1970

Construct Validity of the Length-Complexity Index (LCI)

Martha Morrill Hon
Eastern Illinois University

PAPER CERTIFICATE #2

TO: Graduate Degree Candidates who have written formal theses.

SUBJECT: Permission to reproduce theses.

The University Library is receiving a number of requests from other institutions asking permission to reproduce dissertations for inclusion in their library holdings. Although no copyright laws are involved, we feel that professional courtesy demands that permission be obtained from the author before we allow theses to be copied.

Please sign one of the following statements.

Booth Library of Eastern Illinois University has my permission to lend my thesis to a reputable college or university for the purpose of copying it for inclusion in that institution's library or research holdings.

August 6, 1970
Date

Author

I respectfully request Booth Library of Eastern Illinois University not allow my thesis be reproduced because _____

_____

_____

_____ _____

Date Author

CONSTRUCT VALIDITY OF THE LENGTH-COMPLEXITY INDEX (LCI)

(TITLE)

BY

Martha Morrill Hon

THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

Master of Science

IN THE GRADUATE SCHOOL, EASTERN ILLINOIS UNIVERSITY
CHARLESTON, ILLINOIS

1970
YEAR

I HEREBY RECOMMEND THIS THESIS BE ACCEPTED AS FULFILLING
THIS PART OF THE GRADUATE DEGREE CITED

5-6-70
DATE

ADVISER

May 6 1970
DATE

ACKNOWLEDGEMENTS

The writer wishes to acknowledge the interest and assistance of many persons. Without their help, she could not have successfully completed this thesis.

The writer would like to express since appreciation to Dr. Jerry Griffith and Mr. Lynn Miner for serving as thesis advisers and supervising the formulation of the research design, the collection of data, especially the statistical procedures, and the writing of the paper. Acknowledgement is given to Dr. Wayne Thurman for serving as both undergraduate and graduate adviser and as a member of the graduate committee.

Acknowledgement is given to fellow graduate students and professional friends for their help and support. Thanks is also given to the Department of Speech and the participating instructors at Eastern Illinois University for their cooperation.

Special thanks is given to the writer's husband without whose understanding, confidence, and encouragement the completion of these months of graduate study would not have been possible. Gratitude is also expressed to the writer's parents whose assistance made higher education possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

Figure — Page

CHAPTER I

STATEMENT OF THE PROBLEM

Introduction

Disillusionment with traditional methods of evaluating linguistic performance in children has motivated recent experimentation testing the value of new language measures. One such language measure is the length-complexity index (LCI) as first proposed by Shriner (1967). While the LCI scoring procedures (Miner, 1969) and the temporal reliability (Barlow and Miner, 1969) have been discussed in the literature, its construct validity remains an unanswered question.

Construct validity may be defined as the psychological meaningfulness of the test (Lyman, 1963). From construct validity the results of a test which logically should be obtained can be predicted if the test is valid. The prediction is stated in terms of a coefficient of correlation which lends itself to a statistical test of significance. In this way, a check is made of the validity of both the test and its underlying theory. For the purpose of this investigation, construct validity will be used. Construct validity involves a correlation between test scores and values of another variable; however, the outside variable is

not really a criterion, even though it is a variable which should relate logically to the test. Traditionally, another test which purports to measure the same parameters of the test in question is used as the outside variable. However, in this instance, such a procedure seems unwarranted since the validity and reliability of existing language measures is questionable. Minifie, Darley, and Sherman (1963) found relatively low temporal reliability for the language measures (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 scores, and the type-token ratio) they investigated. It appears that language measures obtained from 50 response language samples are not consistent from day to day. Shriner's (1969) research indicates that response length does not appear to be a significant indicator of expressive language for children who are approximately five years of age and older, because of increased response variability. Therefore, researchers must resort to a more meaningful outside criterion in order to assess LCI construct validity.

Sherman, Shriner, and Silverman (1965) have suggested that the impression language makes upon others might serve as a useful outside criterion. This approach is based on the assumption that measuring language development is primarily a perceptual phenomenon; that is, evaluations of language ability, in the final analysis, are based on judgments from

human observers. As a consequence, if it can be demonstrated that observers' judgments are predictable, the validity of psychological evaluations can be assessed in terms of the amount of agreement among observers. The observers' judgments can be transformed into measurements according to various psychological scaling methods. If observers repeatedly generate diverse scale values, they obviously have different referrential systems for assigning numbers to stimuli. On the other hand, high observer agreement would be interpreted to mean (1) that essentially the same standard was utilized in assigning scale values to stimuli and (2) the stimuli have basically the same perceptual impact on the observers. Observer agreement is the variable that is logically related to LCI scores; therefore, it is appropriate to use this variable to compare with LCI values. The LCI could be considered a valid measure of expressive language ability if it could be demonstrated that LCI scores have a high positive correlation with psychological scale values derived from observers' ratings of expressive language ability. This high positive correlation also would satisfy the definition of construct validity.

Recent research by Sherman, Shriner, and Silverman (1965) and Shriner (1967) utilized psychological scale values obtained from observers' ratings to assess the developmental level of verbal output in children. The general approach in both studies was to make comparisons among correlation

coefficients obtained for the purpose of estimating relationships between various measures of language development for the same set of 50 samples of children's language. In both studies, judges rated entire response segments to stimulus pictures, not individual utterances. The Sherman et al. study (1965) concluded by questioning the validity of the structural complexity score (Templin, 1957) and by suggesting that psychological scaling of children's language could provide new and useful tools for the study of and the assessment of children's language development. Utilizing a multiple-regression procedure, Shriner and Sherman (1967) found that the best single predictor of degrees of language development was the mean length of response (MLR). In a follow-up study, Shriner (1967) used four linear multiple-regression analyses to determine the best composite of several language measures for predicting scale values of language development derived from observers' ratings of child language samples for four different categories. He found that a combined length-complexity measure remained as the single, best predictor of psychological scale values of language development for children of five years of age and younger; that is, a length-complexity index (LCI) more sensitively reflected the impression language makes upon observers than traditional, independent language measures for children of this age category.

The multiple-regression analyses by Shriner (1967)

specified the combination of parameters (sentence length and complexity) which may correlate highly with observers' judgments of linguistic maturity in children. Shriner reported a correlation of 0.87 between the LCI and psychological scale values. In this instance, observers were again rating entire response segments to stimuli; whether a correlation of the same magnitude would be obtained with individual utterances as the test stimuli is not known. Moreover, the multiple-regression technique by itself does not mean that the derived parameters are necessarily the only significant dimensions influencing observers' ratings. There may be other parameters that correlate highly with those derived from the multiple-regression analyses. In other words, while the multiple-regression equation predicts which relevant variables should correlate highly with observers' ratings, it does not, essentially, confirm or reject this prediction. In order to validate the results of multiple-regression analyses, the parameters thought relevant would need to be systematically varied to see if the outcome of observers' ratings can be predicted. In its current stage of development, it is not known whether the LCI can adequately predict observers' ratings of the degree of language development.

In a recent study, Miner and Silverman (1969) assessed the ability of observers to reliably rate single utterances for an attribute of language development and to assess the

influence of mode of stimulus presentation (visual or auditory) upon these ratings. Statistical analyses of the date indicated that observers can reliably scale single utterances for the language attribute rated. The intraclass correlation coefficient for assessing the reliability of the scale values for both auditory and visual presentations exceeded 0.98. The correlation between sets of scale values for auditory and visual presentations of the stimuli was 0.96, which indicates that both modes of stimulus presentation result in a similar ordering of the stimuli. The authors concluded by recommending that additional psychological scaling experiments are needed in which observers assign scale values to stimuli that hold sentence length constant and permit complexity to systematically vary.

In summary, the results of several recent investigations (Nelson, 1966; Sherman, Shriner, and Silverman, 1965; Shriner, 1967; Shriner and Sherman, 1967; Sherman and Silverman, 1968; Miner and Silverman, 1969) provide strong evidence that psychological scaling can be useful for various purposes in the assessment of children's language development, including its use as an outside validity criterion for the evaluation of new measures of linguistic performance. This study represents a systematic continuation of the research initiated by the above investigators. The general purpose of this investigation was to assess the construct validity of the LCI. Specifically, the following

questions were posed at the outset of this study:

1. Can observers reliably scale single utterances representing 57 different grammatical structures obtained from child language samples?
2. What is the relationship between LCI scores and observers' judgments of intricacy of language?
3. Based on the results of this study, what, if any, changes in the LCI scoring procedure are indicated?

CHAPTER II

REVIEW OF LITERATURE

Psychological scaling methodology enables one to quantify the perceptual impact that various speech and language disorders have on observers (Young, 1969). Examples include studies such as those of Morrison (1955), Sherman and Moodie (1957), and Sherman and Morrison (1955) where they quantified or gave a numerical value to articulation severity; studies such as those of Sherman and Lewis (1951), Sherman and Trotter (1956), and Cullinan, Prather, and Williams (1963) applied psychological scaling methodology to the auditory characteristics, frequency, and severity of stuttering. Psychological scaling procedures have also been used to evaluate other speech disorders; Sherman and Linke (1952) and Rees (1958) used an interval scale to determine whether the variation of vowel count had any effect on perceived harshness.

It has been demonstrated that psychological scaling can be useful for various purposes in the assessment of children's language developmn t as the result of recent investigations (Elliott, Hirsh, and Simmons, 1968; Nelson, 1966; Sherman, Shriner, and Silverman, 1965; Shriner, 1967; Sherman and Silverman, 1968). For these investigations the

method of equal-appearing intervals (Edwards, 1957) was used. The psychological rating scale methods evaluated by Sherman and Silverman (1968) for their usefulness in measuring a specific aspect of children's language development--intricacy of language usage were equal-appearing intervals, successive intervals, and direct magnitude estimation.

Recent experiments testing the value of new language measures were prompted by disillusionment with conventional means of assessing child language. Although mean length of response seems adequate for some purposes, the arbitrary weighting system, the structural complexity score, proposed by Templin (1957) to evaluate the grammatical categories of children's language development has been questioned (Darley and Moll, 1960; Minifie, Darley, and Sherman, 1963). Numerous investigators have analyzed length of response independently of complexity of response. Because language production increases in length as well as in complexity with increasing chronological age, and because the weighting system used to assess complexity of response was questioned (Darley and Moll, 1960; Minifie, Darley, and Sherman, 1963), a procedure which combines both length and complexity of response into a single measure may prove to be more useful for research or clinical purposes than either of these measures used independently (Shriner, 1969). One of the most widely used measures of children's language is the Mean Length of Response (MLR). However, recently certain

investigations (Shriner, 1969; Minifie, Darley, and Sherman, 1963) have indicated that the MLR is not a valid measure of language development. Shriner (1967) in comparing selected measures with psychological scale values of language development found that in the multiple-regression analysis the length-complexity measure remained as the single, best predictor of psychological scale values of language development. As the mean age of the groups for analysis increased, MLR lost significance as a predictor. Barlow and Miner (1969) assessed the temporal reliability of the Length-Complexity Index (LCI) and the MLR. They found the intra-class correlation coefficient for MLR was $r_1 = 0.65$ compared to $r_1 = 0.80$ for the LCI, for the individual child's responses on subsequent retests of single 50-response language samples. This indicates that there is considerable variability of MLR as a measure of a child's daily verbal language performance.

A length-complexity measure was formed (Shriner, 1969) by relying on the research of Menyuk (1964a) and Cazden (1965). Menyuk (1964a) reported that complexity was not related simply to increasing sentence length or proportion of usage of what has been termed compound or complex sentences. Increasing complexity, according to Menyuk, is proceeding from the most general rule to the application of increasingly differentiating rules. She reported, for example, that to conjoin two sentences, or to delete and substitute as in relative clauses requires the appreciation

of certain rules. If a child uses a rule to generate a sentence and then proceeds to conjoin two or possibly three similar sentences, the utterance would be obviously increasing in length; however, the utterance would not be increasing in complexity.

Further experimentation with transformations or psychological scaling procedures may help to develop a weighting method with equal units that will eventually prove worthwhile in clinical evaluation. As Carroll (1961, p. 334) has stated: "If such developmental scales could be established, they would probably be more meaningful than such indices of language development as mean sentence length." As a result of this need, the length-complexity index (LCI) has been proposed as a more sensitive measure of verbal maturity in children than the mean length of response or the structural complexity score.

Psychological scaling methods have been employed as a means of assessing the psycholinguistic reality of this measure. The LCI is a linguistic measure designed to make a composite analysis of sentence length and sentence complexity. Both length and complexity are considered together (not independently) according to a numeric weighting system. It is a modified combination of two previous measures, the mean length of response (McCarthy, 1954, chapt. 9) and the structural complexity score (Templin, 1957, p. 81). The LCI measure is based on the research of Menyuk (1964a),

Cazden (1965), and Bellugi (1964) and was first synthesized by Shriner (1967). The child's final LCI score is the sum of his noun phrase (NP) points plus verb phrase (VP) points plus additional points (AP) for each sentence divided by the number of sentences (NS). Put differently, $LCI = \frac{NP_1 + VP_2 + AP}{NS}$ (Miner, 1969).

Psychological scaling methods have been employed as an outside validity criterion for measures of expressive language ability in children. In other words, it serves as a means of assessing the validity of newly developed measures of verbal output. Psychological rating-scale methods thus might provide measures useful for evaluation of the validity of the indices currently used. When doing psychological scaling experiments a number of procedural problems arise. One must first determine which scaling method to utilize. Sherman and Silverman (1968) found in their study that the three sets of scale values derived by the method of equal-appearing intervals and the method of successive intervals rank ordered the samples in almost identically the same manner ($r = 0.995$). Since the two methods result in such closely related scale values, the method of equal-appearing intervals, because of the simpler computational procedures is usually the preferred one. The correlation of the direct magnitude estimation mean scale values with the equal-appearing intervals and successive intervals scale values was high (0.92). Scale values obtained by the

three methods appear to differ very little in their usefulness, at least for the kind of stimulus used in this study.

In equal-appearing intervals, the observer divides his psychological continuum into categories of equal width, then assigns a category number to each stimulus. Direct magnitude estimation requires that the observer state the ratio between each sample and some standard stimulus, e.g. twice as severe, half as severe. In both procedures the numerical ratings are usually average over observers rather than over repeated judgments of the same observer. The category scaling method of equal-appearing intervals is the most popular technique because of its ease of administration, reliability of scale values, and minimal underlying assumptions concerning the observers' abilities (Young and Downs, 1968). Other procedural problems concern the nature of the stimuli to be scaled, such as auditory versus visual presentation and the rating of single utterances versus rating of entire response segments.

In psychological scaling methodology there are basic assumptions to be made when it is applied to speech disorders. Psychological scaling procedures when applied to speech disorders differ from their classical usages in some important ways (Young, 1969). The stimulus dimensions of disordered speech are nonmetric and multidimensional. This is not handicapping if it can be demonstrated that the observer judgments of a particular class of nonmetric

events are as predictable and manipulable as if the stimulus dimension had a direct physical correlate. The multidimensional nature of most speech stimuli is a more important problem. Speech stimuli usually differ from one another in more than one respect. For this reason, validity of such psychological measurements should be examined in terms of the amount of agreement among the observers. If observers' response numbers for the same stimulus are grossly dissimilar, then one could question whether the dimension being evaluated was sufficiently unidimensional for the numbers so generated to have any operational validity, or whether the observers were able to ignore sufficiently the extraneous characteristics of the speech sample.

There is little in the literature to assist the examiner in deciding the number of judges he needs to use with the possible exception of information pertaining to the magnitude of reliability coefficients which have been reported for scaling experiments in which different numbers of judges were used (e.g., Edwards, 1957, pp. 94-95). Such information is of limited usefulness since the number of judges required to attain a specific level of reliability would be expected to vary. In the "typical" scaling experiment, the size of the judging panel is fixed prior to beginning the experiment and reliability of the scale values is permitted to vary. An alternative approach would be to fix the minimum level of reliability desired for the scale

values prior to beginning the experiment and permitting the size of the judging panel to vary. This is referred to as the method of sequential sampling (Silverman, 1968). In addition to providing control over reliability, this solution would permit the size of the panel to be reduced to a minimum.

Miner and Silverman (1969) evaluated the relationship between length-complexity index scores and scale values of degree of language development derived from observer ratings. The language samples to be scaled were presented to the observers according to two different modes: (1) auditorily via playback of a tape recording prepared by the experimenters and, (2) visually via a typed manuscript. All individual utterances were rated by the method of equal-appearing intervals (Edwards, 1957) on a seven point scale of degree of language development. They found that either auditory or visual presentation of the stimuli will yield comparable results (r = 0.956); observers can reliably scale individual utterances from children's language samples (r = 0.984); and, a high relationship exists between LCI scores and observers' judgments of degree of language development (r = above 0.90).

This review of the literature seems to warrant the following conclusions:

1. Psychological scaling has provided a methodological tool for the assessment of various attributes of speech

and language behavior on observers (Young, 1969; Morrison, 1955; Sherman and Trotter, 1956; Rees, 1958; and Shriner, 1967).

2. The LCI appears to be a more sensitive measure in assessing a child's verbal maturity than traditional methods (Miner, 1969).
3. Psychological scaling methods can be employed as an outside validity criterion for measures of expressive language ability in children (Sherman and Silverman, 1968).
4. The method of equal-appearing intervals is preferred because of its ease of administration, reliability of scale values, and minimal underlying assumptions concerning observers' abilities (Young and Downs, 1968).
5. Auditory and visual presentations of stimuli yield comparable results when they are rated as single utterances of children's language in determining intricacy of language usage (Miner and Silverman, 1969).
6. Observers can reliably scale individual utterances from children's language samples (Miner and Silverman, 1969).

CHAPTER III

SUBJECTS, PROCEDURE, EQUIPMENT

Subjects: Transcripts of tape recorded language samples from the speech of 17 subjects, 10 males and 7 females, within two months of age five were available from another experiment (Barlow and Miner, 1969). Each of the subjects had essentially normal intelligence as measured on the Peabody Picture Vocabulary Test (Dunn, 1965; mean IQ = 101.4, s.d. = 7.9), had normal hearing for the speech frequencies, exhibited no obvious neuromuscular impairment and was of lower middle socioeconomic status (Warner, Meeker, and Eells, 1949). All of the subjects were selected from the Sullivan, Illinois Public School system. The language samples obtained from the children were evoked by reading readiness pictures. A total of 2,550 utterances constituted the corpus from which the items to be scaled were selected. Each utterance was analyzed according to the LCI scoring procedures (Miner, 1969). This subject population comprised all of the children available who could meet the criteria for selection.

Preparation of Stimuli: The stimuli from which the psychological scale values of language development were obtained consisted of four pairs of grammatically matched

utterances randomly selected for each LCI point value one through ten. Each pair of utterances at each point level consisted of a different type of syntactic structure. This constitued an initial corpus of 80 stimuli. In addition, 22 examples of developmental language acquisition data as discussed by Brown and Bellugi (1964) and Cazden (1965) were included to test the psychological reality of the sequence of emergence data. These samples were included to determine if they would be rated according to the sequence of emergence by the judging panel; that is, those appearing late, according to Brown and Bellugi would receive higher values while those appearing early would receive lower value judgments. Samples were taken directly from Brown and Bellugi's (1964) and Cazden's (1965) work. Twelve samples of the two categories of questions and four categories of negatives taken directly from Miner's (1969) LCI scoring procedures were included to see if the judging panel would rate the samples in the same manner as the scoring procedures suggest; that is, assign scale values that are proportionate to scoring weight. The stimuli were randomly assigned to the answer sheet. A total of 114 individual utterances were scaled.

Description of Scaling Method: The psychological scaling method of equal-appearing intervals (Edwards, 1957) was selected as the preferred measurement tool. Prior research (Elliot, Hirsh, and Simmons, 1968; Nelson, 1966;

Sherman, Shriner, and Silverman, 1965; Shriner, 1967; and Silverman, 1968) has demonstrated that this method is reliable for purposes of scaling child language samples. A seven-point equal-appearing intervals scale of intricacy of language usage was used with one representing least intricacy of language usage and seven representing most intricacy.

Selection of Judging Panel: Judges who rated the experimental samples were undergraduate students in the Department of Speech at Eastern Illinois University. The single restriction placed upon their selection was the elimination of any student who had previously been enrolled in a course in language development. This restriction seemed necessary in order that ratings would not be unduly influenced by specific and extensive knowledge of the particular language measure under study.

Presentation of Stimuli: The samples to be scaled were presented to the observers visually, via a typed manuscript. Each language sample was preceded by a number. The judges recorded their judgments on the answer sheet to the left of the identifying number of the language sample. A sample answer sheet is included in Appendix I. The instructions to the judges are shown in Appendix II.

Analyses of Judges' Ratings: The method of sequential sampling (Silverman, 1968) was used to determine the number of judges for this experiment. In this approach,

the minimum level of reliability desired for scale values is fixed prior to beginning the experiment and the size of the judging panel is permitted to vary. The desired level of reliability for this experiment was set at 0.95. A total of 33 judges rated 114 stimuli. The judges' ratings were transferred from the answer sheet to IBM data cards from which statistical computation was made. In order to evaluate the reliability of scale values, an intraclass correlation coefficient for averages (Winer, 1962) was computed. To determine the relationship between LCI scores and observers' judgments, a Pearson Product-Moment correlation coefficient was computed. Both correlation analyses were performed by an IBM 360 computer. To test for significance of differences obtained in mean scale values for each of the classes of grammar scaled, a *t* test for significance was applied. This, too, was performed by computer.

CHAPTER IV

RESULTS--DISCUSSION

The purpose of this investigation was to assess the construct validity of the LCI. Specifically, three questions were posed at the outset of this study. This chapter lists those questions, reports the statistical computations, and interprets the results.

1. Can observers reliably scale single utterances representing 57 different grammatical structures obtained from child language samples?

To answer the question posed, an intraclass correlation coefficient for averages (Winer, 1962) was computed for the scale value ratings by the 33 judges. The obtained r was 0.97. This value was interpreted to mean that if the experiment were to be repeated with another random sample of observers from the same population rating the same set of stimuli, the resulting correlation between the ratings obtained from the judges would again be approximately 0.97. The obtained r of 0.97 suggests a high degree of reliability among the judges used for the scaling task. The conclusion is drawn that the observers can reliably scale single utterances

representing 57 different grammatical structures obtained from five year old child language samples.

2. What is the relationship between LCI scores and observers' judgments of intricacy of language usage?

The relationship between LCI scores and observers' judgments of intricacy of language usage ($\overline{MSV}$) was assessed by means of a Pearson Product-Moment correlation coefficient. Intricacy of language usage for the purpose of this experiment was defined as the ability to string words together for the purpose of conveying information. The resulting r was 0.87. This correlation coefficient was interpreted to mean that the two variables rank ordered themselves in approximately the same manner. This suggests that the LCI is a highly sensitive indicator of observers' judgments of intricacy of language usage when those judgments are based upon single utterances. This lends additional support to the construct validity of the LCI.

In an effort to further analyze the relationship between LCI scores and observers' judgments, the stimuli were subdivided into two different groups. The first group of stimuli rated by the judging panel consisted of four pairs of grammatically matched utterances (N = 80) randomly selected for each of the LCI point values one through ten. Each pair of utterances at each point level consisted of a different

type of syntactic structure. The second group of stimuli consisted of various types of questions, negatives, noun phrases, and verb phrases that were not included for analysis in the Miner and Silverman (1969) study. These stimuli (N = 34) were included in order to experimentally test the construct validity of that portion of the LCI scoring system that is based upon the sequence of emergence data reported by Bellinger (1964), Brown and Bellugi (1964) and Cazden (1965). Further discussion of the second group of stimuli will appear in a later section of this chapter.

The relationship between mean scale values and the LCI values of the first group of stimuli (N = 80) was assessed by means of a Pearson Product-Moment correlation coefficient. The resulting correlation coefficient was 0.85. This was interpreted to mean that the two variables rank ordered themselves in approximately the same manner. Recall that the overall correlation coefficient between mean scale values and LCI scores (N = 114) was 0.87. This correlation between mean scale values and LCI scores was based on 114 stimuli. The difference between the 0.85 and the 0.87 correlation coefficients was not statistically significant (z = 0.47).

In this first group of stimuli, the question arises as to whether or not observers would rate the stimuli according to their semantic properties rather than their intricacy of language usage as instructed. That is, do observers assign comparable scale values to utterances syntactically matched

but semantically different? To answer this question, a *t* test was computed between the matched pairs of utterances in order to determine if a significant difference exists between the scale values for the two subsets of stimuli. The resulting *t* value (1.85; df = 38) was not significant at the .05 level of confidence. Apparently, the semantic values of the stimuli did not appreciably influence the judges' rating. Again, this is further evidence of the construct validity of the LCI.

The relationship between scale values and LCI scores is graphically portrayed as a frequency polygon in Figure 1. The frequency polygon is a visual presentation of the relationship between two variables. Figure 1 shows the degree of association between LCI scores and observers' judgments of intricacy of language usage. A relatively linear relationship exists between these two factors based on a sample size of 33 judges. The reader will recall that the method of sequential sampling (Silverman, 1968) was used to determine the number of judges needed to attain a reliability level of 0.95 or better. A reliability level of 0.97 was achieved with only 33 judges.

3. Based on the results of this study, what, if any, changes in the LCI scoring procedure are indicated?

The scoring procedure of the LCI was based in part on the research of Brown and Bellugi (1964) concerning sequence

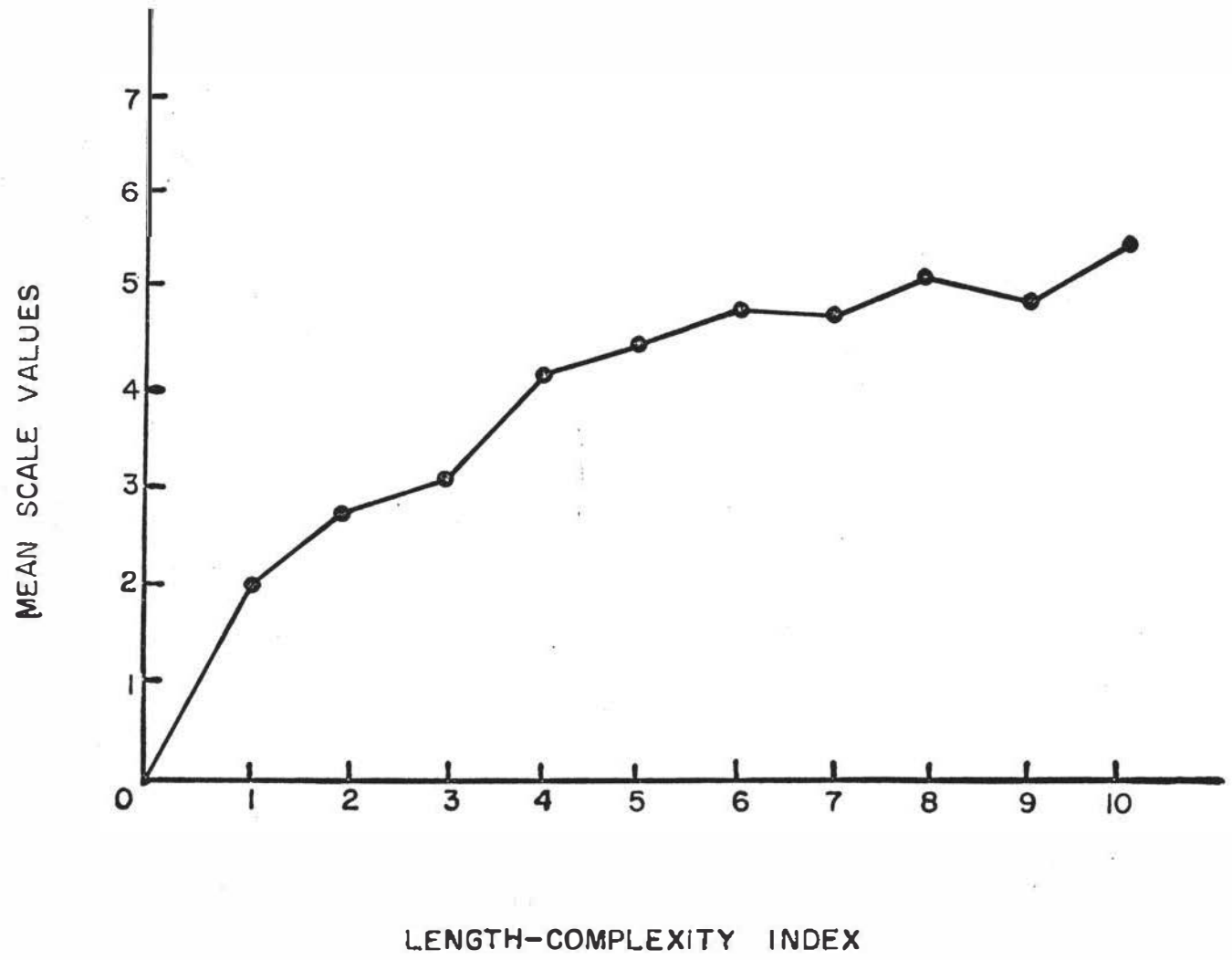

Figure 1.--Frequency polygon depicting the relationship between LCI scores and observers' judgments of intricacy of language usage.

of emergence. There are also two areas of the LCI, negation and question, which are difficult to score because of their differing effects on observers. The data gathered from these three areas will be reported and discussed in the following sections of the chapter.

SEQUENCE OF EMERGENCE

The second group of stimuli ($N = 34$) exemplify data on the developmental sequence of language acquisition; they were included to test the construct validity of sequence of emergence as discussed by Brown and Bellugi (1964), Cazden (1965), and Bellugi (1964). Cazden (1965) used both sequence of emergence and structural complexity criteria in evaluating the child language samples incorporated in her study. Language samples from these two structures, noun phrase and verb phrase, were included in the present investigation in order to determine how the sequence of emergence data would be rated by the judging panel; that is, those noun phrase utterances emerging later, according to Brown and Bellugi (1964), should receive higher scale values while those appearing early would receive lower scale values. In Brown and Bellugi's (1964) research they found that in the first stage of noun phrase emergence, any modifier could be used with any noun. When the differentiation process begins, articles are separated out of the general class of modifiers. Only later do children use two modifiers other than articles before a

noun. Cazden (1965) reports that children use unmodified verbs before they use auxiliaries. Therefore, a weighted index would assign more points to an auxiliary plus a verb than to a verb alone. With our present state of knowledge, it is not known when the past tense appears in relation to other forms but its period of emergence is definitely later. Samples for noun phrases were taken directly from Brown and Bellugi's work and samples for verb phrases were taken from Cazden's work.

Table 1 shows utterances considered by Brown and Bellugi to be representative of the developmental sequence of language acquisition data for noun phrases. Table 1 also lists the LCI values and the mean scale values for each pair of matched utterances.

Inspection of these data reveals that the mean scale values rank order themselves in a manner consistent with Brown and Bellugi's developmental sequence, i.e. the judges rated those stimuli appearing later as higher than those appearing earlier. In addition, the mean scale values are consistent with the LCI's computed for the sample. As the LCI increased so did the scale values assigned. To determine the significance of the differences in mean scale values between pairs of stimuli, that is between *a* and *b*, *b* and *c*, and so forth, a *t* test was applied. The results are reported in Table 2. All differences were statistically significant at the .05 level except for the differences between c_1 and d_1

and for d_1 and e_1. In light of the firm and significant trend shown in all other comparisons, this result may be described as spurious. The impact of the utterance nice flower on the judges was in some way not consistent with the psychological set they had for the other utterances.

TABLE 1.--Developmental sequence of language acquisition (noun phrase) from Brown and Bellugi (1964)

Utterance	LCI	$\overline{MSV}$	Q
a_1. flower	1	1.55	0.07
a_2. dog	1	1.52	0.05
b_1. a flower	2	2.12	0.32
b_2. a dog	2	1.94	0.33
c_1. the flower	2	2.36	0.91
c_2. the dog	2	2.00	0.39
d_1. nice flower	2	2.73	0.74
d_2. big dog	2	2.58	0.75
e_1. a blue flower	3	3.06	0.77
e_2. the big dog	3	3.03	0.68
f_1. my blue flower	4	3.39	0.26
f_2. my big dog	4	3.70	0.76

On the basis of these results it would appear that observers tend to rate the intricacy of noun phrase usage in a manner similar to linguistic findings regarding developmental sequence of emergence. Moreover, the noun phrase sequence of emergence data is rank ordered in the same manner by both methods of analysis (LCI and equal-appearing intervals).

Finally it should be noted that the dispersion of scale values generally increases as the length and complexity of the utterance increases. However, in no case did Q exceed 0.91, suggesting relatively high observer agreement for each stimulus.

TABLE 2.--Values of t for tests of significance of differences in mean scale values between noun phrase utterances for developmental sequence of emergence

Comparisons	$\overline{MSV}$	t
a_1	1.55	4.18*
b_1	2.12	
a_2	1.52	3.68*
b_2	1.94	
b_1	2.12	1.24
c_1	2.36	
b_2	1.94	0.49
c_2	2.00	
c_1	2.36	1.61
d_1	2.73	
c_2	2.00	2.70*
d_2	2.58	
d_1	2.73	1.82
e_1	3.06	
d_2	2.58	2.60*
e_2	3.03	
e_1	3.06	2.35*
f_1	3.39	
e_2	3.03	3.02*
f_2	3.70	

*$p < .05 = 2.03$; df = 32

Below is a table of the developmental sequence of language acquisition data for verb phrases. Again the reader will find the LCI values and mean scale values listed for the paired matched utterances. If the difference between the mean scale values of a pair of matched utterances was greater than 0.75 a *t* test was computed.

TABLE 3.--Developmental sequence of language acquisition (verb phrase) from Cazden (1965)

Utterance		LCI	$\bar{MSV}$	Q	*t*
a_1.	I drop	2	2.39	0.89	
a_2.	I jump	2	2.88	0.93	
b_1.	I dropping	3	2.55	0.72	
b_2.	I jumping	3	3.18	0.98	
c_1.	I'm dropping	4	2.67	0.78	2.63*
c_2.	I'm jumping	4	3.52	1.14	
d_1.	I dropped	3	2.36	0.64	
d_2.	I jumped	3	2.97	0.91	
e_1.	He drops	2	2.67	0.80	
e_2.	He jumps	2	3.03	0.73	

*$p < .05 = 2.03$; df = 32

The difference between mean scale values for grammatically matched pair *c* was analyzed by means of a *t* test. The resulting *t* value for pair *c* (2.63; df = 32) was statistically significant at the .05 level of confidence. The writer hypothesizes that the statistically significant difference between syntactically matched pair *c* occurred on

a semantic basis. There had to be some semantic attribute that caused observers to scale the utterances differently. Two factors could have been involved: (1) It could have been a matter of frequency of occurrence. According to a spoken word count for five year olds (Wepman and Hass, 1969) jump occurs more frequently than drop; therefore, I'm jumping was scored higher than I'm dropping, and (2) I'm jumping may be a stereotyped response having reduced propositional value; therefore, judges might have felt it was not as complex as its matched utterance. Conceivably the observers felt that this phrase does not show much ability to string words together for the purpose of communication. It could also have been due to the wide dispersion for this stimuli. It is concluded that on the basis of the small sample size, any further attempts to explain the difference between these utterances would be speculative.

Inspection of the data in Table 4 reveals that the mean scale values do not rank order themselves in a manner consistent with Cazden's developmental sequence, i.e. the judges did not rate those stimuli appearing later higher than those appearing earlier. In addition, the mean scale values are not consistent with the LCI values computed for the sample. To determine the significance of the difference in mean scale values between pairs of stimuli, that is, between a and b, b and c, and so forth, a t test was computed. The results are reported in Table 4. All differences were not statistically

significant at the .05 level except for difference between c_2 and d_2. Because of the non-significant trend shown in all other comparisons, this result may best be described as spurious. The impact of the utterances I'm jumping and I jumped on the judges was in some way not consistent with the psychological set they had for the other utterances.

TABLE 4.--Values of t for tests of significance of differences in mean scale values between verb phrase utterances for developmental sequence of emergence

Comparisons	$\bar{M}SV$	t
a_1	2.39	0.66
b_1	2.55	
a_2	2.88	1.67
b_2	3.18	
b_1	2.55	0.46
c_1	2.67	
b_2	3.18	1.32
c_2	3.52	
c_1	2.67	1.67
d_1	2.36	
c_2	3.52	2.45*
d_2	2.97	
d_1	2.36	1.03
e_1	2.67	
d_2	2.97	0.27
e_2	3.03	

*$p < .05 = 2.03$; df = 32

On the basis of these results it would appear that observers do not rate the intricacy of verb phrase usage in a manner similar to linguistic findings regarding developmental sequence of emergence.

The writer raises the following points:

1. The change in subjects of some of the stimuli could have been an influencing factor. Both first and third person pronouns were used.
2. Only the contracted form of the auxiliary am was used. Since the observers didn't see any difference between I dropping and I'm dropping, it seems to indicate that the contracted form of the auxiliary doesn't show more linguistic maturity.
3. How much does grammaticality influence observers' judgments? This is a question that has not been empirically explored, but needs to be.
4. The results of this study indicate that revision of the LCI scoring procedures for verbs needs to be explored. These data did not rank order themselves in a manner the investigator expected. Because the sample size was small, further research needs to be done before definite suggestions concerning revision can be made.

NEGATIVES

Eight utterances containing negatives were included in the stimuli to determine if the judging panel would rate

the samples in the same manner as the LCI scoring procedure suggests: that is, assign scale values that are proportionate to scoring weight. Negative utterances were included because their construct validity has not been demonstrated. The LCI point system for negatives was based on the research of Bellugi (1964). Four different point levels were operationally defined for the use of negatives:

1. When the negation appears either at the beginning or at the end of the utterance, not within, and consists of no or not and the rest of the sentence, score as one point.

2. Two auxiliary verbs appear in the negative form, can't and don't. The negative element now appears within the sentence, but may or may not be connected to an auxiliary verb. Nominal + no, can't, don't + main verb is scored as two points. Furthermore, at this point level, the negative also appears in the demonstrative form at the beginning of a sentence in the imperative form. Demonstrative + no or not + nominal is observed as well as don't + main verb.

3. When the negative form appears between the noun phrase and the present participle, a weighting value of three points is assigned (NP + Ng + PrPt).

4. The last level exemplifies the adult version of the negative. The sentence includes appropriate intonation and is scored as four points. Auxiliaries are contrasted with the negative n't. These sentences are of the form: Nominal + Aux + Ng + V. In child language the verb be is often missing but is now optional.

Below is Table 5 containing the negation stimuli used in this investigation. The LCI values, negative values, mean

scale values, and the semi-interquartile ranges are listed. If the difference between mean scale values of a pair of matched utterances was greater than 0.75 a *t* test was computed.

TABLE 5.--Stimuli representing the four levels of Negatives as scored in the LCI and defined by Bellugi

	Utterance	LCI	Ng	$\overline{MSV}$	Q	*t*
a_1.	No wash	2	1	2.06	0.26	
a_2.	Wear mitten no	3	1	2.45	0.68	
b_1.	I don't know	5	2	4.39	1.95	
b_2.	I no bit you	6	2	3.94	1.14	
c_1.	Me not crying	6	3	3.76	0.84	
c_2.	I no peeking	6	3	3.15	0.75	
d_1.	No, it isn't	7	4	3.33	0.85	7.24*
d_2.	I am not a doctor	8	4	5.27	1.43	

*$p < .05 = 2.03$; df = 32

The resulting *t* value for pair *d* (7.24; df = 32) was significant at the .05 level of confidence. The investigator hypothesizes that the statistically significant difference between grammatically matched pair *d* was simply a matter of semantics. Although the negative element of these two structures is scored the same, the entire structures are assigned differing LCI scoring weights. One is a more linguistically complex utterance than the other and the writer feels this is what the observers based their judgments on when they rated one utterance higher than the other.

Examination of these data reveals that the mean scale values did not rank themselves in a manner consistent with the four scoring levels operationally defined by Bellugi. The mean scale values are not consistent with the LCI's computed for the sample. To determine the significance of the differences in mean scale values between pairs of stimuli, that is, between a and b, b and c, and so forth, a t test was applied. The results are reported in Table 6. All differences were statistically significant at the .05 level except for the differences between b_1 and c_1 and c_1 and d_1. Since a significant trend was shown in all other comparisons, these results may be described as spurious. The impact of the utterance Me not crying on the judges was in some way not consistent with the psychological set they had for the other utterances. In both instances I don't know and No, it isn't are stereotyped responses. Although Me not crying shows a higher level of negation, it has a lower level of grammaticality which may offset the higher negation level. The matched stimuli of each of the above pairs is a better example of its particluar negative level and these examples shown a significant difference.

On the basis of these results it would appear that observers did not tend to rate the intricacy of negative usage in a manner similar to that indicated by Bellgui's research.

TABLE 6.--Values of t for tests of significance of differences in mean scale values between negative utterances

Comparisons	$\overline{MSV}$	t
a_1	2.06	6.47*
b_1	4.39	
a_2	2.45	5.39*
b_2	3.94	
b_1	4.39	1.92
c_1	3.76	
b_2	3.94	3.64*
c_2	3.15	
c_1	3.76	1.49
d_1	3.33	
c_2	3.15	7.13*
d_2	5.27	

*$p < .05 = 2.03$; df = 32

It is suggested by this investigator that the four level point system for negatives be reduced to a two level system. It is recognized that some usages of negatives are more complex than others but not as Bellugi's four level system suggests. The less complex structures as defined in levels one and two could be combined to become level one. The more complex structures as defined in level three and four could be combined to become level two. It seems to the writer that this would be less confusing to the person assigning scoring values and be a more accurate picture of how these types of utterances affect observers.

This revision of the four level point system for negatives was empirically assessed by the writer. Revised scoring values were assigned to the negative stimuli and a Pearson Product-Moment correlation was computed between mean scale values, original LCI values, and revised LCI values. The resulting correlation of mean scale values and original LCI values was 0.35 and the correlation of mean scale values and revised LCI values was 0.37. The correlation between original and revised LCI values was 0.98. This was interpreted to mean:

1. The revised LCI scoring procedures for negatives did not rectify the discrepancies found between LCI scores and mean scale values.
2. The revised LCI scoring procedures for negatives were not drastically changed.
3. The LCI scoring procedures, as they now stand, are not assessing negative stimuli included in this investigation as observers perceive the same stimuli.

Below is Table 7 showing the mean scale values as compared to original and revised LCI scoring values for negatives.

In comparison to the total number of stimuli, only a few samples of negation were included for rating. The results of this investigation are only speculative and represent an attempt by the author to interpret what actually happened. More research is needed in this area using more stimuli dealing only with negatives to determine more in

detail the manner in which this type of stimuli affects observers. It is recognized that this is a difficult area in which to assign point values as well as for observers to give a perceptual rating; therefore, more research would aid in setting a more accurate assessment of this aspect of child language.

TABLE 7.--Original compared to revised LCI values for negatives

Utterances	$\bar{M}SV$	Original LCI	Revised LCI
a_1. No wash	2.06	2	2
a_2. Wear mittens no	2.45	3	3
b_1. I don't know	4.39	5	4
b_2. I no bit you	3.94	6	5
c_1. Me not crying	3.76	6	5
c_2. I no peeking	3.15	6	5
d_1. No, it isn't	3.33	7	5
d_2. I am not a doctor	5.27	8	6

QUESTIONS

Four question utterances were included in the stimuli to see if the judging panel would rate the samples in the same manner as the scoring procedure suggests. These utterances were included because their construct validity had not been demonstrated. The point system for questions was based on the research of Bellugi (1964). She distinguishes two levels of questions:

1. There are no auxiliaries and no subject-verb inversion. There are a few negative questions. All are scored as one point.

2. Yes-no questions contain an auxiliary or some form of do. These are scored as two points (Aux + Nominal + V + ?). The auxiliary component can have an optional negative attachment (Aux + Ng + Nominal + V + ?). Sometimes the auxiliaries are not inverted. The auxiliary is optional in wh questions.

Below is Table 8 showing the question stimuli rated by the observers. The LCI, question, mean scale values, and semi-interquartile range values are listed.

TABLE 8.--Stimuli representing the two levels of questions as scored in the LCI as defined by Bellugi

	Utterance	LCI	?	$\overline{MSV}$	Q
a_1.	Mommy eggnog?	3	1	2.45	0.86
a_2.	What cowboy see?	4	1	2.91	0.38
b_1.	Is mommy talking?	6	2	4.33	0.94
b_2.	What he is writing?	7	2	4.06	0.35

Inspection of these data reveals that the mean scale values rank order themselves in a manner consistent with the two scoring levels operationally defined by Bellugi. To determine the significance of the difference in mean scale values between pairs of stimuli, that is, between a and b, b and c, and so forth, a t test was computed. The results are reported in Table 9. All differences were statistically

significant at the .05 level. On the basis of these results it would appear that observers tend to rate the intricacy of question usage in a manner similar to what Bellugi's research indicates.

TABLE 9.--Values of t for tests of significance of differences in mean scale values between question utterances

Comparisons	$\bar{M}SV$	t
a_1	2.45	5.78*
b_1	4.33	
a_2	2.91	3.53*
b_2	4.06	

*$p < .05 = 2.03$; df = 32

Further research in the form of replication needs to be done to determine the effects of perceived values of questions on observers' ratings, using many more samples than were included here to see what really does occur in this aspect of child language. Again, this investigation was only an attempt to determine how observers perceive this aspect of child language and the results are tentative.

In summary, the results of this investigation indicated that observers can reliably ($r = 0.97$) scale single utterances representing 57 different grammatical structures obtained from child language samples. The correlation (0.87) between LCI scores and observers' judgments of intricacy of

language usage suggests that the LCI is a highly sensitive indicator of the impact child language has on observers. The construct validity of the LCI was demonstrated. At this time no changes in the LCI scoring procedure are indicated. The LCI possesses sufficient reliability and validity characteristics for assessing expressive language abilities in children five years of age and younger.

CHAPTER V

SUMMARY AND CONCLUSIONS

Investigators disenchanted with established methods of evaluating child language have begun testing the usefulness of new measures to analyze verbal output. One such method of language assessment is the length-complexity index (LCI) as first proposed by Shriner (1967). There are discussions in the literature of the LCI scoring procedure (Miner, 1969) and its reliability (Barlow and Miner, 1969; Griffith and Miner, 1969) but its construct validity has not been demonstrated.

Until this time the LCI had not been shown to assess child language as it is perceived by observers. By demonstrating its construct validity a researcher could show the LCI to assess children's language development appropriately. This would involve comparing LCI scoring values and observers' judgments.

The general purpose of this investigation was to assess the construct validity of the LCI. Specifically, the following questions were posed at the outset of this study:

1. Can observers reliably scale single utterances representing 57 different grammatical structures obtained from child language samples?

2. What is the relationship between LCI scores and observers' judgments of intricacy of language?
3. Based on the results of this study, what, if any, changes in the LCI scoring procedure are indicated?

Psychological scaling has been proven useful in assessing children's language development as demonstrated in several recent investigations (Nelson, 1966; Sherman, Shriner, and Silverman, 1965; Shriner, 1967; Shriner and Sherman, 1967; Sherman and Silverman, 1968; and Miner and Silverman, 1969). For these investigations the method of equal-appearing intervals (Edwards, 1957) was used. The psychological scaling method of equal-appearing intervals was chosen for use in this investigation. A seven-point equal-appearing intervals scale of intricacy of language usage was employed with one representing least intricacy and seven representing most intricacy.

Language production increases in length as well as in complexity with increasing chronological age. The weighting system used to assess complexity of response was questioned (Darley and Moll, 1960; Minifie, Darley, and Sherman, 1963) and recent investigations have indicated that the Mean Length Response (MLR) is not a valid measure of language development (Shriner, 1969; Minifie, Darley, and Sherman, 1963). Therefore, a procedure which combines both length and complexity of response into a single measure may prove to be more useful than either of the above measures used

independently (Shriner, 1967). In comparing selected measures with psychological scale values of language development Shriner (1967) found that in the multiple-regression analysis the length-complexity measure remained as the single, best predictor of psychological scale values of language development.

A length-complexity measure was formed (Shriner, 1969) by relying on the research of Menyuk (1964a) and Cazden (1965). It is a linguistic measure designed to make a composite analysis of sentence length and sentence complexity. Both length and complexity are considered together (not independently) according to a numeric weighting system. Barlow and Miner (1969) assessed the temporal reliability of the LCI and the MLR and found the intraclass correlation coefficient for MLR was $r_1 = 0.65$ compared to $r_1 = 0.80$ for the LCI.

Transcripts of tape recorded language samples from the speech of 17 subjects, 10 males and 7 females, within two months of age five were available from another experiment (Barlow and Miner, 1969). Each utterance was analyzed according to the LCI scoring procedures (Miner, 1969). Portions of these utterances served as the stimuli for this investigation. The stimuli rated by the observers consisted of four pairs of grammatically matched utterances randomly selected for each LCI point value one through ten. Each pair of utterances at each point level consisted of a different type of syntactic structure. This constituted an initial

corpus of 80 stimuli. In addition, 22 examples of developmental language acquisition data (Brown and Bellugi, 1964; Cazden, 1965) and 12 examples of the two categories of negatives taken from Miner's (1969) LCI scoring procedures were included. A total of 114 individual utterances were scaled.

The judging panel was comprised of 33 undergraduate students in the Department of Speech at Eastern Illinois University. The stimuli to be scaled were presented visually, via a typed manuscript.

In answer to question one, can observers reliably scale single utterances representing 57 different grammatical structures obtained from child language samples, the resulting r was 0.97 as determined by an intraclass correlation coefficient. This was interpreted to mean that observers can reliably scale single utterances representing 57 different grammatical structures obtained from child language samples.

In answer to question two, what is the relationship between LCI scores and observers' judgments of intricacy of language usage, the resulting r was 0.87 as determined by means of a Pearson Product-Moment correlation coefficient. This suggests that the LCI is a highly sensitive indicator of observers' judgments of intricacy of language usage when those judgments are based upon single utterances.

In answer to the third question, based on the results of this study, what, if any, changes in the LCI scoring procedure are indicated, a Pearson Product-Moment correlation coefficient was used to assess the relationship between mean scale values and the LCI values for the first group of stimuli (N = 80) and the resulting $r = 0.85$. This was interpreted to mean that the two variables rank ordered themselves in approximately the same manner.

The second group of stimuli (N = 34) dealt with developmental sequence of noun phrase emergence, verb phrase emergence, negation, and question utterances as scored according to the LCI. Data dealing with noun phrases revealed that the mean scale values rank ordered themselves in a manner consistent with Cazden's developmental sequence. Also the mean scale values were not consistent with the LCI values computed for the sample.

Negative utterances were included from the LCI scoring procedure to determine if the judging panel would rate the samples in the same manner as the scoring procedure suggests. The data revealed that the mean scale values did not rank order themselves in a manner consistent with the four scoring levels operationally defined by Bellugi.

Question utterances were also included in the stimuli to see if the judging panel would rate the samples in the same manner as the scoring procedure suggests. The data revealed that the mean scale values rank ordered themselves

in a manner consistent with the two scoring levels operationally defined by Bellugi.

It is concluded that at this time on the basis of this investigation, no changes need to be made in the scoring procedure of the LCI until further research is done with larger sample sizes. Further research is need on specific types of verb phrases and negatives. The LCI is a beneficial tool that will aid the speech pathologist in analyzing language development in children five years and younger.

APPENDIX I

Name _____ Age _____ Date _____

Children's Oral Language Samples

____ 1. he jumps

____ 2. what cowboy see

____ 3. her teddy-bear fall

____ 4. I no peeking

____ 5. wanna

____ 6. there's a dish and there's a cup

____ 7. 'cause the dog wanted it

____ 8. he's runnin' back in the garage

____ 9. he's gettin' out of the box

____ 10. gonna

____ 11. the girl, she's lookin' at the kitty-cat

____ 12. the bunny-rabbit got into the dog's food

____ 13. Sally

____ 14. rain

____ 15. an' the dog's hidin' in the flowers

____ 16. he'll splash all over 'em

____ 17. lookin' at the boy

____ 18. he ate it

____ 19. he drops

____ 20. mad

____ 21. 'n' she's takin' a pencil in her book

____ 22. the flower

____ 23. no wash

____ 24. there's a car and a kid fell out

____ 25. I no bit you

____ 26. catch him

____ 27. settin' down

____ 28. I jumping

____ 29. watchin' her

____ 30. me not crying

____ 31. I don't know

____ 32. an' a frog was just settin' there

____ 33. 'cause it's raining

____ 34. it's rainin'

____ 35. paintin'

____ 36. and he's playin' on the swing-set

____ 37. spank him

____ 38. the big dog

____ 39. give it to me

____ 40. I jump

____ 41. she's gonna find the dog

____ 42. he likes it

____ 43. she's gonna, she's gittin' the umbrella away from the dog

____ 44. and a barn, and a tree, and a fence

___ 45. in the house

___ 46. the dog

___ 47. uh--it's a lion-tamer and the lion got out

___ 48. she's takin' the umbrella away from the doggy

___ 49. I dropping

___ 50. 'n--'n--she's lookin' at a book

___ 51. the dog is pullin; on it

___ 52. they're washin' the dog in the pan

___ 53. and she is paintin'

___ 54. I'm jumping

___ 55. hurts

___ 56. my blue flower

___ 57. back home

___ 58. watching him

___ 59. wear mitten no

___ 60. I am not a doctor

___ 61. 'cause he was eatin' his food

___ 62. flower

___ 63. and she's cuttin' a picture for her sister

___ 64. runnin' after

___ 65. he's hidin' under his book

___ 66. he runned home

___ 67. no, it isn't

___ 68. an' the daddy was just standin' there

___ 69. dog

___ 70. the dog got in the bunny's food

____ 71. wake up

____ 72. he's hangin' onto the bucket

____ 73. the kitten, the kitten is lookin' lookin' at it

____ 74. nice flower

____ 75. a blue flower

____ 76. I dropped

____ 77. git

____ 78. I'd go in after him

____ 79. Mommy eggnog

____ 80. they're pastin'

____ 81. pull hard

____ 82. for the rain

____ 83. take him back home

____ 84. her doll fell

____ 85. he's gonna catch a--gonna catch--rabbit

____ 86. my big dog

____ 87. oh--um--walkin' in the rain

____ 88. a flower

____ 89. is mommy talking

____ 90. they're takin' the book off him

____ 91. there's a girl and there's a mom

____ 92. talkin'

____ 93. uh--he's bustin' a puddle of mud

____ 94. he's puttin' a face on the kitty

____ 95. 'cause he's eatin' her food

____ 96. a dog

____ 97. an' the dog's layin' in the flowers

____ 98. she's makin' a rabbit

____ 99. and a dog, and a girl, and a boy

____100. I'm dropping

____101. he's goin' at his bed

____102. I jumped

____103. she's comin' home

____104. big dog

____105. four

____106. she dropped her teddy-bear

____107. Spot's carryin' the umbrella

____108. about paintin'

____109. he wants it

____110. she whipped her dolly

____111. what he is writing

____112. she's chasing him

____113. 'cause the dog went home

____114. I drop

APPENDIX II

Instructions to Observers

You are asked to judge a series of utterances of children's oral language which are presented in written form. You are to judge each sample in relation to a seven-point scale of "Intricacy of Language Usage." Intricacy of language usage, for purposes of this experiment is defined as the ability to string words together for the purpose of conveying information. For example, consider the following four utterances which might be judged to vary with respect to intricacy of language usage as defined here:

a. dog
b. the big dog
c. the big dog is running
d. the big dog is running around the house

It is obvious that these examples vary with respect to type of word order arrangement for purposes of conveying information.

Make your judgment on the basis of each individual utterance. Avoid being influenced by grammatical correctness; for example, "we was" and "we were" while different grammatically do not differ with respect to 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 vocabulary.

The scale is one of equal intervals--from 1 to 7--with 1 representing least intricacy of language usage and 7 representing most intricacy; 4 represents the midpoint between 1 and 7 with respect to intricacy; the other numbers fall at equal distance, along the scale. Do not attempt to place samples between any two of the seven points, but only at these points.

Each language sample is preceded by a number. Your task will be to record your judgment on your answer

sheet to the left of the identifying number of the language sample.

Following there will be 114 utterances to be rated on the 7 point scale. These utterances were obtained by requesting children to respond to picture stimuli. They were also encouraged to speak by asking them questions and by making comments as needed. These questions and comments are not included in the material you are to judge. All of the utterances are in response to the same set of pictures.

Before you record any judgments, read quickly through the 114 utterances in order to acquaint yourself with the experimental task and the range of utterances which you are requested to judge with respect to the intricacy of language usage.

After you have acquainted yourself with the range and the task, make a judgment on every sample. If you are somewhat doubtful, make a guess as to the most suitable scale position.

BIBLIOGRAPHY

Books

Berki, J. The Child's Learning of English Morphology. As cited in: Psycholinguistics: A Book of Readings, edited by S. Saporta. New York: Holt, Rinehart, and Winston, Inc., 1961.

Carroll, J. B. Language Development in Children. As cited in: Psycholinguistics: A Book of Readings, edited by S. Saporta. New York: Holt, Rinehart, and Winston, Inc., 1961.

Downie, N. M. and Heath, R. W. Basic Statistical Methods. New York: Harper and Row, Publishers, 1965.

Edwards, L. Techniques of Attitude Scale Construction. New York: Appleton-Century-Crofts, Inc., 1957.

Garrett, H. E. Statistics in Psychology and Education. New York: Longmans, Green, and Company, 1954.

Koenker, R. H. Simplified Statistics. Bloomington, Illinois: McKnight and McKnight Publishing Company, 1961.

Lyman, H. B. Test Scores and What They Mean. Englewood Cliffs, N. J.: Prentice-Hall, Inc., 1963.

McNeill, D. Developmental Psycholinguistics. As cited in: The Genesis of Language, edited by F. Smith and G. A. Miller. Cambridge: MIT Press, 1966b.

Williams, Frederick. Reasoning with Statistics. Chicago: Holt, Rinehart, and Winston, Inc., 1968.

Winer, J. Statistical Principles in Experimental Design. New York: McGraw-Hill Book Company, 1962.

Articles

Barlow, M. C. and Miner, L. E. "Temporal Reliability of Length-Complexity Index," Journal of Communication Disorders, 2, 1969, 241-251.

Bellugi, U. The Emergence of Inflections and Negation Systems in the Speech of Two Children, paper presented at New England Psychological Association, 1964.

Brown, R. and Bellugi, U. "Three Processes in the Child's Acquisition of Syntax," Harvard Educational Review, 34, 1964, 133-151.

Cazden, C. B. Environmental Assistance to the Child's Acquisition of Grammar, Harvard University, unpublished doctoral dissertation, 1965.

Cowan, P. A., Weber, J., Hoddinott, B. A., and Klein, J. "Mean Length of Spoken Response as a Function of Stimulus, Experimenter, and Subject," Child Development, 38, 1967, 191-203.

Cullinan, W. L., Prather, E. M., and Williams, D. E. "Comparison of Procedures for Scaling Severity of Stuttering," Journal of Speech and Hearing Research, 6, 1963, 187-194.

Darley, F. L. and Moll, K. L. "Reliability of Language Measures and Size of Language Sample," Journal of Speech and Hearing Research, 3, 1960, 166-173.

DeMuth, Blaise. Comparison of Articulation Severity Ratings by Mothers, Teachers, and Speech Pathologists. Eastern Illinois University, unpublished master's thesis, 1969.

Elliott, L., Hirsh, I., and Simmons, A. "Language of Young Hearing-Impaired Children," Language-Speech, 10, 1967, 141-158. As cited in: Sherman, D. and Silverman, F., "Three Psychological Scaling Methods Applied to Language Development," Journal of Speech and Hearing Research, 11, 1968, 837-841.

Griffith, J. and Miner, L. E. "LCI Reliability and Size of Language Sample," Journal of Communication Disorders, 2, 1969, 264-267.

James, S. E. Transformational Abilities of Culturally Advantaged and Disadvantaged Children. Eastern Illinois University, unpublished master's thesis, 1968.

Lee, L. L. "Developmental Sentence Types: A Method for Comparing Normal and Deviant Syntactic Development," Journal of Speech and Hearing Disorders, 31, 1966, 311-330.

Menyuk, P. "Syntactic Rules Used by Children from Preschool Through First Grade," Child Development, 35, 1964a, 533-546.

Miner, L. E. "Scoring Procedures for the Length-Complexity Index: A Preliminary Report," Journal of Communication Disorders, 2, 1969, 224-240.

Miner, L. E. Psychological Scaling Methods as Validity Criterion in Language Tests. University of Illinois, unpublished paper, 1969.

Miner, L. E. and Silverman, F. H. Psychological Scaling of Children's Language Development: Influence of Mode of Stimulus Presentation and Response Length Upon Scale Values. Paper presented at the American Speech and Hearing Convention, 1969.

Minifie, F., Darley, F. L., and Sherman, D. "Temporal Reliability of Seven Language Measures," Journal of Speech and Hearing Research, 6, 1963, 139-149.

Morrison, S. "Measuring the Severity of Articulation Defectiveness," Journal of Speech and Hearing Disorders, 20, 1955, 347-351.

Nelson, D. Tests of the Validity of a Prediction Equation for Assessing Language Development. University of Iowa, unpublished master's thesis, 1966.

Rees, M. "Some Variables Affecting Perceived Harshness," Journal of Speech and Hearing Research, 1, 1958, 155-168.

Sherman, D. and Lewis, D. "Measuring the Severity of Stuttering," Journal of Speech and Hearing Disorders, 16, 1951, 320-326.

Sherman, D. and Linke, E. "The Influence of Certain Vowel Types on Degree of Harsh Voice Quality," Journal of Speech and Hearing Disorders, 17, 1952, 401-408.

Sherman, D. and Moodie, C. E. "Four Psychological Scaling Methods Applied to Articulation Defectiveness," Journal of Speech and Hearing Disorders, 22, 1957, 606-706.

Sherman, D. and Morrison, S. "Reliability of Individual Ratings of Severity of Defective Articulation," Journal of Speech and Hearing Disorders, 20 1955, 352-356.

Sherman, D., Shriner, T., and Silverman, F. H. "Psychological Scaling of Language Development in Children," Proceedings of the Iowa Academy of Science, 72, 1965, 366-371.

Sherman, D. and Silverman, F. H. "Three Psychological Scaling Methods Applied to Language Development," Journal of Speech and Hearing Research, 11, 1968, 837-841.

Sherman, D. and Trotter, W. D. "Correlation Between Two Measures of the Severity of Stuttering," Journal of Speech and Hearing Disorders, 21, 1956, 426-429.

Shriner, T. "A Comparison of Selected Measures with Psychological Scale Values of Language Development," Journal of Speech and Hearing Research, 10, 1967, 828-835.

Shriner, T. "A Review of Mean Length of Response as a Measure of Expressive Language Development in Children," Journal of Speech and Hearing Disorders, 34, 1969, 61-68.

Shriner, T. and Miner, L. E. "Morphological Structures in the Language of Disadvantaged and Advantaged Children," Journal of Speech and Hearing Research, 11, 1968, 605-610.

Shriner, T. and Sherman, D. "An Equation for Assessing Language Development," Journal of Speech and Hearing Research, 10, 1967, 41-43.

Silverman, F. H. "An Approach to Determining the Number of Judges Needed for Scaling Experiments," Perceptual Motor Skills, 27, 1968, 1333-1334.

Silverman, F. H. "Correspondence Between Mean and Median Scale Values for Sets of Stimuli Scaled by the Method of Equal-Appearing Intervals," Perceptual Motor Skills, 25, 1967, 727-728.

Stevens, S. S. "Measurement, Statistics, and the Schemapiric View," Science, 161, 1968, 849-856.

Young, M. "Observer Agreement: Cumulative Effects of Rating Many Samples," Journal of Speech and Hearing Research, 12, 1969, 135-143.

Young, M. "Observer Agreement: Cumulative Effects of Repeated Ratings of the Same Samples and of Knowledge of Group Results," Journal of Speech and Hearing Research, 12, 1969, 144-155.

Young, M. and Downs, T. "Testing the Significance of the Agreement Among Observers," Journal of Speech and Hearing Research, 11, 1968, 5-17.

Manuals

Dunn, L. M. Manual for the Peabody Picture Vocabulary Test. Minneapolis: American Guidance Service, 1965.

McCarthy, D. A. Language Development in Children. As cited in: Manual of Child Psychology, edited by L. Carmichael. New York: Wiley and Sons, 1954.

Warner, W. L., Meeker, M., and Eells, K. Social Class in America (a manual of procedure for the measurement of social status). Chicago: Science Research Association, 1949.

Wepman, J. M. and Hass, W. A Spoken Word Count (Children--ages 5, 6, and 7). Chicago: Language Research Associates, 1969.