Comparing Two Methods of Evaluating Aided Speech Recognition Performance for Single Cases

Adrienne Rubinstein, Rochelle Cherry, and Kim Schur
Department of Speech
Brooklyn College, City University of New York

Two methods used to assess changes in speech recognition performance for individual subjects are the binomial distribution (BD) tables of Thornton and Raffin (1978) and single-case design. The present study was designed to determine if data analyzed using both approaches would result in different outcomes. The speech recognition ability of 20 adults with hearing loss was evaluated in a rapidly-alternating treatments design under 2 conditions: with a frequency response approximating gain as recommended by the Revised National Acoustics Laboratory (NAL) procedure, and 1 using 6-dB less gain in the low frequencies and 3-dB more gain in the high frequencies. Results revealed that the information obtained during the single-case analysis did not change the outcome as compared with the binomial model when both analyses were performed using 100-word lists. Using the BD with 25-word and 50-word lists, however, did result in different outcomes. The findings should be treated as preliminary and replicated in the context of a more powerful independent variable.

With rising concern over health care costs, professionals providing hearing services must acknowledge the demand for accountability regarding the value of expensive circuitry in hearing aids. Successful clinical service delivery requires the development of methods to assess performance characteristics of circuitry designed to achieve a certain goal (Beck, 1991; Humes, Christensen, Bess, & Hedley-Williams, 1997). The typical method for assessing new hearing aid circuitry is through the use of group experimental designs. Despite the obvious advan-

Kim Schur is currently at the League for the Hard of Hearing, New York, NY 10010.

Correspondence concerning this article should be addressed to Adrienne Rubinstein, Department of Speech, Brooklyn College, City University of New York, 2900 Bedford Avenue, Brooklyn, New York 11210. Telephone (718) 951-5186. Fax: (718) 951-4363.

tages to group designs, there are a number of problems. A major issue is the obscuring of individual clinical outcome in group averages. When only a segment of the subjects is affected by the treatment, statistical procedures will average out clinical effects along with changes due to unwanted sources of variability. Because results from group studies do not reflect changes in individual patients, these results are not readily translatable to the practicing clinician (Barlow & Hersen, 1984).

The assessment of individual data is needed not only in research but in hearing clinics where clinicians are required to make decisions regarding hearing aid characteristics and circuits without necessarily knowing at the time which qualities or attributes of the client (and his auditory performance) make him suitable for a given electroacoustic characteristic. Subjects may be similar in terms of known attributes yet they may respond differently in terms of satisfaction with a particular circuit (Bentler, 1991).

Research in the area of hearing aids acknowledges this problem, and even in studies where group designs are used, attempts are often made to evaluate data from individuals (Kompis & Diller, 1994; Stein & Dempsey-Hart, 1984; Wolinsky, 1986). This approach can be problematic because it may introduce flaws which were meant to be eliminated by the group design. Several methods have been used to assess differences between conditions in an individual's data. One popular approach has been to use the computer-generated tables of Thornton and Raffin (1978) to establish statistical significance of speech recognition score differences using phonemically balanced (PB) word lists. Phonemically-balanced word lists continue to be the most widely used material for assessing speech recognition (Martin, Armstrong, & Champlin, 1994; Olsen, Van Tasell, & Speaks, 1997). The Thornton and Raffin tables, based on the binomial distribution (BD), have been applied to several areas in audiology (Brimacombe, Arndt, & Staller, 1995; Hochberg, Boothroyd, Weiss, & Hellman, 1992; Sandridge, Goldberg, & Workman, 1994; Wolinsky, 1986).

Walden, Schwartz, Williams, Holum-Hardegen, and Crowley (1983) used the binomial model to compare speech recognition performance within two groups of hearing aids: one group which was electroacoustically similar and preselected for the loss being fit, and the other group which was electroacoustically dissimilar and unlikely to be preselected. They found that significant interaid differences occurred frequently only when electroacoustic characteristics were very different. They concluded that due to test-retest reliability, monosyllabic word lists may not be able to detect interaid differences when comparing circuitry more similar and appropriate for the loss.

Dillon (1982, 1983) reviewed the limitations of the binomial theorem to predict test-retest variability for speech recognition assessment. He noted that the binomial model assumes that the degree of variability depends only on (a) the speech recognition score obtained, and (b) the number of items in the test. Therefore, predictions based on this model are independent of attributes of the particu-

lar subject under test. In reality, these assumptions may be violated which may result in an inaccurate prediction by the model. First, the use of test items of differing degrees of difficulty can be shown to result in an underestimation of the true variability by the model. The second violation is in the assumption of constant probability over time, due to the presence of such factors as fatigue or learning, which would result in greater variability than predicted by the model. Whereas the consequences of these two violations tend to be in opposition, they offset one another. Based on his review, Dillon concluded that the predictions of the binomial distribution should fall close to empirical values for the average subject, however, it may not lead to valid results for a particular individual.

An alternative to the above approach is the use of single-case research design. Intrasubject variability is highlighted through repeated and frequent measurement of the dependent variable. Stein, McGee, and Lewis (1989) were among the first to demonstrate how a single-subject design could be applied to issues in audiology, in the context of a noise-reduction study. Single-case design is being applied with greater frequency to issues in audiology (Brainard & Lesner, 1992; Chmiel & Jerger, 1995; Foust & Wynne, 1991; Parent, Chmiel, & Jerger, 1998). An advantage of the single-case design is that unlike the BD, it provides information on the performance stability for the specific individual being tested (Chmiel & Jerger, 1995), avoiding the potential pitfall of using averaged and thus, less accurate data. For example, subjects with more stable data might demonstrate clear differences in performance with similar circuitry using a single-case design, which might not be revealed using the BD.

The main disadvantage of the single-case design, however, is the time required to take multiple measures. This is especially problematic in the clinical environment, and what makes the BD so appealing. The purpose of the present pilot study was to determine if findings comparing two electroacoustically similar aided conditions would result in different outcomes when data are analyzed using a single-case design versus a BD approach.

METHOD

Subjects

Twenty adults with postlingual hearing loss served as subjects. Selection criteria included: (a) unaided speech recognition threshold of at least 65 dB HL in the better (test) ear, and (b) unaided speech recognition score[1] of at least 50%. Ten subjects were hearing aid users. Table 1 shows the age, ear tested, pure-tone air-conduction thresholds, speech recognition score for the test ear, hearing aid status, and signal-to-noise (SN) ratio used for each subject during the experiment. In cases where both ears met selection criteria, the ear with the higher speech recognition score was chosen.

[1]Speech recognition evaluated at 40 dB above an individual's speech reception threshold.

Table 1

Patient Ages (in Years) As Well As Their Pure Tone Air Conduction Thresholds (in dB HL) and Speech Recognition Scores (SRS) in the Test Ear, Ear Tested, Hearing Aid Status, and Signal-to-Noise (SN) Ratio Used

		Frequency in kHz								Hearing	SN
Subject	Age	0.25	0.50	1.00	2.00	4.00	8.00	SRS	Ear	aid	ratio
1	87	40	40	45	60	75	95	84%	R	NO	9
2	68	70	60	60	55	80	85	96%	R	NO	5
3	74	40	45	25	25	80	80	68%	R	YES	2
4	69	35	40	30	50	70	95	84%	R	YES	12
5	75	35	45	40	50	75	95	88%	R	NO	0
6	72	45	40	45	30	20	60	96%	R	NO	3
7	64	20	30	55	55	60	110	68%	R	YES	2
8	69	20	25	30	60	70	90	78%	R	NO	14
9	75	15	25	45	45	40	70	84%	R	YES	2
10	78	25	40	40	50	60	70	86%	R	YES	2
11	91	30	45	40	45	60	70	80%	L	YES	7
12	75	45	55	50	50	70	70	98%	L	YES	4
13	74	55	60	60	65	80	105	54%	L	YES	10
14	80	45	50	45	50	70	90	82%	L	YES	6
15	84	25	35	40	40	65	85	80%	R	NO	3
16	70	20	15	30	65	75	95	60%	R	NO	6
17	65	15	25	20	70	85	95	84%	L	NO	4
18	64	30	35	55	55	55	75	92%	R	YES	-3
19	85	50	40	40	40	55	75	76%	R	NO	7
20	70	15	25	40	40	45	70	84%	L	NO	-2

Equipment

A noise-reduction strategy was simulated using a programmable multichannel system (PMC) and a Siemen's Programmable Triton 3004 postauricular hearing aid. Comparisons were made between two frequency responses: (a) a frequency response aimed at approximating gain as recommended by the Revised National Acoustics Laboratory (NAL) prescriptive procedure, and (b) a frequency response approximating 6-dB less gain in the low-frequency band, no change in the mid-frequency band, and 3-dB more gain in the high-frequency band, aimed at noise reduction (NR). The change in frequency response was based on research by Kuk and Pape (1992) who found this to be a discriminable difference for 80% of their subjects. An Audioscan RM500 real ear measurement system/hearing aid analyzer was used to verify the gain of the hearing aid for each subject and each condition and to assure that output fell within ±1 dB of target levels. Speech material was played back on a Technics RS-TR272 stereo cassette player via a Grason Stadler 16 audiometer.

Procedure

Following the fitting of amplification, each subject was seated 1 m from the loudspeaker at a 0° azimuth in a double-wall sound-treated booth, and was asked to face directly towards the speaker during presentation of the stimuli. To ensure monaural presentation, an E.A.R. plug was inserted into the nontest ear. All speech material was commercially produced by Auditec of St. Louis.

An adaptive pretesting procedure was performed for each subject to attempt to establish a signal-to-noise ratio which would approximate a 50% score for the NAL condition (Levitt, 1971). The goal of this pretesting was to avoid ceiling/floor effects during the study. Pilot data had revealed that this could be accomplished by using a CID W-22 list and whole-word scoring. The subsequent test protocol included NU-6 word lists with phoneme scoring. All eight NU-6 lists were required for the test protocol, thus a W-22 list was used for pretesting. Although the W-22 lists produce higher scores than NU-6 lists, this difference was offset by the whole-word versus phoneme scoring. Phoneme scoring was adopted to increase the sensitivity of the measure by increasing sample size (Olsen et al., 1997). The speech level was set at 50 dB HL with cafeteria noise initially set at 30 dB HL. Noise was increased in 8-dB steps until the first error, decreased in 4-dB steps until a correct response was obtained, and varied in 2-dB steps thereafter. The midpoints of the last eight reversals in 2-dB step size were averaged to establish the appropriate noise level for each subject. The spectrum of cafeteria noise is dominated by low-frequency information, but also contains some high-frequency components and simulates a real-life listening environment (Humes et al., 1997).

An experimental rapidly-alternating treatment (RAT) single-case design was used (Tawney & Gast, 1984). This design may be distinguished from the more

familiar ABA design. The ABA design typically compares a no-treatment (A) baseline and a treatment (B) condition, followed by another no-treatment (A) phase. In the RAT design, the comparison is more often made between two treatment conditions. The major difference, however, is that in the ABA design, several data points are collected for each condition before moving to the next condition, whereas in the RAT design, single data points may be collected as one alternates several times between the two treatment conditions. This offers an advantage in that results can be obtained with fewer data points because it requires neither a formal baseline phase nor multiple measures within each phase. This is especially useful when time is of the essence and there are limitations in the number of available lists.

The test protocol was carried out in a single session. Frequency-response conditions were alternated until each had been presented four times, using 25 words for each presentation. The decision to use 25 words was made due to its common clinical practice (Martin et al., 1994; Wiley, Stoppenbach, Feldhake, Moss, & Thordardottir, 1995) and in consideration of availability of materials and time needed for making multiple measures. The choice of condition as well as list presented first was randomized among subjects. Phoneme recognition scores were calculated in percent correct for each of the eight presentations (4 lists × 2 conditions) with 75 phonemes (25 words × 3 phonemes) per list. To determine if a critical difference had been exceeded between the two conditions, the score from the first list used for one condition was compared with the score from the first list from the second condition, based on the BD calculations of Thornton and Raffin (1978). An adjustment was made because it would have been incorrect to assume 75 independent pieces of information. The effect of phonemic constraints was taken into account by using the equations derived by Boothroyd and Nittrouer (1988) who compared recognition probability scores of the whole word to the probability scores of the individual phonemes, as follows:

$$p_w = p_p{}^j \tag{1}$$

where p_w is the probability of recognition of the whole word, and p_p is the probability of the recognition of an individual phoneme when j is known. j is calculated by:

$$j = \log(p_w)/\log(p_p) \tag{2}$$

Boothroyd and Nittrouer found that for CVC words, the j value was 2.5, which may be interpreted to mean that recognition of only 2.5 of the phonemes was needed for correct recognition of the entire word. Boothroyd and Nittrouer pointed out that whereas their findings revealed the general magnitude of the effect for CVC words, recalculation of j values is advised with changes in data by such factors as test material and subject pool. In the present study, an average j value was calculated from the j values obtained from each subject, and multiplied

by 25 to obtain the correct number of independent sources of information for subsequent analysis by the BD model.

The preferred method for establishing significance in single-case paradigms is controversial. Opinions range from those who strongly recommend the use of inferential statistics to those who vehemently object to their use (Kratochwill & Levin, 1992). Critics of visual inspection object to the element of subjectivity in the judgments, which can be shown to affect interrater agreement (Ottenbacher, 1993). Proponents, on the other hand, can point to the lack of agreement which can occur from using more than one statistical treatment (Nourbakhsh & Ottenbacher, 1994); furthermore, once the statistical analysis is performed, one remains with the critical subjective decision regarding the clinical relevance of the results.

Levin (1992) recommended visual inspection and analysis for legitimate exploratory-research vehicles, reserving statistical analysis for confirmatory research studies. Visual inspection was chosen for the analysis of results in the present study, an approach supported by McReynolds and Kearns (1983) who argued that results should be easily observable to be clinically significant. Barlow and Hersen (1984) reported that, in analyzing alternating treatment designs using visual inspection, a conservative approach was adopted by most investigators, accepting as significant only clear divergence between conditions. According to Barlow and Hersen, this criterion is met when the two series of data points, one for each condition, are found to be nonoverlapping. These conservative criteria were adopted in the analysis of results for the present study. Three experienced researchers, two authors and one individual independent of the project, inspected the data for treatment effects. A difference between conditions was accepted as significant only if agreement was unanimous.

RESULTS AND DISCUSSION

Table 2 presents the speech recognition score obtained for each subject in each condition for each trial, which in turn is displayed in Figure 1. Based on the strict criteria cited by Barlow and Hersen (1984) and recommended by McReynolds and Kearns (1983), conditions were judged as significantly different in the single-subject analysis only for Subject 15. Agreement was unanimous among judges for all subjects except in the case of Subject 2.

An average word score and an average phoneme score were obtained for each subject from all 200 stimuli in the NU-6 lists. Equation 2 was applied to these data to obtain a j value for each subject. The mean j value across subjects was found to be to 2.2, which is similar to the results found by Boothroyd and Nittrouer (1988). This resulted in the assumption of 55 independent sources of information (25×2.2). A BD chart was generated as described by Thornton and Raffin (1978), based on an n of 55.

Analysis using the BD on the first two data points for each subject revealed

Table 2

Speech Recognition Scores (in Percent Correct) for NU-6 Word Lists With Phoneme Scoring for Each Subject for the National Acoustics Laboratory (NAL) and Noise Reduction (NR) Conditions During Each Test Period

		Test Period			
Subject	**Condition**	**Test 1**	**Test 2**	**Test 3**	**Test 4**
1	NAL	70	68	69	68
	NR	70	72	77	67
2	NAL	68	73	70	74
	NR	60	63	53	77
3	NAL	68	67	64	68
	NR	57	63	64	65
4	NAL	57	67	70	65
	NR	70	65	71	72
5	NAL	31	36	26	41
	NR	12	21	48	29
6	NAL	49	43	70	67
	NR	52	48	51	76
7	NAL	49	45	34	48
	NR	45	39	47	33
8	NAL	80	85	74	73
	NR	81	84	86	84
9	NAL	40	49	31	36
	NR	36	32	37	44
10	NAL	43	45	44	69
	NR	45	45	57	49
11	NAL	42	52	63	52
	NR	37	57	57	61
12	NAL	57	53	65	60
	NR	61	59	68	60
13	NAL	63	56	56	60
	NR	57	56	51	45
14	NAL	45	44	37	44
	NR	41	40	53	33
15	NAL	55	61	51	57
	NR	41	51	44	45
16	NAL	40	47	43	30
	NR	51	39	40	53
17	NAL	70	66	61	77
	NR	73	61	68	55

Continued on next page

Table 2 *continued from previous page*

Subject	Condition	Test Period			
		Test 1	Test 2	Test 3	Test 4
18	NAL	29	40	21	44
	NR	49	35	29	44
19	NAL	55	51	61	74
	NR	60	49	59	56
20	NAL	35	40	39	59
	NR	41	37	37	41

two subjects for whom the two conditions were significantly different (Subjects 5 and 18). In both cases, the line graphs of these two subjects are characterized by overlap and wide intrasubject variability. Subjects whose data showed significant differences based on the binomial model did not have curves with clear divergence in the single-case data, and the subject whose data showed divergence between conditions in the single-case data did not have significantly different re-

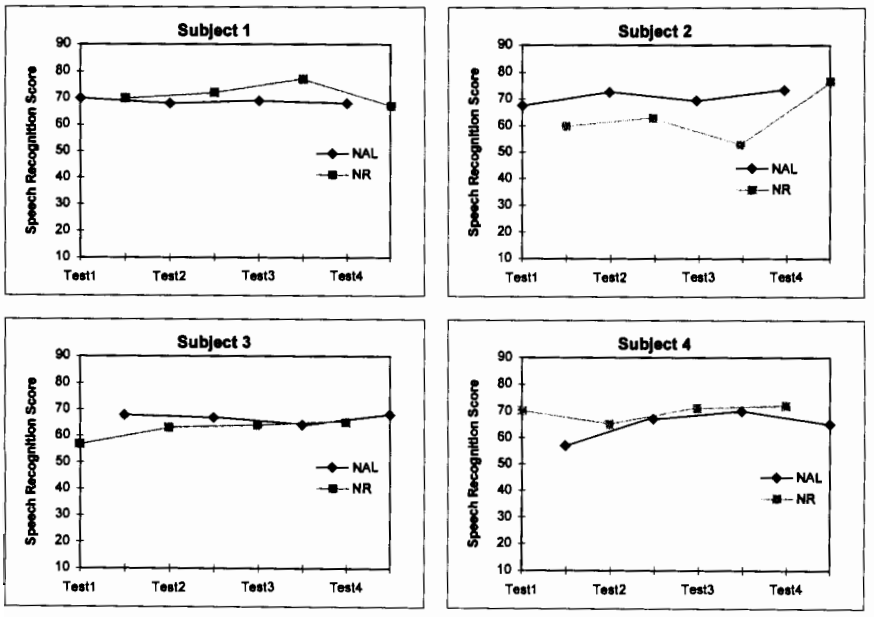
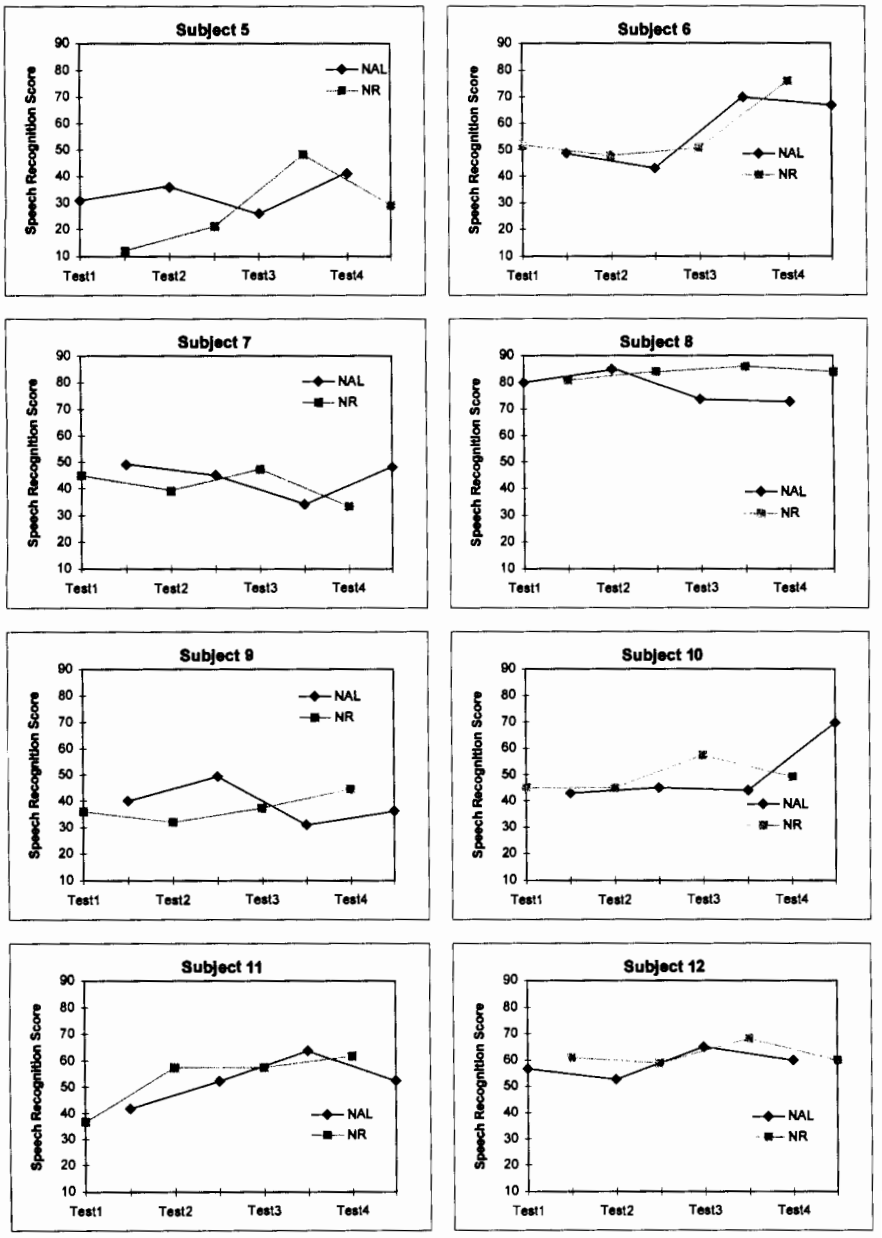


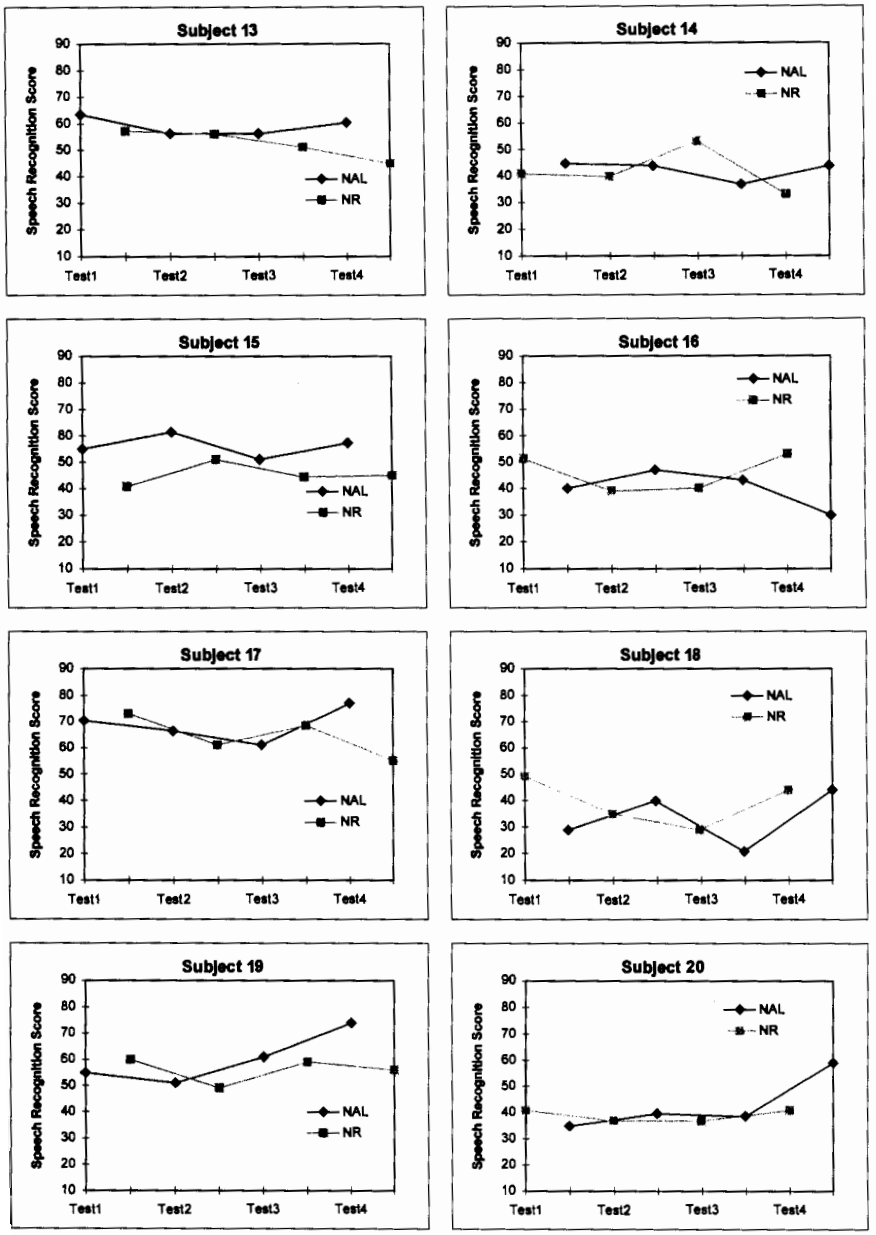
Figure 1. Line graphs representing test results for twenty subjects. Diamonds symbolize speech recognition scores for four presentations using the Revised NAL frequency response. Squares symbolize the speech recognition scores for the noise reduction (NR) frequency response. *(Continued on next page)*

Figure 1 *Continued from previous page*



Continued on next page

Figure 1 *Continued from previous page*



sults based on the binomial model. Therefore, in no case was clear divergence between conditions found in single-case data along with significant differences based on the BD analysis in the same subject.

The present findings suggest that different methods of establishing who benefits from a particular circuit or frequency response may lead to different conclusions. There are a number of possible explanations for the difference in the outcomes. One may be that the BD is based on the stability of a percent-correct score whereas the single-case analysis considers the performance stability of the particular subject (Chmiel & Jerger, 1995). A second possibility, however, is that different outcomes were due to the different number of items used in each procedure. To explore this question, a second BD analysis was performed on all 100 words per condition. As before, the BD analysis was based on 2.2 bits of information per word ($100 \times 2.2 = 220$). In this analysis, results revealed conditions to be significantly different only for Subject 15. It may be recalled that this is the same subject identified by the single-case design evaluation. Since identical outcomes were obtained when different methods were used, it suggests that difference in outcomes from the original analysis was in fact due to number of items used in the evaluation. A final BD analysis using the first 50 words ($50 \times 2.2 = 110$) revealed an intermediary stage in which only the results from Subject 5 demonstrated significant differences.

The results of the present study reveal that the added information on the performance stability for each subject as assessed in the single-case analysis did not change the outcome as compared with the BD when item number was held constant. Whereas in the present study only one subject was identified as being significantly affected by the change in conditions, it would be advisable to replicate this study in the context of a more powerful independent variable in order to determine whether BD and single-case data would identify the same subjects. A study which uses a more powerful independent variable should provide more opportunities to compare the BD and single-case design by increasing the frequency of significant outcomes. In the present study, a subtle difference between conditions had been chosen intentionally to evaluate the possibility that a single-case design might more readily differentiate between two potentially clinically-relevant conditions, although this was not shown to be true. The overall lack of significant findings may be due to insensitivity of the measurements or the fact that the differences in conditions were simply not clinically meaningful.

The use of a more powerful independent variable can also provide an opportunity to further explore the use of 25-word lists. The case against using 25-monosyllabic word lists has been cited frequently in the literature, and was recently reviewed by Wiley et al. (1995). Surveys of clinical practice, however, reveal increases over time in the frequency of use of half lists (Martin et al., 1994; Martin & Forbis, 1978). A frequently cited concern over the use of 25-word lists (Wiley et al., 1995) is the decrease in sensitivity of the test, reducing the likelihood of es-

tablishing true differences in performance (i.e., a Type II error). This is reflected in the use of wider confidence limits in the BD to account for the wider variability with the shorter lists. In the present study, Subject 15 was not identified as performing differently under the two aided conditions until a 100-word list was evaluated. Whereas the findings of one subject are insufficient to draw conclusions, they support further investigation. It would also be interesting to determine if the opposite occurs. In the present study, the data of two subjects which were shown to reveal significantly different scores for the two conditions from the initial (25-word) BD analysis became insignificant when the entire 100-word list was used.

In summary, results of the present study suggest that adding information on the performance stability for each subject (as obtained during the single-case analysis) did not change the outcome as compared with the BD analysis when test item number was 100 words. Using the BD with 25- and 50-word lists, however, did result in other outcomes. These findings should be treated as preliminary and should be replicated in the context of a more powerful independent variable.

ACKNOWLEDGMENTS

This research was supported by a grant from The City University of New York PSC-CUNY Research Award Program.

REFERENCES

Barlow, D.H., & Hersen, M. (1984). *Single case experimental designs: Strategies for studying behavior change* (2nd ed.). New York: Pergamon Press.

Beck, L. (1991). Where do we go from here? In G. Studebaker, F. Bess, & L. Beck (Eds.), *The Vanderbilt hearing aid report II* (pp. 1-9). Parkton, MD: York.

Bentler, R. (1991). Clinical implication and limitations of current noise reduction circuitry. In G. Studebaker, F. Bess, & L. Beck (Eds.), *The Vanderbilt hearing aid report II* (pp. 78-91). Parkton, MD: York.

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

Brainard, S., & Lesner, S. (1992). Assessment of telephone amplifiers using an alternating treatments design. *Journal of the Academy of Rehabilitative Audiology, 25,* 123-129.

Brimacombe, J., Arndt, P., & Staller, S. (1995, May). Multichannel cochlear implants in adults with residual hearing. In *Cochlear implants in adults and children.* Paper presented at the NIH Consensus Development Conference, Bethesda, MD.

Chmiel, R., & Jerger, J. (1995). Quantifying improvement with amplification. *Ear and Hearing, 16,* 166-175.

Dillon, H. (1982). A quantitative examination of the sources of speech discrimination test score variability. *Ear and Hearing, 3,* 5-58.

Dillon, H. (1983). The effect of test difficulty on the sensitivity of speech recognition tests. *Journal of the Acoustical Society of America, 73,* 336-344.

Foust, T., & Wynne, M. (1991). Effectiveness of supplemental parent training in hearing aid checks. *Journal of the Academy of Rehabilitative Audiology, 24,* 85-96.

Hochberg, I., Boothroyd, A., Weiss, M., & Hellman, S. (1992). Effects of noise and noise suppression on speech perception by cochlear implant users. *Ear and Hearing, 13,* 263-271.

Humes, L., Christensen, L., Bess, F., & Hedley-Williams, A. (1997). A comparison of the benefit provided by well-fit linear hearing aids and instruments with automatic reductions of low frequency gain. *Journal of Speech, Language, and Hearing Research, 40,* 666-685.

Kompis, M., & Diller, N. (1994). Noise reduction for hearing aids: Combining directional microphones with an adaptive beamformer. *Journal of the Acoustical Society of America, 96,* 1910-1913.

Kratochwill, T., & Levin, J. (1992). *Single-case research design and analysis.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Kuk, F., & Pape, N. (1992). The reliability of a modified simplex procedure in hearing aid frequency response selection. *Journal of Speech and Hearing Research, 35,* 418-429.

Levin, J. (1992). Single-case research design and analysis: Comments and concerns. In T. Kratochwill & J. Levin (Eds.), *Single-case research design and analysis* (pp. 213-224). Hillsdale, NJ: Lawrence Erlbaum Associates.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *Journal of the Acoustical Society of America, 49,* 467-477.

Martin, F., Armstrong, T., & Champlin, C. (1994). A survey of audiological practices in the United States. *American Journal of Audiology, 3,* 20-26.

Martin, F., & Forbis, N. (1978). The present status of audiometric practice: A follow-up study. *Asha, 20,* 531-541.

McReynolds, L., & Kearns, K. (1983). *Single-subject experimental designs in communicative disorders.* Baltimore: University Park Press.

Nourbakhsh, M., & Ottenbacher, K. (1994). The statistical analysis of single-subject data: A comparative examination. *Physical Therapy, 74,* 768-776.

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

Ottenbacher, K. (1993). Interrater agreement of visual analysis in single-subject decisions: Quantitative review and analysis. *American Association of Mental Retardation, 98,* 135-142.

Parent, T., Chmiel, R., & Jerger, J. (1998). Comparison of performance with frequency transposition hearing aids and conventional hearing aids. *Journal of the American Academy of Audiology, 9,* 67-77.

Sandridge, S., Goldberg, D., & Workman, C. (1994). The effectiveness of acoustically tuned earmolds. *Journal of the Academy of Rehabilitative Audiology, 27,* 61-72.

Stein, L., & Dempsey-Hart, D. (1984). Listener-assessed intelligibility of a hearing aid self-adaptive noise filter. *Ear and Hearing, 5,* 199-204.

Stein, L., McGee, T., & Lewis, P. (1989). Speech recognition measures with noise suppression hearing aids using a single-subject experimental design. *Ear and Hearing, 10,* 375-381.

Tawney, J., & Gast, D. (1984). *Single subject research in special education.* Columbus, OH: Charles E. Merrill Publishing Co.

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

Walden, B., Schwartz, D., Williams, D., Holum-Hardegen, L., & Crowley, J. (1983). Test of the assumptions underlying comparative hearing aid evaluations. *Journal of Speech and Hearing Disorders, 48,* 264-273.

Wiley, T., Stoppenbach, D., Feldhake, L., Moss, K., & Thordardottir, E. (1995). Audiologic practices: What is popular versus what is supported by the evidence. *American Journal of Audiology, 4,* 26-34.

Wolinsky, S. (1986). Clinical assessment of a self-adaptive noise filtering system. *Hearing Journal, 30,* 29-32.