 to 1.0. The item separation reliability for the 250-item SSPDT was found to be .77. This result indicates the calibration values of the test items spread out over a considerable range, attributing breadth and conceptual significance to the variable that they collectively define. The magnitude of the item separation reliability statistic reported in this study is attenuated because the range of speech recognition abilities in the participant sample was restricted (i.e., the participant sample primarily comprised individuals with normal-hearing sensitivity and hearing losses in the mild-to-moderate range). Inclusion of participants extending across a broader range of hearing loss would increase the magnitude of the item separation reliability statistic. The mean item difficulty value was 0.00 logit ($SD = 1.24$), with item difficulties spanning a range of -2.43 to $+4.68$ logits. Mean item difficulty expressed as proportion correct was .89 ($SD = .09$, minimum = .30, maximum = .98).

When the fit statistics for the 250 test items were examined, only 4 items were identified as misfitting. In the fit analysis, we evaluated the extent to which the observed responses to items (across persons) agreed with those predicted by the Rasch measurement model. From this analysis we concluded there is substantial

Table 4. Mean Speech Sound Pattern Discrimination Test performance for normal-hearing and hearing-impaired participants in quiet and noise listening conditions. Raw score means scaled from 0 to 125 are displayed, with standard deviations shown in parentheses.

Condition	Normal hearing	Hearing impaired	F
Quiet	120.8 (2.5)	102.9 (25.5)	10.65*
+5 dB SNR	104.7 (5.8)	81.2 (24.1)	20.16*
0 dB SNR	83.8 (5.8)	60.4 (22.6)	22.30*
-5 dB SNR	71.0 (11.3)	40.8 (24.2)	29.62*

Note. $df = 1, 51$ for each condition.

* $p < .01$.

conformity between observation (data) and prediction (model). In other words, the fact that 246 of the 250 items had fit statistics that did not deviate from model expectations demonstrates that the SSPDT data are well-fitted by the measurement model.

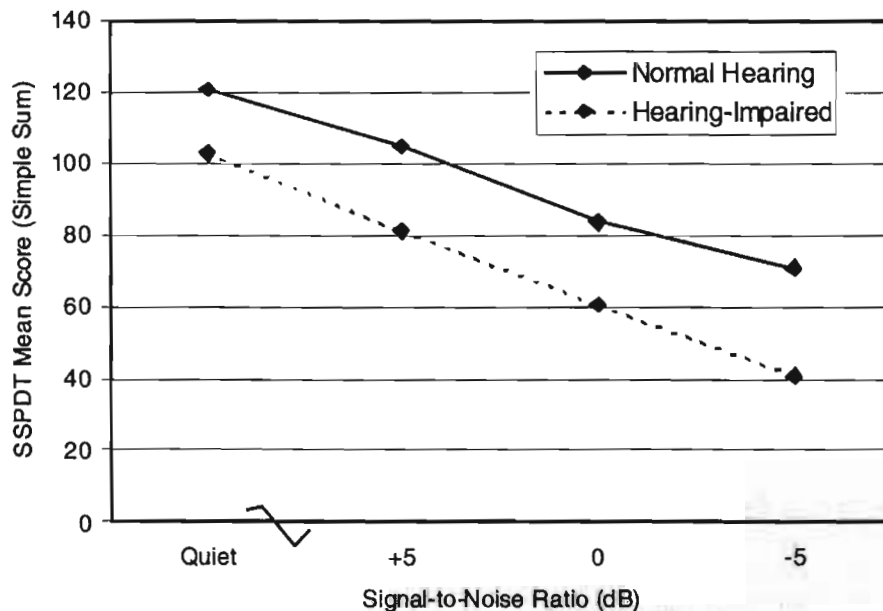
For the quiet listening condition, the mean score of all 53 participants on List A was 121.6 ($SD = 23.4$) and the mean score on List B was 122.0 ($SD = 20.7$). Equivalent forms reliability (correlation between scores on Lists A and B) was .96. Therefore, Lists A and B were equally difficult, with listeners performing similarly on each list.

The average performance on the SSPDT for all participants across the different listening conditions is summarized in Table 4. To simplify the presentation of

results across listening conditions, raw scores in the quiet listening condition were scaled from 0 to 125 by tallying each participant's performance across the 250 test items and multiplying the result by 0.5. This adjustment resulted in a common raw score scale extending from 0 to 125 across listening conditions. Had this adjustment not been made, comparisons across listening conditions would have been complicated by the fact that 250 items were administered to each listener in the quiet condition whereas 125 items were administered to each listener in the degraded conditions. The means in Table 4 are presented in this raw score metric, with standard deviations shown in parentheses. For each listening condition, the results of a one-way analysis of variance (ANOVA; with hearing status as a factor) indicated that the difference in performance between normal-hearing and hearing-impaired participants was statistically significant ($p < .01$). In each instance, the hearing-impaired participants performed more poorly than did the normal-hearing participants.

The information provided in the preceding table is presented in a graphical format in Figure 1. Performance on the SSPDT follows a uniform pattern, with the performance functions for the normal-hearing and hearing-impaired participants being roughly parallel, declining systematically with decreasing SNR. The decrement in test performance of the hearing-impaired participants across the testing conditions corresponds to a nearly constant increase in test difficulty (in the nonlinear proportion-correct or p metric) of about 17%. That is, a performance decrement of about 20 to 21 raw scale units

Figure 1. Mean SSPDT performance for normal-hearing and hearing-impaired participants across the quiet and degraded listening conditions. The y-axis represents the raw score means scaled from 0 to 125.



on the SSPDT corresponds to a change in average item difficulty of approximately .17 (proportion correct) across the +5 to -5 dB SNR conditions.

The mean difficulty of SSPDT items for hearing-impaired participants under the quiet, +5, 0, and -5 dB SNR listening conditions was .82, .65, .48, and .33 (*p* metric), respectively. Determining the optimum level and distribution of item difficulties in a test depends, in part, on the purpose of the test. Whenever the purpose of a test is to differentiate among respondents on the trait or ability measured, it is generally agreed that the "best" test is one composed of items of medium difficulty. Research evidence suggests that this design increases the variance of test scores and the reliability of the test (Tinkelman, 1971).

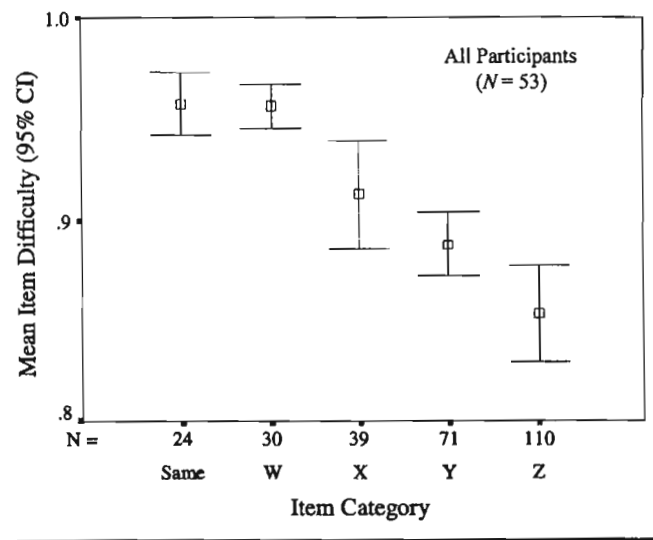
For choice-type items such as those characterizing the SSPDT (for which the possibility of guessing exists), an item of medium difficulty is defined as one on which the proportion of correct responses is midway between the expected chance proportion and 100% (Lord, 1952). By that logic, we can expect items with an average difficulty value of approximately .63 (midway between 25% and 100% in our research) to yield maximal information about the respondents (here, individuals with mild-to-moderate hearing loss). This would suggest that the quiet and +5 dB SNR testing conditions are most favorable. The 0 and -5 dB SNR conditions pose demands outside the capabilities of listeners with mild-to-moderate hearing loss, yielding data that are less informative of hearing for speech.

Item Content Validity

The analysis of item content is concerned with phonetic classification of the contrast occurring within each item. The test items have been classified in terms of four categories. As described earlier, each category is defined in terms of a specific set of phonetic properties. Three categories involve contrasts in a single sound segment occurring within one syllable, classifying item content in terms of phonetic features such as place, manner, and voicing. The fourth category involves multiple contrasts encompassing both segmental and nonsegmental features that can extend across syllable (and word) boundaries, classifying item content with reference to time-intensity variations in the speech waveform envelope and prosodic features. The four categories are ordered from most difficult (Category Z) to least difficult (Category W).

Trials involving replication of the standard stimulus (i.e., *same* trials), by definition, do not involve a phonetic contrast of any sort. In our previous research (Bochner, Garrison, & Palmer, 1992), *same* trials were found to be inherently easier than their counterparts.

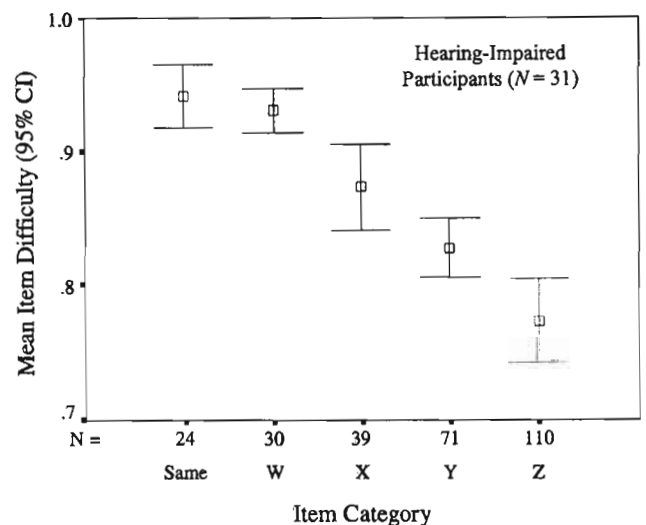
Figure 2. Mean difficulty across item categories for all participants. Error bars indicate 95% confidence intervals (CI).



In the present study, the stimulus array included 24 foils comprised of two *same* trials. The 24 foils have been included in the analysis of item content to study the generality of our earlier findings.

The relations between item placement into the categories described above and the difficulty values of the items are shown in Figures 2 and 3. The error bar plots in Figures 2 and 3 are those resulting from analyses of data derived from administration of the SSPDT in quiet. Thus, the item difficulty values are not confounded by the added difficulty brought about by the introduction

Figure 3. Mean difficulty across item categories for hearing-impaired participants. Error bars indicate 95% confidence intervals (CI).



of background noise into the testing paradigm. In Figure 2, the difficulty values of items are those determined from the responses of all 53 participants and are reported in the p metric. The error bar plots extending to either side of the mean item difficulty values for each of the categories shown in the figure symbolize a 95% confidence interval. Results of a one-way ANOVA (with item category as the factor) indicated that the difficulty values of the items differ significantly across content categories, $F(4, 269) = 11.2, p < .001$. As we observed in our earlier research (Bochner et al., 1992), the same trials continue to be the easiest. Scheffe tests were conducted to compare differences between all possible pairwise combinations of item categories. The results of this analysis indicated that 4 of the 10 possible combinations of item categories were significantly different ($p < .05$).

In Figure 3, we illustrate the error bar plots derived from the responses of the 31 hearing-impaired participants. Results of a one-way ANOVA (with item category as the factor) of these data corroborate the foregoing findings relating the categorization and difficulty of SSPDT items, $F(4, 269) = 17.7, p < .001$. Again, Scheffe tests were conducted to compare differences between all possible pairwise combinations of item categories. The results of these statistical tests indicated that 5 of the 10 possible combinations of item categories were significantly different ($p < .05$). In general, the results of the Scheffe tests conducted following each ANOVA indicated that items in Categories Z and Y were significantly more difficult than were items in the other categories. The general form of the function relating item categorization to difficulty reported in the current research replicates a finding reported in Bochner et al. (1997) and extends this earlier finding to listeners having mild-to-moderate hearing losses.

Hearing-impaired listeners "generally show considerably better perception for vowels than for consonants" (Revoile & Pickett, 1982, p. 33). Indeed, inspection of our item difficulty values indicated that items comprising vowel contrasts tended to be easier than those comprising consonant contrasts. These items exhibited a tendency to cluster in the range of difficulty observed for items classified in Category X. To account for this observation, items comprising vowel contrasts in Categories Z and Y were reclassified as members of Category X. As a result of this refinement in the classification scheme, Category X was redefined to include contrasts in voicing for consonants and tongue position/advancement, tongue height, and tenseness/length for vowels. Categories Z and Y, then, were redefined to include consonant contrasts in the phonetic features of place and manner, respectively. Refining the item classification scheme in this way increased the separation

among Categories X, Y, and Z for the full sample of 53 participants, as well as for the sample of 31 hearing-impaired participants. In each instance, the results of a one-way ANOVA (with item category as the factor) indicated that item difficulties differed significantly across all categories of item content, $F(4, 269) = 15.0, p < .001, n = 53; F(4, 269) = 24.4, p < .001, n = 31$. The results of Scheffe tests indicated that 5 of the 10 combinations of comparisons were significantly different for the full sample of 53 participants and that 6 of the 10 combinations of comparisons were significantly different for the sample of 31 hearing-impaired participants ($p < .05$). Again, the results of the Scheffe tests generally showed that items in Categories Z and Y were statistically more difficult than the other classes of items. The reclassification of items containing vowel contrasts, therefore, accurately accounts for empirical data observed in this investigation. This reclassification of the data is consistent with findings reported in studies reviewed by Revoile and Pickett (1982) and, as such, constitutes an improvement in the item categorization scheme.

Criterion-Related Validity

Performance on measures included in the audiological evaluations obtained from participants with normal-hearing sensitivity and from those with hearing loss are summarized in Table 5. Included in the table are means and standard deviations, as well as minimum and maximum values attained on the measures. All of the table entries are in the dB metric, with the exception of scores on the Hearing Handicap Scale (simple sum) and scores on the W-22 word lists (percentage correct).

Audiometric thresholds for the right (test ear) and left ears, including the PTA, high-frequency PTA (HFPTA), and speech reception threshold (SRT), were

Table 5. Audiological data summary. All table entries are presented in the dB HL metric (re: ANSI, 1996) with the exception of scores on the W-22 test (percentage correct) and on the Hearing Handicap Scale (simple sum).

Measurement	<i>M</i>	<i>SD</i>	Minimum	Maximum
Right ear PTA	11.54	14.60	-5	63.33
Right ear HFPTA	21.15	23.40	-6.67	75
Right ear SRT	13.11	14.29	-5	60
Left ear PTA	13.08	15.22	-5	66.67
Left ear HFPTA	21.96	24.10	-5	75
Left ear SRT	12.64	14.26	-5	60
W-22 dB HL	51.89	11.19	35	70
W-22 word lists (%)	93.58	14.87	12	100
QuickSIN	3.88	5.26	0.5	25.5
Hearing Handicap Scale (simple sum)	17.43	3.55	7	20

correlated for all 53 participants. PTA was based on 0.5, 1, and 2 kHz thresholds. HFPTA was based on 2, 4, and 8 kHz thresholds. These correlational analyses were performed to detect a hearing asymmetry between ears. The correlation between left and right ear PTA was .95 ($p < .01$), and for left and right ear HFPTA the correlation was .96 ($p < .01$). The correlation between left and right ear SRT was .92 ($p < .01$). These findings indicate there was not a strong hearing asymmetry in the sample of participants studied.

Simple correlations between the three measures of speech recognition were obtained (i.e., W-22, QuickSIN, and SSPDT). These tests differ in the materials they use, as well as in the manner in which they are scored. The SSPDT measure (quiet listening condition) correlated .67 ($p < .001$) with the W-22 measure and $-.69$ ($p < .001$) with the QuickSIN measure. The relations between the SSPDT measures and W-22 test performance declined across the degraded listening conditions, reaching a minimum Pearson r of .51 ($p < .001$) for the -5 dB SNR condition. Similarly, the relations between the SSPDT measures and QuickSIN test performance evidenced declines with decreasing SNR (Pearson $r = -.59$, $p < .001$ at -5 dB SNR). The correlation between W-22 and QuickSIN test performance was found to be $-.86$ ($p < .001$), indicating a large amount of shared variance between these measures.

Predictive Validity

We asked whether the information contained in the audiological evaluation might be organized in such a way as to confirm the clinical classification of the 53 participants into normal-hearing and hearing-impaired groups. This amounts to asking which audiological measures are most effective for estimating the hearing status of the respondents. To answer this question, we used a logistic regression procedure (stepwise variable selection method). As some of our hearing-impaired participants had a single threshold that was at 25 dB HL or above, we were surprised that 100% of the respondents were classified correctly into normal-hearing and

hearing-impaired groups using three predictor variables from the audiological data. These predictor variables were, in order of importance (a) HFPTA; (b) W-22 dB HL, the level at which W-22 was administered; and (c) simple sum score on the Hearing Handicap Scale, a self-report measure of hearing handicap. The results of the logistic regression analysis enabled us to create a categorical variable corresponding to membership in either the normal-hearing or hearing-impaired group (Group).

Construct Validity

Multiple linear regression analyses were performed to establish the relationships of the three speech measures (dependent variables) to the nonspeech measures (independent variables) obtained in the audiological evaluation. The results are summarized in Table 6. Also included in these analyses as independent variables were Group, (see above) and age of the respondent (Age). These analyses enable us to make direct comparisons among the speech measures modeled on the same set of independent variables. Again, a stepwise variable selection method was used in each instance.

The independent variables that were entered into a stepwise linear regression solution for the W-22, QuickSIN, and SSPDT (measures obtained under the four quiet and signal-to-noise conditions) are shown in Table 6. Also shown in the table is the multiple R , as well as the proportion of the total variance accounted for by linear combinations of the independent variables. For W-22, there is a simple relation between test performance and the HFPTA ($R^2 = .33$). No other independent measure satisfied the criterion for entry into the stepwise equation. High-frequency hearing loss is similarly associated with QuickSIN scores, with R^2 increasing to .45.

Four independent variables combined to produce a model of speech recognition ability in the quiet condition for the SSPDT and these variables accounted for 80% of the variance. In this model, the HFPTA is complemented by the SRT in the test ear. Advancing age (Age) and the HFPTA appear to be associated, consistent with the observations of Maurer and Rupp (1979) that hearing loss for pure tones increases both with chronological age and in the high frequencies. Group, a variable that includes a self-reported measure of hearing handicap, also entered the prediction equation.

The model of speech recognition ability fitted under the quiet listening condition generalized to the $+5$ dB SNR condition. Thereafter, as the listening conditions moved increasingly off-target from optimal difficulty, the SSPDT became less and less useful as a measure of hearing for speech. For the 0 dB SNR condition, HFPTA and

Table 6. Model comparisons among three measures of speech recognition.

Dependent variable	Predictors (stepwise entry)	Multiple R	R^2
W-22	HFPTA	.57	.33
QuickSIN	HFPTA	.67	.45
SSPDT (Quiet)	HFPTA, SRT, Age, Group	.89	.80
SSPDT (+5 dB SNR)	HFPTA, SRT, Age, Group	.90	.81
SSPDT (0 dB SNR)	HFPTA, Age	.86	.73
SSPDT (-5 dB SNR)	Age	.84	.71

Age are associated with SSPDT performance. However, the only predictor of SSPDT performance for the -5 dB condition is Age.

Discussion

The results of this study indicate that the SSPDT is a highly reliable and valid measure of speech recognition, holding considerable promise for future audiological evaluation and management practice. The person-separation reliability and item-separation reliability indices support the reliability of the test, indicating a high degree of internal consistency and a considerable range of difficulty within the item pool. Data also demonstrate that the test can effectively differentiate among listeners with normal-hearing sensitivity and those having mild-to-moderate hearing loss.

The results of multiple regression analyses indicate that performance on the SSPDT is strongly related to high-frequency hearing loss, SRT, listener age, and a simple attribute variable (derived from a set of variables) enabling participants to be categorized as being either normal hearing or hearing impaired (multiple $R = .89$ for quiet listening condition; multiple $R = .90$ for +5 dB SNR listening condition). In contrast, performance on CID Auditory Test W-22 and QuickSIN was only related to high-frequency hearing loss ($r = .57$ and $r = .67$, respectively). Furthermore, the hierarchy of item difficulty observed in the stimulus array reflects the degree to which individuals experience difficulty in processing specific properties of the speech signal (e.g., phonetic features of place, manner, and voicing). The latter finding paves the way for the test to provide diagnostically relevant information concerning the benefits derived from differential interventions or treatments (e.g., deciding among competing assistive devices, evaluating and fitting hearing aids, and determining the efficacy of cochlear implants) for listeners who differ in degree/configuration of hearing loss.

The results of this study provide the foundation for construction of a computerized, adaptive-testing system that can serve as an efficient and precise tool for the clinical assessment of speech recognition. An assessment procedure of this sort uses an "up-down" method of selecting items on the basis of their information value. Like the stimuli used in adaptive psychophysical procedures, the stimuli used in adaptive testing are scaled along a continuum arranged in an ordered fashion, extending from low to high degrees of magnitude. However, rather than representing a physical construct such as the intensity of a sound, the continuum in this instance represents a domain of human performance. Evidence presented in previous research (Bochner et al., 1997) supports the interpretation of this continuum as

a domain of human performance reflecting the construct of speech processing ability. A mathematical model is used to quantify the difficulty of test items scaled along a continuum representing this domain of human performance. The mathematical model enables the discovery of this continuum and the scaling of item difficulty. The continuum of item difficulty, in turn, provides an implicit hierarchy that enables the adaptive testing procedure to use an up-down method of item selection to array listeners on the same continuum as the items.

Adaptive testing offers an efficient and precise approach to the measurement of human abilities. The approach is efficient because it requires a minimum number of items, and it is precise because standard errors are specific to individual examinees and are estimable (and controllable) during the testing process. The results of previous research with an earlier version of the SSPDT suggest that adaptive testing can provide a meaningful measure of speech recognition ability with about 15 items and 5 min of administration time (Bochner et al., 1997). Findings from the present investigation can be used to establish multiple entry points for beginning the adaptive procedure to enhance its efficiency and effectiveness. That is, data from this investigation can be used to establish a simple and direct decision criterion for determining the position in the stimulus hierarchy where adaptive testing should begin for each listener.

Compared to conventional approaches to the clinical assessment of speech recognition (i.e., word recognition testing using PB-50 word lists), adaptive procedures will significantly enhance measurement precision because of the systematic manner in which they determine the ability level of each examinee (cf. Gelfand, 1998; Thornton & Raffin, 1978). In adaptive testing, the items administered to each examinee will, for the most part, be selected from within a relatively narrow range of difficulty circumscribing their level of ability. In addition, the standard error associated with each respondent's performance will be strictly contained within specified limits. The notion of measurement precision for individuals (as opposed to population) represents a major advance within psychometrics. This notion holds considerable promise for the field of audiology because it will enhance the accuracy and integrity of clinical data.

Acknowledgments

The research reported in this article was supported with a Phase I Small Business Innovation Research (SBIR) grant from the National Institute on Deafness and Other Communication Disorders. The authors are most grateful for the valuable support and assistance provided by Liz Laczi. We also want to acknowledge the contributions of Sue Roberts, Ken Johnson, and Jim Orr.

References

- Bochner, J., Garrison, W., & Palmer, L. (1992). Simple discrimination isn't really simple: A confirmatory analysis of the Speech Sound Pattern Discrimination Test. *Scandinavian Audiology, 21*, 37-49.
- Bochner, J., Garrison, W., Palmer, L., MacKenzie, D., & Braveman, A. (1997). A computerized adaptive testing system for speech discrimination measurement: The Speech Sound Pattern Discrimination Test. *Journal of the Acoustical Society of America, 101*, 2289-2298.
- Elliott, L. L., Busse, L., Partridge, R., Rupert, J., & DeGraff, R. (1986). Adult and child discrimination of CV syllables differing in voice onset time. *Child Development, 57*, 628-635.
- Etymotic Research.** (2001). *QuickSIN Speech-in-Noise Test* (Version 1.3). Elk Grove Village, IL: Author.
- Frank, T., & Craig, C. H. (1984). Comparison of the Auditec and Rintelmann recordings of the NU-6. *Journal of Speech and Hearing Disorders, 49*, 267-271.
- Frisina, D. R., & Frisina, R. D. (1997). Speech recognition in noise and presbycusis: Relations to possible neural mechanisms. *Hearing Research, 106*, 95-104.
- Gelfand, S. A. (1998). Optimizing the reliability of speech recognition scores. *Journal of Speech, Language, and Hearing Research, 41*, 1088-1102.
- Gelfand, S. A. (2001). *Essentials of audiology* (2nd ed.). New York: Thieme.
- High, W. S., Fairbanks, G., & Glorig, A. (1964). Scale for self-assessment of hearing handicap. *Journal of Speech and Hearing Disorders, 29*(3), 215-230.
- Laitakari, K. (1996). Speech recognition in noise: Development of a computerized test and preparation of test material. *Scandinavian Audiology, 25*, 29-34.
- Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 40*, 467-477.
- Lord, F. M. (1952). The relation of the reliability of multiple-choice tests to the distribution of item difficulties. *Psychometrika, 17*, 181-194.
- Martin, F. N., Armstrong, T. W., & Champlin, C. A. (1994). A survey of audiological practices in the United States. *American Journal of Audiology, 3*(2), 20-26.
- Maurer, J. F., & Rupp, R. R. (1979). *Hearing and aging: Tactics for intervention*. New York: Grune & Stratton.
- Mendel, L. L., & Danhauer, J. L. (1997). *Audiologic evaluation and management and speech perception assessment*. San Diego, CA: Singular.
- Milenkovic, P. (2000). *TF32-2000* [Computer software]. Madison: University of Wisconsin.
- Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America, 27*, 338-352.
- Peterson, G. E., & Barney, H. L. (1952). Control methods used in a study of the vowels. *Journal of the Acoustical Society of America, 24*, 175-184.
- Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18*, 43-52.
- Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago: University of Chicago Press. (Originally published 1960)
- Revoile, S. G., & Pickett, J. M. (1982). Speech perception by the severely hearing-impaired. In D. G. Sims, G. G. Walter, & R. L. Whitehead (Eds.), *Deafness and communication: Assessment and training* (pp. 25-39). Baltimore: Williams & Wilkins.
- Shaw, F. (1991). Descriptive IRT vs. prescriptive Rasch. *Rasch Measurement, 5*(1), 131.
- Studdert-Kennedy, M., & Shankweiler, D. P. (1970). Hemispheric specialization for speech perception. *Journal of the Acoustical Society of America, 48*, 579-594.
- Thornton, A., & Raffin, M. (1978). Speech discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research, 21*, 507-518.
- Tinkelman, S. N. (1971). Planning the objective test. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., p. 63). Washington, DC: American Council on Education.
- Tyler, R. S. (1994). The use of speech-perception tests in audiological rehabilitation: Current and future research needs. In J. -P. Gagne & N. Tye-Murray (Eds.), *Research in audiological rehabilitation: Current trends and future directions* [Monograph Supplement]. *Journal of the Academy of Rehabilitative Audiology, 27*, 47-66.
- Wright, B. D., & Linacre, J. M. (1991). *BIGSTEPS: Rasch model computer program* (Version 2.0) [Computer software]. Chicago: MESA Press.
- Wright, B., & Stone, M. (1991). *Separation statistics* (Research Primer No. 1). Wilmington, DE: Jastak Associates.

Received April 23, 2002

Accepted January 14, 2003

DOI: 10.1044/1092-4388(2003)069

Contact author: Joseph H. Bochner, PhD, National Technical Institute for the Deaf, Rochester Institute of Technology, 52 Lomb Memorial Drive, Rochester, NY 14623-5604.
E-mail: jhbnep@rit.edu