1965

Psychological Scaling of Language Development of Children

Dorothy Sherman

Thomas Shriner

Franklin Silverman

Such research may very well reveal hitherto unsuspected relationships of these factors—birth order, family size, and sex of siblings—to certain traits and interests among siblings in various family constellations.

Literature Cited

Cohen, F. "Psychological Characteristics of the Second Child as Compared with the First," *Indian Journal of Psychology*, XXVI (1951), 79-84.

Kimble, Gregory A., and Norman Garmezy. *Principles of General Psychology*. New York: The Ronald Press Company, 1963.

Koch, Helen L. "Some Emotional Attitudes of the Young Child in Relation to Characteristics of His Sibling," *Child Development*, XXVII (1956).

Lindquist, E. F. *Design and Analysis of Experiments in Psychology and Education*. Boston: Houghton Mifflin Company, 1953.

Schacter, S. *Psychology of Affiliation*. Stanford, Calif.: Stanford University Press, 1959.

Schooler, Carmi. "Birth Order and Schizophrenia," *Archives of General Psychiatry*, IV (1961), 91-97.

Psychological Scaling of Language Development of Children

DOROTHY SHERMAN, THOMAS SHRINER and FRANKLIN SILVERMAN

Abstract: Certain aspects of the use of psychological rating scale methods for measuring degrees of language development in the speech of children are evaluated. That typed samples from children's speech can be scaled reliably is demonstrated. Comparisons are made among correlation coefficients which were obtained for the purpose of estimating relationships among three measures of language development for the same set of 50 samples of children's language: structural complexity scores obtained by analysis of the samples; scale values of intricacy of language usage obtained by the psychological scaling method of Equal-Appearing Intervals; and mean estimates of age derived from sophisticated observers' judgments. The conclusion was drawn that psychological scaling of various aspects of children's language could provide new and useful tools for the study of and the assessment of children's language development.

The basic problem is to evaluate certain aspects of the use of psychological rating-scale methods for the purpose of measuring the degrees of language development exhibited in samples of children's speech. Both for experimental and clinical purposes a method for assessing children's language development is often needed by those who are concerned with speech pathologies. Presently, however, no single measure has been used which appears to be completely satisfactory for this purpose.

Indices which have been used have been based upon various single aspects of language which reasonably might be expected to change with increasing age. The ones primarily used have been these: mean length of response, mean of the five longest responses, number of one-word responses, standard deviation of response length, number of different words, structural complexity score, and the ratio of the different words over the total number of words (Johnson, Darley, and Spriestersbach, 1963). Each of these measures is based upon an analysis of 50 oral responses which have been obtained from a child in a standard manner (Winitz, 1959).

The above-named indices have often been used under the assumption that they provide a satisfactory basis for evaluating language development. These indices, although they apparently have been useful, have not been established as valid for their intended purpose; they have not been studied in relation to any outside criterion. The one obviously useful outside criterion is the impression language makes upon others. Psychological rating-scale methods thus might provide measures useful for evaluation of the validity of the indices currently used; they also might provide, in the form of speech samples scaled for degree of language development, a tool useful for evolving new indices which not only would be more valid than are those previously named but also would be more reliable. Satisfactory temporal reliability, that is, adequacy of sampling in obtaining the 50 responses from which the currently used indices are derived, has been questioned (Minifie, Darley, and Sherman, 1963).

The present experiment was designed for two purposes: first, to determine whether given samples of children's language will be consistently judged to display greater or lesser degrees of language development than will certain other samples; and, second, to evaluate the validity of a frequently used measure, the structural complexity score, by estimating the relationships, for the same samples, among three sets of measures: structural complexity scores, scale values of intricacy of language usage derived from observers' judgments, and mean estimates of age derived from sophisticated observers' judgments.

<div align="center">PROCEDURE</div>

Language Samples

The language samples to be scaled were prepared for presentation to the observers in typed, mimeographed form. This method of presentation was chosen mainly to eliminate the influence of certain irrelevant cues which might operate if

observers were to make their ratings on samples presented by tape recordings. It seems likely that variables such as pitch usage, rhythm, and articulation skill might seriously contaminate the desired responses.

Transcripts of tape-recorded language samples from the speech of 96 children, 24 boys and 24 girls within two months of the age of five and one-half years, and 24 boys and 24 girls within two months of the age of eight years, were available from a previous experiment (Minifie, Darley, and Sherman, 1963). These samples, each consisting of 50 verbal responses, had been elicited from the children in response to Children's Apperception Test Cards.

From this larger pool of 50-response samples, 25 were chosen at random for experimentation aimed at determining the feasibility of employing the psychological scaling method of Equal-Appearing Intervals to obtain data for assessing language development by means of median scale values derived in the way described by Thurstone and Chave (1929). Each of these 25 samples was taken from a longer 50-response sample and consisted of the first 150 words to the nearest complete response.

Another set of 50 speech samples to be scaled with reference to some aspect of language consisted of portions of 50 of the longer 50-response samples. For the purpose of minimizing irrelevant influences upon observers' rating it seemed desirable to keep the topic constant for all samples; and, for this reason, the 50 portions consisted of the verbal output of each of 50 children in response to the same stimulus card. One result of this limitation was variation in lengths of samples to be rated, a result considered desirable because of the possibility that amount of verbal output may be an important and relevant factor with reference to certain aspects of language usage if measures are intended to reflect impressions characteristic of a true situation of communication. These 50 language samples were the experimental stimuli used for obtaining scale values of intricacy of language usage and also for obtaining estimates of chronological age.

The longer 50-response samples, portions of which were used for scaling intricacy of language usage, were analyzed to obtain corresponding structural complexity scores for each of the "intricacy" scale values. The structural complexity measures were derived by the usual method of assigning a weight of 0, 1, 2, 3, or 4 to each response according to differing classifications, such as "functionally complete but structurally incomplete", "simple sentence without phrase(s)", "simple sentence with phrase(s)" with several subcategories, "compound sentence" with subcate-

gories, and "complex sentence" with subcategories. Exact instructions for the usual method of deriving these scores have been reported by Johnson, Darley, and Spriestersbach (1963).

Rating of Language Samples

To obtain data for evaluation of the reliability of scale values of language development obtained from observers' responses to typed language samples, 39 students of speech pathology were instructed to respond to 25 language samples by rating them on a 7-point equal-appearing-intervals scale extending from *1* for least language development to *7* for most language development. After an interval of eight to ten days the same 25 samples arranged in a new random order were again presented to each of the 39 students and the rating procedure was repeated.

An additional 29 students rated 50 samples on intricacy of language usage. None of these students had had any extensive course work in the language development of children; the purpose was to avoid the possibility that rating would be influenced by prior training in the use of the structural complexity score and its derivation. Observers were instructed to respond by the procedure already described. Since it was necessary for the purposes of the experiment to avoid artificial weighting of various factors which might influence the observers' responses, the following definition and instructions were included in "Instructions to Observers":

"Intricacy of language usage, for the purposes of this experiment, is defined as the intricacy of the arrangement of words for the purpose of conveying information. For example, consider the following four sets of words, which, without reference to the specific meanings, might be judged to vary with respect to intricacy of language usage as here defined:

    a) two good little boys
    b) boys in our school
    c) boys who are orphans
    d) really good little boys

Although each of the above sets contains four words, it is obvious that they vary with respect to type of arrangement of words for the purpose of conveying information.

"Make your judgment on the basis of the whole sample. Avoid being influenced by grammatical correctness; for example, 'we was' and 'we were' do not differ with respect to the intricacy of word arrangement. Also, do *not* give a rating based upon a judgment of the extent of vocabulary; for example, 'big size' and 'extensive area' are equivalent as far as the intricacy of arrangement is concerned, but they probably would not be considered

equivalent if judged for the purpose of rating extent of vocabu-
lary."

Five sophisticated judges estimated the age of the 50 children
from whom the 50 samples were elicited. All had had extensive
associations with children of elementary school age. One was a
supervisor of practice teaching, another was a country supervisor
in education work, and the other three were elementary school
principals. They were instructed to assume that each language
sample was selected from the speech of a child with average
intelligence and with average home environment. The experi-
mental task consisted of recording for each of the samples an
estimate of the age of the child who had spoken.

RESULTS AND DISCUSSION

The two sets of median scale values of degree of language
development for 25 samples rated by 39 observers on two trials
separated by eight to ten days are closely related. The Pearson
r for estimating the relationship is .96. For the first trial a Q-
value, the semi-interquartile range, which is a measure of the
dispersion or scatter of judgments, was calculated for each of
the 25 samples. These values were satisfactorily small, with a
range from .47 to 1.02 and a mean of .78. Two additional sets
of scale values were derived for the first trial by randomly
assigning the 39 observers, and the corresponding raw data, to
two groups. The Pearson r obtained for estimating relationship
between these two additional sets of scale values is .90. Thus,
with respect to placing samples in relative positions on the 7-
point scale of language development, the measures obtained
by the method of this experiment appear to be satisfactorily
reliable.

Mean differences between the two sets of scale values for
(a) the two trials and (b) the two smaller groups of the first
trial were in both cases small, .36 and .49, respectively. These
differences, however, according to results of a t test for related
measures, were both significant beyond the .01 level. Scale
values were quite consistently slightly lower for the second
trial than for the first. Scale values of one of the two additional
sets derived for the first trial were quite consistently lower than
scale values of the other set. Certain precautions, then, in the
interpretation of exact values of obtained scale positions are
necessary. Pooling of scale values derived from responses of more
than one judging session for the same group of observers, or for
two groups of observers, make it necessary to follow a procedure
which will ensure that the observers use the same standards for
judging. If maintenance of the same standards is not possible,
scale values of two sets of scale values might be pooled in those

instances when one of the sets can be "adjusted" by the addition of an appropriately determined constant.

A set of 50 scale values of intricacy of language usage was derived from responses of 29 naive observers to 50 language samples, all on the same topic and varying in length. For the same samples, mean estimates of the ages of the children were obtained by averaging the estimates of five sophisticated observers. Structural complexity scores were computed from analyses of the 50 corresponding, longer 50-response samples.

Interrelationships among the three sets of measures were estimated by the Pearson r procedure. The Pearson r's for estimating these relationships for the indicated pairs of variables are as follows: structural complexity scores and scale values of intricacy of language usage, .63; structural complexity scores and mean estimates of age, .70; scale values of intricacy of language usage and mean estimates of age, .90. The high r of .90 is evidence not only of strong relationship between the last-mentioned sets of measures but also of their reliability. Comparisons among these three coefficients lead to the inference that the structural complexity score, as derived, may not be a good measure of the aspect of language for which it has been used. As previously mentioned, the validity of the weighting procedure has been questioned. The question arises also as to whether "complexity" may be at least partially dependent upon factors other than those used in derivation of structural complexity scores. The present results provide definite evidence leading to both questions. Possibly neither the categories of responses used in deriving these scores nor the weighting of them are satisfactory for the intended purpose.

That an extension of this experiment could result in useful new tools for the study of and the assessment of language development of children appears to be a reasonable assumption.

Literature Cited

Johnson, W., F. L. Darley, and D. Spriestersbach. 1963. *Diagnostic Methods in Speech Pathology.* New York: Harper and Row, 160-185.

Minifie, F. D., F. L. Darley, and D. Sherman. 1963. J. Speech Hearing Res., 6, 139-149.

Thurston, L. L. and E. J. Chave. 1929. *The Measurement of Attitude.* Chicago: University of Chicago Press.

Winitz, H. 1959. J. Speech Hearing Res., 5, 337-386.