ARTICULATION JUDGMENTS: SOME PERCEPTUAL CONSIDERATIONS

LAWRENCE D. SHRIBERG

University of Wisconsin, Madison, Wisconsin

Five experienced articulation judges scored the tape-recorded responses of four children with articulatorily deviant /r/s under two response arrangement conditions. The tape conditions contrasted intact list scoring with a dubbed arrangement to investigate perceptual bias. The "intact" tape condition presented allophones of the target phoneme which varied from item to item, as the responses were originally obtained; the "dubbed" tape condition rearranged these responses so that responses to similar stimuli occurred successively. The /r/ phoneme data indicated significant interactions among the tape conditions, allophone error-type, and severity of error. Implications for reliability training and sampling are discussed.

In a review of articulation testing, Noll (1970) has correctly noted that a "test" of articulation refers to an evaluative process carried out by a listener. Although lists of stimuli may evoke those articulatory behaviors of importance to a particular research or clinical interest, the test of any response invokes a judge's reference to an internalized "phonemic space." A substantial body of studies has attempted to isolate the many stimulus, response, and methodological variables that affect both the validity and reliability of such judgments (see reviews by Shine, 1969; Winitz, 1969). This report considers some additional variables that may affect a judge's perceptual competence.

Current programs for articulatory acquisition typically include serial presentations of word lists which sample a target phoneme in different phonetic contexts. Although some evidence suggests that children generate different test performances and acquisition curves depending on the allophone type (Curtis and Hardy, 1959; Diedrich, 1971; Perkins, 1952; Sommers, Leiss, and Gerber, 1969), response data are typically plotted only by the total number or percent correct for each word list administration. At least two sources of bias are possible in scoring the responses to intact lists.

One source of bias, whether scoring live or from a tape recording, involves "errors of habituation" (Manning and Rosenstock, 1968). The tendency to develop a perceptual disposition to a series of stimuli has been demonstrated in perceptual scaling tasks (Manning and Rosenstock, 1968) and is well known to the clinician and reliability judge who attempt to keep successive articulatory judgments independent. Even when a judge is kept unaware of the therapy status of a child, scoring a succession of baseline responses or post-

management probes would assumedly maximize conditions for a persistence effect of "incorrect" or "correct," respectively.

A second, and perhaps more subtle, source of unreliability considers lists item arrangement. Since differing allophones of a target phoneme are typically arranged randomly in each list, judges must continually shift their allophonic standard with each item. Thus, both the standard and the comparative stimuli are varying as a judge proceeds through the list, a condition which, a priori, should increase the difficulty of a perceptual task.

Moreover, if certain immediate or broad phonetic contexts (McDonald, 1964) do facilitate production for a given child so that list acquisition proceeds reliably by an additivity of correctly articulated within-list responses, the measurement, that is, articulation judgment, should be made at the test item level, not the list level. That is, the metric should be applied at the most sensitive level of the treatment effect, item-by-item, much the same as a physician would hold up a series of pre- and post-spot x-rays for contemporaneous comparison.

The purpose of this investigation, therefore, was twofold. The first goal was to determine if there are differences in reliability of total-score judgments (see Siegel, 1962) when judges score intact versus specially dubbed articulation tasks. The second purpose was to determine if there are any differences specific to particular phonemes or allophones of a phoneme.

METHOD

Subject Tapes

In a previous study (Shriberg, 1971), 48 articulatory deviant first- and second-grade children were tested on 30-item Sound Production Tasks (Elbert, Shelton, and Arndt, 1967) once a week for three consecutive weeks. Each 30-item imitative task contained one target phoneme in varying phonetic environments in syllables, words, and sentences. Three randomized lists of the same 30 items were used for each target phoneme. For the present study, the tapes of four children with a deviant /r/ and four with a deviant /s/ were selected from the 48 available tapes. These tapes contained responses to three /r/ lists and three /s/ lists, respectively. Previous judges had scored the responses of these children as variable across the three testing sessions; intrasubject variability ranged frome a one- to nine-item difference in total correct between any two of the three test sessions. Thus, these subjects presented differing responses (correct versus incorrect) to similar items. Additionally, an attempt was made to choose children who averaged 15 correct on the 30-item tasks, that is, 50% correct, to maximize independence of judgments from item to item.

The test responses of each subject were dubbed onto 16 master tapes to comprise two judging conditions for each subject's responses.

Intact Condition. The intact tape condition was simply a 90-item task con-

sisting of the three intact 30-item lists. The ordering of the three lists was randomly rearranged for each master tape with reference to the date of original test administration. The mean pause time between each item was five seconds, with 15 seconds between lists.

Dubbed Condition. The dubbed condition presented the same 90 responses for each subject; however, the three responses to a given item, originally obtained one or two weeks apart, were grouped sequentially. Thus the first three items for the /r/ task were: "He can ride a bike" (five-second pause), "He can ride a bike" (five-second pause), "He can ride a bike," followed by an eight-second pause and the next set of three responses. The order of dubbed responses within each set was also randomized with reference to the date of original test administration.

Instrumentation

The original recordings were made on high quality tape on an Ampex 601 audio tape recorder at a tape speed of 7.5 ips. The transducer was an Electro Voice 664 microphone, positioned four to 10 inches from a child's lips. Dubbing and editing of the second generation tapes were accomplished by feeding the Ampex 601 into an Ampex AG350.

Procedures

Five female speech clinicians with considerable experience in listening studies served as judges. In keeping with a clinical paradigm, they were instructed to score each articulation response as correct or incorrect, using the response definition, "If this child were in speech therapy, how would you score his response?"

Four one-hour judging sessions, three to five days apart, were held in a listening laboratory. Subject tapes were presented via two Rheem Califone Model 770X SS audio tape recorders which fed five listening booths equipped with Koss Pro-4A headphones. The 16 subject tapes were randomized and presented in an order which alternated intact and dubbed conditions. Judges scored five or six 90-item tasks at each session; each judge heard each subject's responses under the two tape conditions at some time within the four listening sessions. Each 90-item task was numbered, rather than identified by subject. The responses of three of the subjects, one /s/ and two /r/, were presented twice in each tape condition to determine intrajudge reliability.

RESULTS

The reliability formula used to calculate both intrajudge and interjudge item-by-item agreement was: percentage of agreement = (number of agreements)/(number of items judged). Thus, interjudge agreement is reported as

the mean percentage of agreement derived from each pair-wise comparison of all judges, while the mean intrajudge agreement represents the mean derived from each judge's agreement with herself.

Preliminary review of the percentage of interjudge agreement across both tape conditions indicated generally poor agreement for /s/ judgments, ranging from 44-94%, with a mean of 69.7%. Given the difficulty in judging some types of /s/ errors from audio tapes without extensive response definition training, such difficulties with this sound were not wholly unexpected. Because of some of the spuriously low mean interjudge agreements in both tape conditions, only the /r/ data were subjected to statistical analyses.

TABLE 1. Descriptive data for the four deviant /r/ subjects.

Dependent Variable	*Subject* 1	2	3	4
Mean intrajudge percentage of agreement	85	86	–	–
Total percent correct	79	82	74	31
Percent prevocalic /r/ items correct	91	84	83	22
Percent postvocalic /r/ items correct	78	92	76	44
Mean interjudge percentage of agreement	79	81	79	70
Mean agreement prevocalic /r/	75	76	79	73
Mean agreement postvocalic /r/	83	86	78	67
Mean interjudge agreement–intact condition	83	77	81	70
Mean interjudge agreement–dubbed condition	75	85	76	70

Table 1 presents the descriptive data for the four /r/ deviant children. Interjudge agreement for /r/ judgments across tape conditions ranged from 70-85%, with a mean of 79.1%. Three-way completely repeated analyses of variance (Weiner, 1962, p. 298), tape condition (intact versus dubbed) × subject (Subject Tapes 1-4) × allophone type (prevocalic /r/ versus postvocalic /r/) were performed individually for the dependent variables of total number correct, number of intrajudge agreements, and number of interjudge agreements. To accomplish the allophone-type analyses, the 30-item lists were divided into 14 prevocalic /r/ and 14 postvocalic /r/ items and the remaining two items from each list were discarded.

The overall mean intrajudge percentage of agreement was 85.5 As a group, each judge's agreement with herself was not significantly different for the two tape conditions, nor did judgments vary significantly by allophone type, children, or any interactions among these parameters.

Analyses of judgments of total number correct indicated that judges found differences among the four children ($F = 54.39$; $df = 3, 12$; $p < 0.001$). Subsequent analyses of simple main effects (Weiner, 1962, p. 232) indicated that judges scored Subject 4 significantly lower than the other three children. This child averaged 31% correct, whereas the mean of the other three children's

average scores was 79.9% correct. The subject tapes also differed on the number of prevocalic /r/ versus postvocalic /r/ items correct ($F = 13.67$; $df = 3, 12$; $p < 0.001$). Separate analyses indicated that this difference was significant for two subjects. Subject 1 was scored as having significantly more prevocalic than postvocalic /r/ items correct, and Subject 4 had significantly more postvocalic /r/ items correct.

The analysis of interjudge agreement indicated that it was significantly higher for the postvocalic than the prevocalic /r/ ($F = 13.36$; $df = 1, 4$; $p < 0.025$). Further, there were significant differences among subjects ($F = 15.29$; $df = 3, 12$; $p < 0.001$). Specifically, interjudge agreement was lower for Subject 4 than for the other three subjects. In addition, there was a significant interaction between allophones and subjects ($F = 10.31$; $df = 3, 12$; $p < 0.005$). The basis of this intereaction of allophones-by-subjects seemed to be that there was little difference among subjects for the prevocalic /r/, while there were large differences among subjects for the postvocalic /r/. Finally, the interaction of tape conditions (intact versus dubbed) and subjects was also significant ($F = 11.10$; $df = 3, 12$; $p < 0.001$). As can be seen from Table 1, interjudge agreement for Subjects 1 and 3 was significantly higher for the intact condition than for the dubbed condition. For Subject 2, on the other hand, interjudge agreement was significantly higher for the dubbed than for the intact condition.

DISCUSSION

Failure to find main effects for tape conditions and allophone type may have been determined, in part, by the distance from the 50% point of the obtained scores. That is, the higher total correct scores for three subjects may have limited the sensitivity to true effects. These scores were higher than expected in relation to the criteria for choosing these tapes; differences may have been due to a more liberal response definition in this study than was used in the Shriberg (1971) study, as well as several differences in judging procedures.

However, the significant interactions of subjects with tape conditions and allophone-type warrants attention. Differing articulatory characteristics of each of the children mediated the significant differences in interjudge agreement. The two children for whom judges had poorer agreement in the dubbed condition than in the intact condition were also the two children who had more postvocalic /r/ items scored as incorrect, although the difference in number correct for each allophone type reached significance for only one child. One interpretation, consistent with the trend of the data, is that the dubbed condition, in which responses to a given stimulus item were grouped, forced more sensitive attention to approximations to the target phoneme. This condition appeared to have had its greatest effects on judgments of postvocalic /r/ items, where /r/ coloring typically has wider perceptual boundaries than those of the usual w/r prevocalic substitution. Note that the high postvocalic /r/ correct child obtained greater agreement in the dubbed condition. In summary, the

probability of judges disagreeing was greatest when the perceptual task had a relatively high frequency of occurrence of incorrect postvocalic /r/ items, and the stimulus arrangement (dubbed) forced sensitive comparison among responses to allophonically similar test items.

Within the experimental constraints of this study and specific to the /r/ phoneme, the intrajudge agreement data indicate that a clinician's reference to her own phonemic competence is equally stable, whether judging several responses to a particular item, or responses with successively changing allophones. However, it should be noted that this perceptual stability may be influenced when clinicians are actually conducting a management program. Diedrich and Irwin (1970) and Peterson[1] report differences in both intra- and interjudge agreement which are apparently related to client-clinician relationships in the therapy process. The judges in the present study did not have personal investment in a child or his therapy progress; evidently the added role of "clinician," which necessitates on-line decisions about which responses to reinforce, adds yet another dimension of influence to intrajudge reliability assessment.

In consideration of these issues and in light of this study's finding that interjudge agreement on this phoneme appears to be very sensitive to individual differences in a child's allophone-level frequency of misarticulation, three suggestions for sampling intrajudge and interjudge reliability are offered.

First, for sampling either form of judge reliability, it seems imperative that researchers use subjects whose overall correct score on any phoneme or allophonic type approaches the 50% point across items. At this point item judgments will be maximally independent and free from a perceptual bias which, in scoring very low or very high correct subjects, will yield high reliability, but possibly poor validity. Alternately, if very low or high correct subjects need to be reported, researchers should present the total correct scores for each subject, as well as separate agreement figures for both correct and incorrect item-by-item judgments.

Second, at present, the only approach to train clinicians or judges to perceptual competence is to employ training tapes scored by "experts." These data suggest that such training samples might be constructed to yield subtotals by allophone type. Requiring a clinician or potential judge to reach a criterion level of agreement on each allophone type should decrease the possibility of perceptual bias due to allophone-level disagreement.

Finally, whether or not calibration is achieved via such training tapes, it might be prudent for researchers to obtain a brief intra- and interjudge reliability statement for each /r/ child involved in some intensive training program. For optimum sensitivity, a sample of responses to a given item should be edited from stages in the program for contemporaneous comparison. Such reliability checks may be useful in parcelling out the variance in articulation-modification programs due to incorrect contingency management by "partially" unreliable clinicians.

[1]D. D. Peterson, personal communication.

ACKNOWLEDGMENT

This study was supported in part by the Bureau of Child Research, Lawrence, Kansas. The author wishes to thank John F. Michel, Deedra Engemann, Roberta Everslage, Nancy Eyre, Linda Few, Joan Gentry, and Robert Hoekenga for their time and capable assistance. Requests for reprints may be addressed to Lawrence D. Shriberg, Department of Communicative Disorders, University of Wisconsin-Madison, 1975 Willow Drive, Madison, Wisconsin 53706.

REFERENCES

CURTIS, J. F., and HARDY, J. C., A phonetic study of misarticulation of /r/. *J. Speech Hearing Res.*, **2**, 244-257 (1959).

DIEDRICH, W. M., Training speech clinicians in the recording and analyses of articulatory behavior. Summary Report, Year 2, U.S. Office of Education Grant OEG-0-9-261293-3406(031), December 1 (1971).

DIEDRICH, W. M., and IRWIN, J. V., Training speech clinicians in the recording and analysis of articulatory behavior. Summary report, Year 1, U.S. Office of Education Grant OEG-0-9-261293-3406(031), November 13 (1970).

ELBERT, MARY, SHELTON, R. L., and ARNDT, W. B., A task for the evaluation of articulation change: 1. Development of methodology. *J. Speech Hearing Res.*, **10**, 281-288 (1967).

MANNING, S. A., and ROSENSTOCK, E. H., *Classical Psychophysics and Scaling*. New York: McGraw-Hill (1968).

MCDONALD, E. T., *Articulation Testing and Treatment: A Sensory-Motor Approach*. Pittsburgh: Stanwix House (1964).

NOLL, J. D., Articulation assessment. In J. Fricke (Ed.), *Speech and the Dentofacial Complex: The State of the Art, ASHA Reports Number 5*. Washington, D.C.: American Speech and Hearing Association (October 1970).

PERKINS, W. H., Methods and materials for testing articulation of /s/ and /z/. *Quart. J. Speech*, **38**, 57-62 (1952).

SHINE, R. E., The influence of selected phonological variables on the consistency of intrajudge and interjudge evaluations of articulation. Doctoral dissertation. Pennsylvania State Univ. (1969).

SHRIBERG, L. D., The effect of examiner social behavior on children's articulation test performance. *J. Speech Hearing Res.*, **14**, 659-672 (1971).

SIEGEL, G. M., Experienced and inexperienced articulation examiners. *J. Speech Hearing Dis.*, **27**, 28-35 (1962).

SOMMERS, R. K., LEISS, R., and GERBER, ADELE, Criteria for dismissal of /r/ defective children from therapy. Paper presented at the Annual Convention of the American Speech and Hearing Association, Chicago (1969).

WEINER, B. J., *Statistical Principles in Experimental Design*. New York: McGraw-Hill (1962).

WINITZ, H., *Articulatory Acquisition and Behavior*. New York: Appleton-Century-Crofts (1969).

Received June 22, 1971.

Articulation Judgments: Some Perceptual Considerations

Lawrence D. Shriberg
J Speech Hear Res 1972;15;876-882

This information is current as of June 19, 2012

This article, along with updated information and services, is
located on the World Wide Web at:
http://jslhr.asha.org/cgi/content/abstract/15/4/876

AMERICAN
SPEECH-LANGUAGE-
HEARING
ASSOCIATION