

# A computerized adaptive testing system for speech discrimination measurement: The Speech Sound Pattern Discrimination Test

Joseph Bochner,<sup>a)</sup> Wayne Garrison, Linda Palmer, Douglas MacKenzie,  
and Amy Braveman

National Technical Institute for the Deaf, Rochester Institute of Technology, P.O. Box 9887, Rochester,  
New York 14623

(Received 13 December 1995; revised 29 November 1996; accepted 12 December 1996)

A computerized, adaptive test-delivery system for the measurement of speech discrimination, the Speech Sound Pattern Discrimination Test, is described and evaluated. Using a modified discrimination task, the testing system draws on a pool of 130 items spanning a broad range of difficulty to estimate an examinee's location along an underlying continuum of speech processing ability, yet does not require the examinee to possess a high level of English language proficiency. The system is driven by a mathematical measurement model which selects only test items which are appropriate in difficulty level for a given examinee, thereby individualizing the testing experience. Test items were administered to a sample of young deaf adults, and the adaptive testing system evaluated in terms of respondents' sensory and perceptual capabilities, acoustic and phonetic dimensions of speech, and theories of speech perception. Data obtained in this study support the validity, reliability, and efficiency of this test as a measure of speech processing ability. © 1997 Acoustical Society of America. [S0001-4966(97)06704-0]

PACS numbers: 43.71.Gv, 43.71.Ky, 43.66.Sr [WS]

## INTRODUCTION

Conventional methods and materials used in the clinical assessment of speech processing may be traced to the development of articulation and intelligibility tests (Egan, 1948; Fletcher, 1929). Specifically, the work of Egan and his colleagues at the Harvard Psychoacoustics Laboratory in the 1940s led to the development of phonetically balanced monosyllabic word-recognition measures, such as the CID Auditory Test W-22 (Hirsch *et al.*, 1952). These so-called PB-50 tests historically have constituted the primary tools of speech audiometry. Other, more recent approaches to the assessment of speech perception, such as the SPIN test (Kallikow *et al.*, 1977) and the MAC battery (Owens *et al.*, 1985), have used different methods and materials designed to provide more useful information about individuals' speech processing ability.

In speech audiometry, tests have also been developed to assess listeners' ability to perceive phonetic segments and patterns. For example, tests of consonant perception in word and nonsense-syllable contexts have been developed (Dubno *et al.*, 1982; Owens and Schubert, 1977). These tests, however, have not seen widespread use in clinical or rehabilitative settings. Given the history and limitations of speech perception tests currently available, a need exists for the advent of new approaches to the measurement of speech processing abilities adapted to the characteristics of individuals receiving audiological services (Tyler, 1994).

At the heart of the adaptive testing procedure is the simple proposition that an examinee is measured most effectively when the test tasks are neither too difficult nor too

easy. For speech discrimination measurement, this condition can occur when an examinee responds correctly to, say, a predetermined percentage of test items administered (e.g., 50%). Consequently, adaptive testing involves the selection of test items during the testing process which are appropriate in difficulty level for a given examinee. Adaptive tests are designed to provide successively refined estimates of an examinee's proficiency or capability, on the basis of his/her responses to items already administered. Using these successive approximations of, in the present discussion, sensory and perceptual capability, decisions are made about what (if any) test items are to be administered next. This iterative procedure generally continues until established criteria for test termination have been met. Because of their computational and branching requirements, adaptive tests most frequently are implemented using computer-interactive methods.

This paper describes the development of a computerized adaptive form of the Speech Sound Pattern Discrimination Test (SSPDT) developed by Bochner *et al.* (1986), and evaluates its reliability and validity from multiple perspectives. The rationale for developing this test was to construct a reliable, valid instrument for measuring speech processing ability that utilizes naturalistic sentence-length utterances and efficiently provides meaningful information about the perceptual capabilities of listeners with a wide range of hearing loss and English language proficiency. The initial version of the SSPDT was a prototype containing only 30 items. The current SSPDT item pool includes 130 items used in an adaptive testing format. The purpose of this study was to evaluate the hypothesis that a computerized adaptive form of the SSPDT can serve as a functional tool for speech discrimination measurement; hence, measurement reliability

<sup>a)</sup>Electronic mail: JHBNCP@ritvax.isc.rit.edu

and validity have been approached from a variety of perspectives.

## I. METHOD

### A. A psychometric model for person measurement

Although it is possible to design and administer computerized adaptive tests without an explicit theory of item responses, psychometric models that represent the influence of specific person and item parameters on the outcome of the person-item interaction have proven useful. Item response theory models are mathematical abstractions based on suppositions or hypotheses about what happens when an examinee responds to a test item. The simplest of these models, and also the basis of the present work, is the generalized Rasch (1980) model for person measurement.

In the situation where examinees' responses to test items can be scored "right/wrong," the Rasch model provides a means of predicting success or failure on specific test items in probabilistic terms. These probabilities reflect the difference between an examinee's position along a continuum of capability, and the difficulty of items scaled along the same continuum. Thus, the Rasch model conceptualizes the result of the person-item interaction in terms of a single person parameter (i.e., ability or capability, used interchangeably in the current discussion), and a single item parameter (i.e., difficulty). These parameters are expressed in a metric known as logits. A person's ability in logits is his/her natural log odds for succeeding on items with "zero" difficulty. The item difficulty scale is centered at zero. An item's difficulty is its natural log odds for eliciting failure from persons with "zero" ability (Wright and Stone, 1979). In the present study, the item difficulty parameter was decomposed further to investigate factors suspected of causing variation in the item difficulty values.

### B. Item pool

The SSPDT is a flexible, broad-range test of speech discrimination ability. It is constructed on a respondent-by-respondent basis from a pool of 130 potential items. An item is a modified discrimination task in which listeners are required to make judgments on whether each of four sentences, presented in succession, is identical to a standard (target) sentence. Accordingly, the term *item*, as used in this paper, denotes a set of four trials. An orthographic representation of the target sentence appears on a computer monitor throughout the presentation of acoustic stimuli. The listener must indicate whether comparison sentences are either the same or different using designated response keys on the computer keyboard. A special keyboard template can be used to prevent access to all but appropriate response keys.

Examinees are instructed that any number of comparison stimuli may match the target, making it possible for 0–4 matches to occur for a given target. Matching stimuli are exact repetitions of the target utterances. Items are scored correct/incorrect as blocks, minimizing the influence of guessing on the estimation of ability. That is, items scored correct are those for which all four discrimination judgments

are correct. An error on any one of the four comparison tasks within a block results in an item score corresponding to incorrect.

### C. Instrumentation

The stimulus sentences were uttered by a male speaker with a General American dialect. The speech was digitized at 10 kHz for storage on computer disk using a 12-bit A/D module (AD12FA) interfaced to a Masscomp 5600 UNIX Workstation. Rockland antialiasing filters with a roll off of 48 dB/oct were set to 5 kHz. A body microphone (Electrovoice RE51) was placed 1 in. from the speaker's mouth. A preamp (Shure M67) and compression limiter (dbx 161) were used to amplify the microphone signals. The speaker was instructed to utter each sentence in an item block with the same pace, clarity, and effort, changing only the portion containing lexical/phonetic contrasts. Each sentence was spoken three times. The best utterances were later selected and edited into individual sentence files. Recordings were made in a double-walled IAC sound booth. Signal levels were monitored on a VU meter throughout the recording to assure that the peaks were not clipped.

### D. Adaptive testing procedure

The adaptive testing system elaborated herein operates from a calibrated item pool and, generally, comprises three components: (1) an item selection routine; (2) an ability estimation technique; and (3) rules for test termination. The calibrated item pool consists of SSPDT items and associated difficulty values. The item calibrations (i.e., difficulty values) were those obtained using the BISGSTEPS Rasch scaling program developed by Linacre and Wright (1994). Data for the estimation of difficulty values were the scored responses to items presented to listeners in this study.

Procedurally, the adaptive testing system operated in the following manner:

(1) The first SSPDT item administered to each examinee had a difficulty value closest to the central reference value of 0.00 logit.

(2) If an examinee responded correctly, an item of greater difficulty was administered next. If an examinee responded incorrectly, an item of lesser difficulty was administered. The difficulty increment (decrement) was set at 0.50 logit.

(3) When the response record contained at least one correct and one incorrect item score (1 and 0, respectively), a finite maximum likelihood estimate of ability, and its associated standard error, was obtained [the reader is referred to Wright and Stone (1979) for solution equations]. The numerical estimation method employed here has the property that an examinee's number right (raw) score is a sufficient statistic for estimation of the ability parameter.

(4) Using the estimate of ability just computed, test items in the pool which had not already been administered were evaluated for their potential to enhance information about the examinee's speech discrimination ability. Specifically, the next item to be administered in the testing sequence was that with difficulty closest to the ability estimate.

with the further provision that the difference between the ability estimate and the selected item's difficulty had to be less than the standard error associated with the ability estimate (i.e., difficulty in the range  $\pm$ one standard error of the ability estimate). Otherwise, items remaining in the pool (not administered) were considered to be out-of-range, and testing was terminated.

(5) After each newly selected test item was presented and the examinee's response evaluated for correctness/incorrectness, the ability estimate was recomputed, making use of the additional information. The new ability estimate was compared, as before, with the difficulty values of items remaining in the pool. Items continued to be administered as long as the difference between their difficulty and each newly computed ability estimate remained within a (decreasing) standard error of measurement.

(6) Testing was terminated when items remaining in the pool were of inappropriate difficulty (i.e., outside the established range), or when a maximum number of items had been administered (25 items here).

### E. Subjects and test administration

Seventy-three adults participated in the study. Seventy-two were students with sensorineural hearing losses enrolled in courses of study at the National Technical Institute for the Deaf and one was a normal-hearing undergraduate. Ages of the study participants ranged from 17 to 49 years, with a mean age of 22 years. Subjects were paid for their participation.

Twenty-eight subjects (sample A) were administered a common set of seventy items (fixed-item nonadaptive format) during the spring, 1993. Adaptive tests for sample A respondents were *simulated* from their responses to these 70 items. Data obtained from these participants were used primarily to estimate the difficulty values of these 70 items. The second purpose was to evaluate a simulated adaptive testing protocol.

Forty-five subjects (sample B) responded to a different, common set of sixty items (fixed-item nonadaptive format). These 60 items were intended to extend the range of difficulty upward from that realized from the sample A testing. Sample B respondents were also administered an *actual* adaptive test with an item pool composed of the 70 items calibrated from the sample A testing. Data obtained from sample B examinees were used primarily, then, to evaluate the adaptive testing system. Simultaneously, these data provided an empirical basis for the calibration of 60 new SSPDT items. Subjects in sample B were tested during the spring and fall of 1994.

Five subjects were common to samples A and B. Consequently, for correlational analyses to be reported later, the final sample size was reduced from 73 to 68 respondents (sample C). Sample C subjects had a mean pure-tone average (ANSI S3.6, 1989) in the better ear for 0.5, 1, and 2 kHz of 73.1 dB HL (s.d.=23.8). The range was 107 dB (minimum=0, maximum=107).

An effort was made to group subjects on the basis of their audiometric configuration using the criteria employed by Dubno *et al.* (1982). A group of 22 subjects was identi-

fied as having flat hearing losses. A group of 13 subjects had gradually sloping high-frequency hearing losses. A group of nine subjects had steeply sloping hearing losses. The remaining subjects were either not tested to 4000 Hz, or their audiograms could not be clearly placed into one of these three configuration classes.

All subjects were tested individually in a sound-treated room and received stimuli dichotically under earphones (TDH-39P) at a comfortable listening level using a clinical audiometer (Madsen OB 822). In order to establish a comfortable listening level for each subject, a sample stimulus was presented repeatedly and its level adjusted until the subject reported it to be most comfortable. For the nonadaptive testing component of this study involving subjects in sample A, the digitized stimulus sentences were converted to analog signals and recorded on a cassette tape recorder (Nakamichi 1000 II). The stimuli were played back to subjects in sample A on another tape recorder (Wollensak 2556 AV) routed through an amplifier (Crown model D 60). A warning light cued subjects to the onset of each stimulus, and the subjects used pen/pencil to indicate their responses ('S' for same, 'D' for different) on an answer sheet.

Subjects in sample B were seated in front of a VT100 computer terminal. The edited sentence files were played back directly from the computer, routed through a 12 bit D/A module (DA04H), and low-pass filtered at 5 kHz (48-dB/oct attenuation). The timing of the stimuli was controlled by a programmable clock module. A flash of light from a visual response box placed next to the terminal alerted subjects to the onset of a sentence, and number prompts displayed on the computer screen marked opportunities to respond 'S' for same or 'D' for different. Practice items were provided to all subjects to ensure that the task was understood. Each testing session took approximately 1 h. The time required for the administration of the adaptive test to subjects in sample B, however, was approximately 10 min.

## II. RESULTS

### A. Model/data conformity

The usefulness of any psychological testing device is gauged, in part, by its success in differentiating respondents. This amounts to asking the question, "How well do test items separate the persons tested?" In Rasch measurement practice, a person separation index has been developed (Wright and Stone, 1991). The person separation reliability (PSR) index, ranging in value from 0.00 to 1.00, provides insight into the extent to which the items comprising a test are members of the same conceptual domain. The reliability index for the 70 items calibrated from sample A examinees was found to be 0.95, indicating that these items separated the 28 respondents very well. We conclude from this finding that these 70 items define a single, dominant variable (i.e., the items are homogeneous and internally consistent).

In Rasch measurement practice, an item separation reliability (ISR) index has also been introduced by Wright and Stone (1991) to evaluate how well respondents differentiate items. This amounts to asking the question, "How broad is the range of difficulty operationalized by an item set?" The

ISR is algebraically similar to the PSR and ranges in value from 0.00 to 1.00. The ISR for the sample A data was found to be 0.90, indicating that the difficulty of these 70 items spread over a considerable range, giving breadth and meaning to the variable being assessed (i.e., speech discrimination). The mean difficulty value was 0.00 logit (s.d.=2.14), with item difficulties spanning a range of -4.61 to +5.00 logits. The distribution of item difficulty values was symmetrical and, essentially, normal.

PSR and ISR are gross statistical indicators, providing a general characterization of test qualities. To guide us in our understanding of more specific person and item attributes, two additional statistics have been introduced. These statistics, usually expressed as INFIT and OUTFIT, are discussed elsewhere at length by their developers (Wright and Linacre, 1991; Wright and Masters, 1982; Wright and Stone, 1979). In general, item fit values reflect the extent to which the observed responses to items (across persons) agree with those predicted by the Rasch measurement model. Examination of the fit statistics associated with the 70 items calibrated from sample A data indicated conformity between observation (data) and prediction (model). That is, there were no misfitting items.

When the INFIT and OUTFIT statistics are applied to persons, we are able to evaluate the extent to which individual examinees' responses to items are in accord with those predicted by the measurement model. One of the 28 individuals included in sample A had an INFIT statistic indicating misfit. Person misfit to the measurement model results when a respondent makes too many correct discriminations on items predicted to be much harder than the examinee is able or, conversely, when a respondent makes many errors on items predicted to be relatively easy. Aberrance within the response vector for this examinee was attributed to "lapse of attention," noted during the testing session by the audiologist who was overseeing the test administrations.

The final analysis performed on the sample A data involved the simulation of adaptive test records for these 28 individuals. Using the adaptive testing procedure detailed earlier, the simulated test record represented scored responses to a *subset* of the 70 items. The first item in the simulated record was the same for all respondents. Thereafter, a variable branching scheme which involved person-ability/item-difficulty comparisons was used to construct the simulated test records. Ability estimates obtained from the adaptive test simulation were compared, then, to the ability estimates obtained in the situation where the same examinees responded to all 70 items.

The mean performance of the 28 subjects on the 70 items was 1.33 logits (s.d.=1.90) and the mean performance of the same individuals on the simulated tests was 1.43 logits (s.d.=2.12). For the group, the difference in performance between simulated and fixed-item conditions is not statistically significant.

When the data were subjected to casewise analysis, 27 of the 28 examinees obtained measures on the simulated tests that were statistically equivalent to the ability estimate computed from their responses to all 70 items. This result is

reflected in a substantial correlation of +0.95 between the score pairs. The statistical test for the equivalence of two measures evaluates the magnitude of a difference in terms of the expected standard error of the difference.

The one individual for whom statistical equivalence of measures was not observed was the same individual suspected of "attention lapses" by the audiologist who supervised test administration. For all 70 items, this subject attained a measure of 2.06 logits. On the simulated test, the same individual attained a measure of 4.61 logits. The aberrance within the item response vector signaled by the magnitude of the person fit statistics in the first instance was not manifested in the latter (i.e., the person fit statistics in simulated test mode failed to reach significance level). Importantly, what we observe for this examinee is a deleterious effect of administering test items which, in the fixed-item format, are off-target (e.g., items which are too easy).

The average length of a simulated test was 14 items (s.d.=3.4). Thus, only 20% of the items in the pool were required to reproduce the ability measures implied by responses to all 70 items. The minimum number of items administered under the simulated condition was 10, the maximum was 23. The frequency distribution of number of items administered was positively skewed, with 75% of the persons tested responding to 15 or fewer items. Testing was terminated for all respondents due to insufficiency of items remaining in the pool which were appropriate in difficulty level.

With the goal of enlarging and extending the range of usefulness of the item bank, and to evaluate the adaptive testing system in an actual rather than simulated setting, additional data were obtained from 45 individuals (sample B). Specifically, individuals in sample B interacted with a computerized adaptive testing system limited to administration of the 70 SSPDT items discussed above. The same persons also responded to a fixed set of 60 new, experimental items (also computer administered, but in a nonadaptive format) which were evaluated for fit to the Rasch measurement model and, hence, addition to the item bank.

The person separation reliability for the 60 new SSPDT items, calibrated from the sample B data, was found to be 0.94, indicating that these items separated the 45 persons tested very well. Mean performance of the same individuals on the 60-item fixed form of the SSPDT was 1.12 logits (s.d.=1.79). The item separation reliability for the sample B data was found to be 0.91, indicating that the difficulties of the 60 new items spread out over a broad range.

Fifty-nine of the sixty new SSPDT items had fit statistics indicating model/data conformity. One item had associated INFIT and OUTFIT values indicating that it should be monitored (i.e., administered, but not scored) presently. Forty-four of the forty-five persons tested fit the Rasch measurement model. That is, in the evaluation of the regularity of individual response vectors, 44 persons responded to items consistent with model predictions; irregularity in the response pattern was observed for one examinee, reflecting too many errors on items predicted to be relatively easy for the respondent.

Mean performance of the 45 persons tested on the adap-

tive SSPDT format was 2.64 logits (s.d.=1.91). The measurements of these persons on the adaptive and fixed-item formats constitute the essential "common person" data for linking the 60 new items onto the reference scale defined by the 70 items calibrated with sample A data. Specifically, the difference in the two ability means [ $2.64 - 1.12 = 1.52$  logits] corresponds to the translation constant necessary to bring the items calibrated with sample A and sample B data onto a common scale.

With this scaling adjustment, the 60 items calibrated using sample B data spanned a difficulty range of  $-1.19$  to  $+5.00$  logits. The intent to extend the range of difficulty of the items upward from that observed with the sample A data was not realized. Rather, the sample B data served to enlarge the item bank in regions defining moderate and very difficult speech discrimination tasks.

For the adaptive test format, the mean number of items administered to sample B respondents was 14 (s.d.=4.0). This is consistent with what was observed with simulated adaptive testing in sample A. All but one examinee exited the testing sequence due to insufficiency of items remaining in the pool with appropriate difficulty levels. One examinee (with normal hearing) attained a perfect score upon reaching the extreme level of item difficulty.

Just as the items were combined into a common bank, the persons tested were pooled into a common sample (sample C). For each individual in sample C, there were two estimates of speech discrimination ability. One estimate, and its associated standard error, was obtained using a common-item test format; the other estimate, and its associated standard error, was the result of adaptive test administration (simulated or real). Sixty-one of the sixty-eight persons in sample C attained statistically equivalent ability estimates. For these 68 individuals, the Pearson correlation between the ability estimates was  $+0.93$ . For sample B respondents only, the comparable correlation was  $+0.92$ . We conclude from analyses presented to this point that the test data are characterized rather well by the Rasch model for person measurement. Further, once we have estimated the item difficulties, we are able to tailor tests to test-takers in a manner which is both informative and efficient.

## B. Item content analysis

The analysis of item content is concerned with the characterization of contrasts occurring within the block of trials comprising an item. The contrasts manifested in each block of trials were characterized according to categories of phonetic features and acoustic properties. In this analysis, consonant contrasts were, for the most part, classified according to phonetic features of place, manner, and voicing, while vowel contrasts were classified according to features of tongue position/advancement and tongue height. These phonetic features are associated with specific acoustic properties of speech, and their perception is known to vary as a function of the portion of the speech spectrum which is audible to listeners (e.g., Dubno and Levitt, 1981; Miller and Nicely, 1955; Pickett *et al.*, 1970). Trials involving replication of the target sentence (i.e., "same" trials) have not been included in this analysis because they do not involve a contrast (i.e., a

difference) in item content and because previous research has shown that these trials are inherently easier than their counterparts (Bochner *et al.*, 1992). Four of the 130 SSPDT items had 4/4 trials which were replications of the target stimulus (i.e., comparisons were ALL "same").

Since an item is scored correct if and only if an examinee's response to each of the four trials is correct, we reasoned that the difficulty of an item would be determined by that of its most difficult trial. An analysis of examinees' responses supported this reasoning, indicating that trials which involved contrasts in the phonetic features and acoustic properties expected to be least audible consistently presented the greatest difficulty for listeners. Within each item, then, the trial involving the least audible contrast may be regarded as dominant because it determines the difficulty of the entire block of four trials comprising the item. In a few items, however, one trial could not be deemed more difficult than others. Such items are, therefore, regarded as having more than one dominant trial.

The test items were characterized in terms of four content categories. Three categories account for one or more segmental contrasts occurring within a single syllable, and characterize item difficulty in terms of differences in phonetic and acoustic content observed between the pair of utterances comprising the dominant trial. The fourth category accounts for gross contrasts extending across two or more syllables, and characterizes item difficulty without reference to phonetic features or the notion of a dominant trial. The four categories are ordered from most difficult (category Z) to least difficult (category W). Examples of items in each category are shown in Table I.

Category Z items are characterized by spectral cues contained in the vicinity of the second formant, such as cues for place of articulation for consonants and tongue position/advancement for vowels. Contrasts among semivowels (e.g., "led"—"wed"—"read") are also included in category Z, as are certain contrasts involving sibilants. Some items classified in category Z contain more than one contrast within a single syllable. Such items may contain a contrast involving the semivowels /y,w,l,r/, an extra /s/ segment, and/or a contrast in place of articulation (e.g., "trap"—"strap," "stray"—"clay," and "throat"—"float"). These contrasts are characterized by phonetic cues contained in the upper regions of the speech spectrum. Since their perception tends to be associated with acoustic cues in the vicinity of the second formant and above, and they occur within a single syllable, these contrasts conform to the criteria for classifying items in category Z.

Category Y items are characterized by spectral and temporal cues appearing in the vicinity of the first formant, such as cues for manner of articulation for consonants (with or without an accompanying place contrast) and tongue height for vowels. Contrasts involving the presence of an extra voiceless consonant (other than an extra /s/ segment as described above), and/or an extra semivowel within a consonant cluster are also included in category Y, as are some items containing more than one contrast within a single syllable (e.g., "car"—"cart," "tie"—"try," "crime"—"time," and "shoes"—"clues"). Category Y, therefore, in-