WRITEDOWN VERSUS TALKBACK SCORING AND SCORING BIAS IN SPEECH DISCRIMINATION TESTING

DAVID A. NELSON *and* JOSEPH B. CHAIKLIN

University of Minnesota, Minneapolis, Minnesota

Writedown and talkback responses to 500 Hz low-pass filtered CID W-22 words were obtained from eight listeners, and their talkback responses were scored by eight experienced and eight inexperienced examiners. Four of the experienced and four of the inexperienced examiners monitored at 70 dB SPL; four experienced and four inexperienced examiners monitored at 60 dB SPL. Comparison of talkback discrimination scores (DSs) with corresponding writedown DSs revealed: (1) Inexperienced examiners awarded significantly higher mean talkback DSs than the mean writedown DS; i.e., they showed a mean correct bias. (2) Experienced examiners produced talkback DSs that were not significantly different from the mean writedown DS. (3) Decreasing the monitoring level from 70 to 60 dB SPL increased inexperienced examiners' mean correct bias but the experienced examiners' mean talkback DSs did not change significantly with monitoring level. (4) Inexperienced examiners made more scoring errors than experienced examiners at both monitoring levels. (5) Most examiners in both groups made both correct bias and incorrect bias scoring errors to produce a net effect on the talkback DS. (6) Distributions of DS differences show individual differences between talkback and writedown DSs as large as +16% and −20% and frequent differences of ±6%, even when the mean DS difference between scoring methods was negligible.

Lovrinic, Burgi, and Curry's (1968) recent investigation of five speech discrimination tests led them to suggest, ". . . verbal responses should be abandoned in favor of written responses." They found differences between discrimination scores (DSs) obtained clinically with verbal (talkback) scoring procedures and DSs obtained experimentally from the same subjects with written (writedown) scoring procedures. The DS differences were in excess of 10% in "at least three" of the 30 cases examined, and in those three cases talkback response scores were higher than writedown scores. They concluded, ". . . it seems possible that the tester is more inclined to hear a correct than an incorrect response in questionable instances. . . ." We will use the term *correct scoring bias* to describe this inclination to overestimate correct responses and the term *incorrect scoring bias* to describe the tendency to score correct responses incorrectly.

Lovrinic, Burgi, and Curry are not alone in their observation of correct scoring bias. Merrell and Atkinson (1965) compared DSs obtained with both writedown and talkback scoring procedures. They had a 25-member panel (two

audiologists and 23 inexperienced examiners) score the tape-recorded talkback responses of a 64-year-old male patient who had responded to recordings of CID W-22 test words. They found that DSs obtained with a talkback scoring procedure were an average of 8.88% higher than DSs obtained with a write-down procedure—a result suggesting a correct scoring bias for their predominantly inexperienced scoring panel. An examination of the Merrell and Atkinson data reveals individual talkback-writedown DS differences as great as 20%. Merrell and Atkinson also found fewer scoring errors when their examiners viewed a printed list of the stimulus words during the scoring process.

It is possible that Merrell and Atkinson's 8.88% mean DS difference was influenced by the fact that the majority of their examiners were inexperienced with the scoring procedures and stimulus words, in contrast with the usual clinical situation in which examiners are familiar with stimuli and scoring procedures.

In addtion, Merrell and Atkinson's results may have been influenced by their examiners' monitoring (scoring) levels. Each of their examiners adjusted the monitoring level to his own "comfort level." The resulting levels were not reported, but if some examiners scored the talkback responses at less than optimum listening levels, they might have been more prone to make listening errors, including both correct and incorrect bias errors.

Finally, an emphasis on talkback-writedown mean DS differences in the studies discussed above, may have obscured the relative effects of correct and incorrect scoring bias on individual DSs. To illustrate, if an examiner exhibits correct bias on four words (+8%) and incorrect bias on another four words (−8%) in the same list, the net biasing effect on the talkback DS would be zero because the two types of bias cancel each other. Although the resulting talkback DS is numerically equal to the true DS (i.e. the writedown DS), considerable bias operated to produce the paradoxically correct but invalidly derived DS.

In an attempt to answer some of the questions raised by previous research, the study reported here investigated correct and incorrect scoring bias through comparison of talkback and writedown scoring procedures used by experienced and inexperienced examiners who monitored talkback responses at two different intensity levels.

METHOD

General Plan. Both talkback and writedown responses were obtained from eight normal-hearing young adult listeners who responded to low-pass filtered recordings of CID W-22 word lists. The listeners performed the task commonly performed by patients during speech audiometry. Their talkback responses were stored on magnetic tape.

Eight experienced and eight inexperienced examiners individually scored the tape-recorded talkback responses of the listeners. The resulting talkback DSs derived by the examiners were then compared to the writedown DSs of the listeners. Mean differences between talkback and writedown DSs, differ-

ences between types of scoring bias, and distributions of talkback-writedown DS differences were analyzed.

Listeners' Responses. Tape-recorded dubs of Technisonic Studios disc recordings of CID W-22 lists 1A and 2A (Hirsh et al., 1952) were passed through a 20 dB per octave, 500 Hz, low-pass filter (Peekel, TF823) as they were presented to four male and four female adult (18-26 years) listeners who had normal hearing (15 dB HL or better, ISO 1964) at 500, 1000, and 2000 Hz. Stimuli were presented 40 dB above each listener's 1000 Hz threshold as he sat in a double-wall sound-treated booth (Industrial Acoustics Co., 1202 A). The listeners responded to each of the 50 stimulus words of list 1A or list 2A by writing their responses on a score sheet, and by saying the same word aloud. Four listeners gave the writedown response before the talkback response; two of the four responded to list 1A and two responded to list 2A. The other four listeners gave the talkback response first; two of them responded to list 1A and two to list 2A. The average writedown DS for all the listeners was 49.75% (range = 40% to 62%).

Listeners were asked to clarify spelling errors or illegible writedown responses at the end of each test session. The writedown DS was determined for each listener by comparing writedown responses to a typewritten master list of the W-22 words.

Talkback Response Tapes. The listeners' talkback responses were recorded with an Ampex tape recorder (Model 1460) fed by a condenser microphone (Altec, 150 BR). Each talkback response recorded on the talkback tapes was preceded by an unfiltered recording of the stimulus word to which the listener had responded. Thus, the talkback response tapes later provided a stimulus format similar to the common clinical format in which a clinician (examiner) hears a recorded W-22 stimulus word in his monitor phone and then hears the patient's (listener's) talkback response to the word. The tapes contained a total of 400 stimulus-response pairs—50 pairs from each of the eight listeners.

Average maximum peak levels of the talkback responses were determined with a VU meter (Weston Electric Instrument Corp., N4-862) and a graphic level recorder (Bruel and Kjaer, 2305). The sound pressure level (SPL) of the speech peaks was estimated with a 1000 Hz pure tone adjusted to zero VU.

Examiner Groups. Sixteen normal-hearing (see above) individuals were selected to serve as examiners to score the talkback response tapes. Eight of the examiners were highly experienced with the test words and with talkback scoring (the Experienced group), and eight were completely inexperienced with the words and talkback scoring (the Inexperienced group). The Experienced group consisted of clinical audiologists who ranged from approximately 24 to 34 years of age and had between one and eight years of full-time clinical experience involving frequent speech discrimination testing. The eight Inexperi-

enced examiners were university students who ranged from approximately 19 to 24 years of age.

Each examiner listened individually to the talkback tapes through a single TDH-39 dynamic earphone in an MX-41/AR cushion with a dummy phone and cushion on the opposite ear as he sat in a control room that had an ambient A-weighted sound level of 40 dB and a C-weighted sound level of 58 dB measured with a sound level meter (Bruel and Kjaer, 2203). During a single session, each examiner listened to and scored four of the eight talkback response lists, received a 15-minute break, and then listened to the other four response lists.

The talkback tapes were presented at 70 dB SPL to four members of the Experienced group and to four members of the Inexperienced group, and at 60 dB SPL to the remaining examiners. Thus, the examiners were divided into four groups, each group containing four examiners: Experienced-70, Experienced-60, Inexperienced-70, and Inexperienced-60.

Standard instructions were read to each examiner instructing him to compare each talkback response to its preceding unfiltered W-22 stimulus word and to make a mark for "incorrect" or "correct" on the score sheet which contained blanks numbered from 1-50; stimulus words were not printed on the score sheet. This listening arrangement is analogous to a clinical test situation in which the clinician uses an earphone to monitor the stimuli as the test progresses, but unlike the clinical situation, there was no opportunity for our examiners to use lipreading cues or to request repetition and clarification of talkback responses that were spoken or heard unclearly.

RESULTS

Mean Talkback-Writedown Differences. Mean talkback DSs and talkback-writedown mean DS differences (talkback minus writedown) for the different examiner groups are shown in Table 1. Since each examiner group was treated as a sample from an independent population, a t test for independent means ($df = 24$) was used to assess mean differences.

All talkback-writedown mean DS differences and intercondition differences were small (see Table 1), reaching statistical significance ($p = 0.05$) only for the two Inexperienced examiner groups. The 1.62% talkback-writedown mean DS difference exhibited by the Inexperienced-70 group increased to 4.25% for the Inexperienced-60 group. The 2.63% mean difference between the Inexperienced-60 and Inexperienced-70 groups was also significant ($t = 2.88$). These results indicate that as the monitoring task was made more difficult by decreasing the monitoring level from 70 dB to 60 dB SPL, the talkback-writedown mean DS difference increased, but only for the Inexperienced group.

In addition, Inexperienced examiners awarded a significantly higher mean talkback DS than Experienced examiners at both monitoring levels—2.5% higher at 70 dB ($t = 3.42$), and 5.25% higher at the 60 dB monitoring level ($t = 5.0$).

TABLE 1. Mean talkback discrimination scores (DSs), mean differences between talkback and writedown DSs, standard errors (SE) of the mean differences and t ratios ($df = 24$) obtained from the four examiner groups ($N = 4$ examiners per group).

Examiner Group	*Mean Talkback DS* (%)	*Talkback-Writedown Mean DS Difference* (%)†	*SE Diff.*	t
Inexperienced–70	51.37	1.62	0.49	3.33**
Inexperienced–60	54.00	4.25	0.79	5.52**
Experienced–70	48.88	−0.87	0.54	1.65*
Experienced–60	48.75	−1.00	0.69	1.47*

*$p > 0.05$
**$p < 0.005$
†Mean writedown DS, 49.75%; writedown DS ranged from 40% to 62%.

Examiner Scoring Bias. As indicated earlier, analysis of mean differences between DSs obtained with talkback and writedown scoring procedures may not tell the whole story. Underlying these differences are two types of bias that have an important influence on DSs. The talkback DS reflects the net effect of these two types of bias. For example, consider the correct and incorrect bias of the Inexperienced-60 examiner group as shown in the second row of Table 2. Each of the four Inexperienced-60 examiners scored eight 50-item talkback tapes—a total of 400 talkback responses scored by each examiner, or 1600 talkback responses in all. Of the 1600 talkback responses, 112 were incorrectly scored as correct against the writedown validity criterion (7% correct bias). Similarly 44 of the 1600 talkback responses were correct with the writedown method but incorrect with the talkback method (2.75% incorrect bias). Thus, the Inexperienced-60 group had a total of 156 scoring bias errors; i.e. they made incorrect decisions on a total of 9.75% of the 1600 talkback responses they monitored. Despite the 9.75% total scoring bias, the net mean difference between talkback and writedown DSs was only 4.25%, a discrepancy attributable to the cancelling effects of the two types of scoring bias.

An inspection of Table 2 shows clearly that both correct and incorrect bias were obtained from Experienced as well as Inexperienced examiners. At each monitoring level, Inexperienced examiners produced more total bias (sum of correct and incorrect bias) than Experienced examiners. Figure 1 shows the total bias obtained from each of the examiner groups at each monitoring level.

It can be seen in Figure 1 that, regardless of monitoring level, Inexperienced examiners made consistently more errors than Experienced examiners in making decisions about the correctness of talkback responses. Figure 1 also shows that examiners consistently made more errors at the more difficult 60 dB SPL monitoring level, regardless of their experience.

The effect of the two types of scoring bias on "net scoring bias" is slightly more complicated. It appears from the data in Table 2 that two bias patterns underlie the Inexperienced examiners' net bias: (1) Inexperienced examiners

TABLE 2. Summary of scoring bias obtained from the four examiner groups ($N = 4$ examiners per group). Correct bias is based on the number of responses that were scored as incorrect with writedown scoring but correct with talkback scoring and is stated as a percentage of the total 1600 responses per group. Incorrect bias is based on the number of responses that were scored as correct with writedown scoring but incorrect with talkback, and is also stated as a percentage of 1600 responses per group. Total bias is the sum of both correct and incorrect bias. Net bias is the algebraic difference between correct and incorrect bias.

Examiner Group	*Correct Bias*	*Incorrect Bias*	*Total Bias*	*Net Bias*
Inexperienced–70 dB	4.25% (68)*	2.63% (42)	6.88% (110)	+1.62% (26)
Inexperienced–60 dB	7.00% (112)	2.75% (44)	9.75% (156)	+4.25% (68)
Experienced–70 dB	1.94% (31)	2.81% (45)	4.75% (76)	−0.87% (−14)
Experienced–60 dB	3.31% (53)	4.31% (69)	7.62% (122)	−1.00% (−16)

*Numbers in parentheses are total errors for each group per 1600 responses.

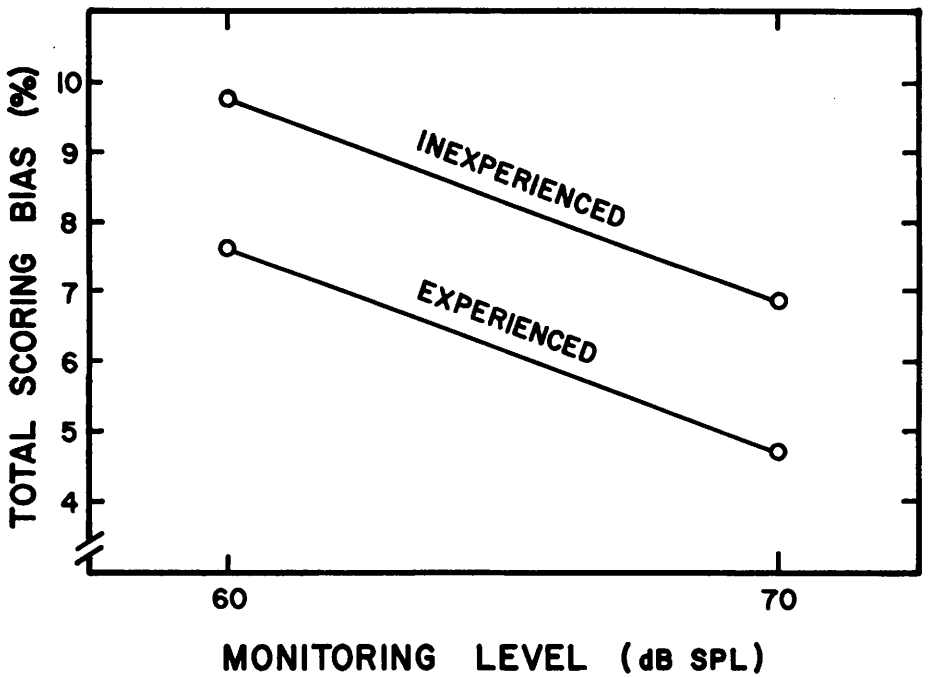

FIGURE 1. Total scoring bias (% of total responses per group) obtained from the Inexperienced and Experienced examiner groups. Total scoring bias is the sum of correct and incorrect scoring bias disregarding the opposite effects of two types of bias have on the talkback discrimination score.

made more correct bias errors than incorrect bias errors—more so at 60 dB than at 70 dB SPL; and (2) Inexperienced examiners' incorrect bias remained small and nearly constant with monitoring level. Since their incorrect bias remained nearly constant with monitoring level, and their correct bias increased at the 60 dB monitoring level, there was proportionately greater net correct bias at the 60 dB level.

On the other hand, Experienced examiners produced nearly equal amounts of correct and incorrect bias at both the 60 dB and 70 dB monitoring levels. The effect of decreasing monitoring level from 70 dB to 60 dB was to increase both types of scoring bias almost equally so that the net scoring bias was very small and nearly identical at both monitoring levels. Thus, the net effect of the 3.31% correct bias and the 4.31% incorrect bias for the Experienced-60 examiners was only −1.0%, and the net effect of the 1.94% correct bias and the 2.81% incorrect bias for the Experienced-70 group was −0.87%.

From the results presented above, one can make inferences only about the types of talkback scoring errors to expect from groups of examiners, rather than from single examiners. In application, one is frequently concerned with a particular examiner and how he scores a particular patient's talkback responses. The following analysis of the distribution of talkback-writedown DS differences obtained by the examiners in this study gives some indication of the discrepancy one might expect between talkback and writedown scoring procedures for a single speech discrimination test.

Distribution of Talkback-Writedown Differences. The frequency histograms in Figure 2 summarize the distributions of talkback-writedown DS differences for each examiner group. Recall that each examiner group produced 32 talkback-writedown difference scores, i.e., the four examiners of each group scored one test from each of eight listeners. A positive sign on the abscissa indicates that the talkback DS was higher than the writedown DS and that there was a net correct scoring bias. A negative sign indicates a net incorrect scoring bias. For example, Figure 2 shows that the Experienced-70 group yielded 0% talkback-writedown difference on only 7 of the 32 tests they scored, and the Inexperienced-70 group yielded 0% difference on 5 of the 32 tests they scored.

There is considerable spread in the talkback-writedown DS differences, especially for the examiners who monitored at the most difficult monitoring level. The range of differences was 18% for examiners who monitored at 70 dB SPL and 26% for examiners who monitored at 60 dB SPL. Interestingly, the ranges were identical for the Inexperienced and Experienced groups. Note that all examiner groups, even the Experienced groups had talkback-writdeown differences greater than ±10% and that differences of ±6% were common even though the mean talkback-writedown difference was relatively small.

The talkback-writedown differences were distributed among all of the examiners in each examiner group, and were not simply the product of only one or two examiners consistently making large scoring bias errors. For example, the thirteen +6% talkback-writedown differences produced by the Inexperienced-

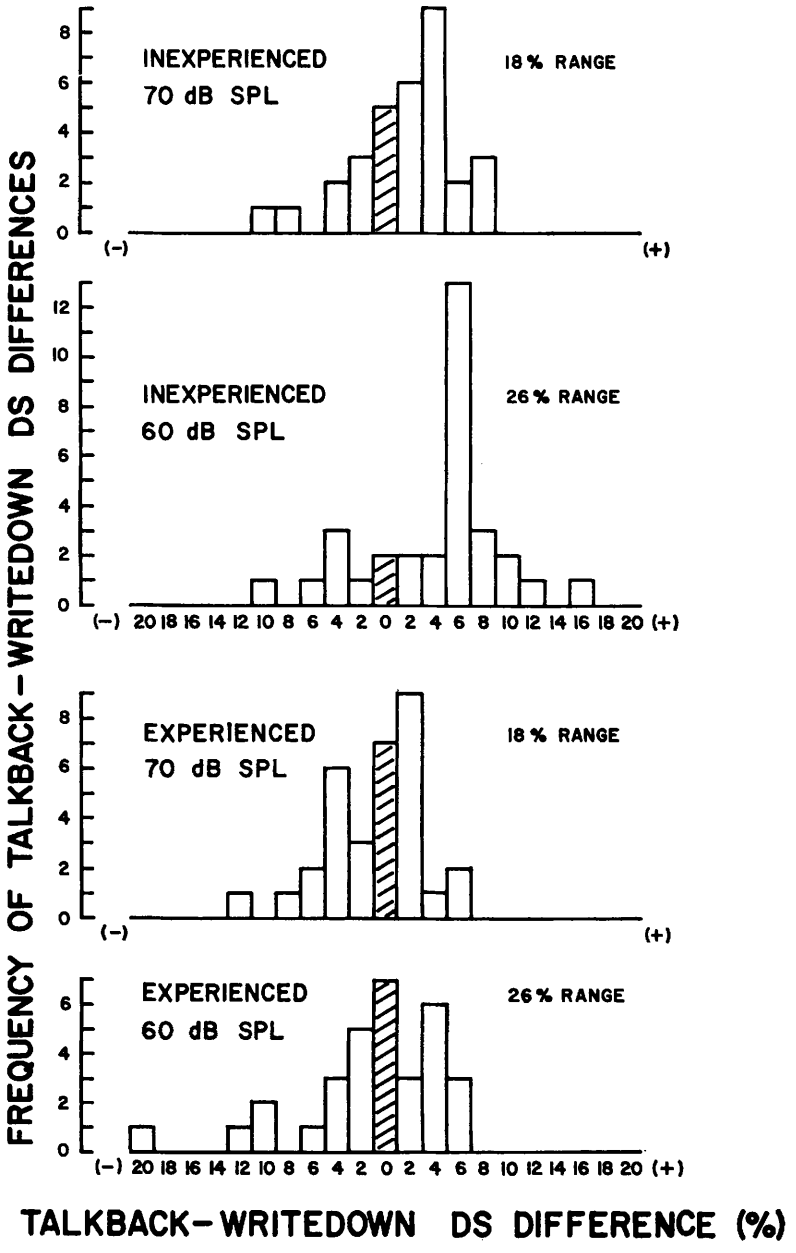

FIGURE 2. Distributions of talkback-writedown discrimination-score (DS) differences. A DS difference is the difference between a DS obtained with the talkback scoring procedure and the corresponding DS obtained with the writedown procedure. Each of four examiners in a group scored talkback responses from each of eight listeners to obtain a total of 32 possible DS differences per examiner group. A positive sign on the abscissa indicates that the talkback DS was higher than the writedown DS (the net correct scoring bias); a negative sign on the abscissa indicates net incorrect scoring bias. The range of talkback-writedown differences is shown for each examiner group.

60 group were distributed among the four examiners as follows: 3, 4, 2, 4; the nine +4% talkback-writedown differences produced by the Inexperienced-70 group were distributed: 3, 2, 2, 2.

DISCUSSION

The mean differences between talkback and writedown DSs in this study (−0.87% to 4.25%) are smaller than the 8.88% mean difference reported by Merrell and Atkinson (1965). Our results indicate that differences in monitoring level may account for a portion of the interstudy difference; hence it seems reasonable to conjecture that other sources of increased discrimination-task difficulty, such as a patient's poor articulation, or noise in a talkback monitoring system, might increase the inexperienced examiners' talkback-writedown DS differences in the same direction.

Our results support Merrell and Atkinson's finding that correct bias occurs in speech discrimination testing when a talkback scoring procedure is used. While Merrell and Atkinson reported that their inexperienced examiners did not perform differently from their experienced examiners, we found that correct bias occurred more often and to a greater extent with our inexperienced examiners. However, since experienced examiners usually administer clinical discrimination tests, the preponderance of correct bias shown by our inexperienced examiners probably does not constitute sufficient evidence for complete abandonment of talkback scoring procedures. In addition, we found little *average* difference between talkback and writedown results when our experienced examiners scored talkback responses in relatively ideal acoustic conditions (a monitoring system with a high signal-to-noise ratio, and stimuli at or above 70 dB SPL).

But one should not conclude that a single talkback DS will not be biased when an experienced examiner scores the talkback responses, even under ideal acoustic monitoring conditions. The distributions of talkback-writedown differences shown in Figure 2 illustrate the difficulty of generalization to an individual talkback DS. Some DSs were influenced by a preponderance of either correct or incorrect bias, occasionally as great as 16% or 20% and frequently between 4% and 10%, even in the Experienced group. In a clinical setting, bias of this magnitude is a serious source of error to add to the inherent unreliability of the test itself.

On the basis of the distributions of talkback-writedown differences discussed here, one may conclude that there is danger of clinically significant scoring bias even with experienced examiners. The range of net scoring bias is large enough to cause serious concern about the clinical use of talkback responses.

An examiner probably can minimize the effects of talkback scoring bias if he requests patients to repeat, spell, or clarify in some other manner, all talkback responses that sound even slightly ambiguous. The poorer a patient's articulation, the more urgent the need for caution with the talkback method. On the other hand, effective use of clarification through repetition presumes an exam-

iner who has a keen respect for test validity, a strong appreciation of the fallibility of his own perceptions, and more than average patience. The need for repetition may be reduced by eliminating examiner distractions, favorable placement of the patient's microphone, adequate monitoring levels, and by watching the patient's face as he responds.

A factor not investigated in this study is the effect of the absolute value of the writedown DS on scoring bias. If an examiner has a tendency to give the patient the benefit of the doubt (correct scoring bias), he has an opportunity to do this only on those talkback responses which are incorrect by the writedown scoring procedure. As the writedown DS increases from 50% to 90% for a 50-item test, the number of talkback responses on which the examiner has an opportunity to make a correct bias error decreases from 25 to 5. Similarly, the number of talkback responses on which he has an opportunity to make an incorrect bias error increases from 25 to 45. If one assumes that his inherent tendency to make one type of bias error as opposed to another remains constant, i.e., that his criterion for making the correctness decision about the talkback response does not change, then the relative amounts of correct and incorrect bias obtained might change markedly depending upon the absolute value of the writedown DS. The mean writedown DS in this study was 49.75%, as close to 50% as we could obtain. We suggest that further investigation is needed to determine the constancy of examiner bias for different values of writedown DSs.

With the evidence at hand, the talkback versus writedown scoring controversy is not completely settled. We have shown examiner experience and monitoring level to be critical variables. Merrell and Atkinson have shown that mean scoring bias is influenced by the examiner having the test words in front of him while he scores TB responses. Visual cues (lipreading) may also be valuable in reducing testing bias, particularly when the acoustic monitoring conditions are poor. These variables should be, but sometimes are not, accounted for in a clinical setting.

ACKNOWLEDGMENT

This article is based on a master's project completed by Nelson under the direction of Chaiklin at the University of Minnesota, Department of Speech Science, Speech Pathology and Audiology.

REFERENCES

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

Lovrinic, J. H., Burgi, E. J., and Curry, E. T., A comparative evaluation of five speech discrimination measures. *J. Speech Hearing Res.*, **11**, 372-381 (1968).

Merrell, H. B., and Atkinson, C. J. The effect of selected variables upon discrimination scores. *J. aud. Res.*, **5**, 285-292 (1965).

Received April 18, 1969.