DOCUMENT RESUME

ED 098 536 CS 001 459

AUTHOR	Harris, Albert J.; Jacobson, Milton D.
TITLE	Revised Harris-Jacobson Readability Formulas.
PUB DATE	Oct 74
NOTE	14p.; Paper presented at the Annual Meeting of the College Reading Association (18th, Bethesda, Maryland, Oct. 31-Nov. 2, 1974)
EDRS PRICE	MF-$0.75 HC-$1.50 PLUS POSTAGE
DESCRIPTORS	Basic Reading; Elementary Grades; *Readability; Reading; *Reading Level; Reading Materials; *Reading Research
IDENTIFIERS	*Harris Jacobson Readability Formula

ABSTRACT

This paper summarizes the work to date on the revised Harris-Jacobson Readability Formulas. The contents include "The Criterion," which discusses the criterion used in the development of the formulas; "Variablos Employed," which includes percent of uncommon words, average sentence length, percent of long words, mean number of letters per word, and spelling patterns; "Validity of the Individual Variables," which presents the correlations obtained with basal reader levels; "Readability Formulas Based on Multiple Correlations," which discusses two-variable and three-variable combinations which have been tried out and reports on the five best formulas; "Corrected Grade Equivalents," which discusses the corrections used for the Harris-Jacobson formula; "Practical Utility," which looks at the practical features of the readability formula; and "Further Research Needed," which discusses a planned testing program and improving the readability formula. (WR)

Albert J. Harris
Wellington G233, Century Village
West Palm Beach, Florida 33401

and

Milton D. Jacobson
Bureau of Educational Research
University of Virginia
Charlottesville, Virginia 22903

REVISED HARRIS-JACOBSON READABILITY FORMULAS

(College Reading Association, Bethesda, Maryland, Oct.31, 1974)

A preliminary report on the Harris-Jacobson Readability Formulas was given in May, 1973, at a convention of the International Reading Association. Since that report much additional work has been completed, and new formulas have been developed. The present paper summarizes the work to date.

The Criterion

Readability has been defined for our work as those characteristics of reading material that make it easy or difficult to comprehend. The criterion used in developing the formulas is based on the average characteristics of six popular series of basal readers. These series had been used in the development of the Harris-Jacobson Basic Elementary Reading Vocabularies (1972). Computer processing made it possible to use a large number of samples. From primer level up, ten about equally spaced samples were chosen from each book, each sample having slightly more than 200 words. At preprimer level as many 200-word samples were taken as the three preprimers of a series could provide. There were 661 samples, totalling about 135,000 words. The samples at a given level were all given the same reading scale value and provided a scale ranging from 1.2 to 6.5. Since the samples at a reader level are not all actually equal in difficulty, there is some inaccuracy in this criterion. In the criterion scale there are seven steps at primary levels but only three steps for grades 4-6.

Variables Employed

Variable 1. (V1) Per cent of uncommon words. Three word lists based on the Harris-Jacobson word lists were tried out. One list, which contained only the 335 first grade words and their inflected forms, was discarded when it was found to lack discriminative ability above first grade. The Short Readability List contains all first-grade and second-grade Core and Additional words and their common inflected forms. It includes 912 root words and 1,880 inflected forms, totalling 2,792 words. The first variable, V1 in the formulas, is the per cent of unique words not in this Short List. Unique means that a word not in the list is counted only once per sample regardless of how many times it may occur in that sample. Unique words was found to have slightly better predictive ability than total number of uncommon words.

A third list, containing all first, second, and third grade words, included 1,925 root words and 4,076 inflected forms, totalling 6,001 words. This was found to be not quite as discriminative at primary levels as the Short List.

For the middle grades the two word lists were about equal, and there were strong indications that for formulas intended to work above sixth grade level, the Long List would be better. Only the Short List is used in the formulas described in this paper.

Variable 2 (V2). Average Sentence Length. The second variable used in the formulas is obtained by dividing the total number of words in a sample by the number of sentences, providing the mean number of words per sentence. This variable is also used in the Lorge, Dale and Chall, and Spache Formulas:

Variable 3 (V3). Per Cent of Long Words. Several ways of measuring word ifficulty directly were tried out. These included counting the number of words that contain more than five letters and dividing by the number of words. A word with six or more letters is considered a long word. This variable is quickly and easily scored by hand as well as by computer.

Variable 4 (V4). Mean Number of Letters per Word. Another way to measure word difficulty is based on the assumption that the longer a word, the harder it is likely to be. Average number of letters per word can be scored almost instantaneously by computer, but is slow and laborious to score by hand. Since it turned out to be slightly inferior to V3 in predictive power, it does not appear in our formulas.

Variable 5 (V5). Spelling Patterns. In our preliminary work we discovered that the per cent of words beginning with the letter e has a substantial correlation with the difficulty of primary reading materials. Dr. Jacobson located over 1,000 spelling rules about spellings which occur characteristically at the beginning, end, or in the middle of words, mainly from Hanna et al. (1966). These rules were combined to form 101 spelling patterns. Dr. Jacobson developed computer programs for identifying and counting these patterns in a sample of reading material, and correlated all 101 patterns with the criterion of reader level. In a recently published paper (Jacobson, 1974) he reported that a combination of 37 spelling patterns correlated .92 with primary reading difficulty.

Since then, Jacobson has located 12 spelling patterns which, when combined by multiple correlation, provide amazingly high correlations with reading level. The 12 best patterns at the primary level have only one pattern in common with the 12 best patterns for grades 4-6. The per cent of words ending with a single letter l increases across the full range from preprimer through sixth reader. Most of the other patterns are effective at primary or middle-grade level but not at both. There will be further discussion of this variable later in this paper.

Validity of the Individual Variables

The first-order Pearson r's with basal reader levels are shown in Table 1, for grades 1-3, and for grades 1-6. Correlations were also obtained for grades 4-6 and grades 3-6. Since there was only one reading level per grade in grades 4-6, the criterion scale provided only three steps for those grades and only five steps with third grade added; the correlations with those very coarse criterion scales were low to moderate.

It may be noted that V1 surpasses V2 for grades 1-6, but V2 has the higher correlation for grades 1-3. V5, which has the highest correlations at both levels, is at present scorable only by computer. Of the three measures of word difficulty that are scorable by hand, V1 and V3 have approximately equal *r*'s for grades 1-3 and V1 is clearly the best for grades 1-6. Variable 3 is consistently superior to V4.

The relationships of Variables 1, 2, and 3 to reader levels are shown in Figures 1, 2, and 3, while that for variable 5 is shown in figures 4 (primary), 5 (elementary) and 6 (grades 1-6). At each reader level the steeper the slope of the line and the smaller the standard deviation, the better the discriminative power of the variable. Variable 1 (Fig. 1) starts off poorly at first grade level but does well over the rest of the range. Variable 2 (Fig. 2) shows a fairly consistent upward slope except for poor discrimination between high third level and fourth reader level. Variable 3 shows steady upward progression but comparatively large standard deviations. Variable 5 shows steady upward progression for the primary and elementary grades with lesser discrimination between the high third level and fourth level readers.

Readability Formulas Based on Multiple Correlations.

The predictive power of a combination of variables depends not only on the correlation of each variable with the criterion, but also on the size of the correlations between the variables; the lower the correlations among variables, the greater the benefit obtained from combining them. A large number of two-variable and three-variable combinations have been tried out, and the five best formulas are reported here. The five formulas in regression equation form are as follows:

Formula 1. Readability level = .094 V1 + .168 V2 + .502
Formula 2. Readability level = .140 V1 + .153 V2 + .560
Formula 3. Readability level = .158 V2 + .055 V3 + .355
Formula 4. Readability level = .070 V1 + .125 V2 + .037 V3 + .497
Formula 5. Readability level = .118 V1 + .134 V2 + .032 V3 + .424

The combination of V1 and V2 provides efficient readability formulas both for primary-grade material (Formula 1) and middle-grade material (Formula 2). These formulas employ the same variables but give them somewhat different weights. Data on the validity and reliability of the five formulas are given in Table 2. Formulas 1, 3, and 4 are for use with materials thought to be below fourth reader level. Formulas 2 and 5 cover the range from preprimer through six but are recommended for use only when the material is thought to be above third grade in difficulty.

The multiple correlation coefficients of the five formulas are shown in the first column of Table 2. The two three-variable formulas are slightly higher in the value of R than the three two-variable formulas. All of the R's are quite similar, ranging only from .888 to .918.

The per cent of the total variance accounted for by a multiple correlation is indicated by R^2, which is shown in the second column of Table 2. Of the three primary-level formulas, Formula 4 is 3.6 per cent better than Formula 1, and Formula 1 is 1.9 per cent better than Formula 3. The standard errors of estimate, in column 3, are in the same order, with Formula 4 having the smallest error, Formula 1 next, and Formula 3 the largest of the three.

Which of these formulas to use for hand computation depends on one's priorities. When maximum validity is more important than speed, the three-variable Formula 4 is the obvious choice. When speed of scoring and computation is most important, Formula 3 may be chosen. For all-around efficiency combining good validity with next-best speed and ease of use, Formula 1 should be preferred.

Formula 2 and Formula 5 are intended for use when the readability of the material is probably above third grade. In terms of R^2, Formula 5 is only .9 per cent better than Formula 2, and the standard errors are very similar. Formula 2 uses two variables and Formula 5 requires three variables. The substantial additional time for scoring and computation that Formula 5 requires does not seem worth the very slight gain in validity. For most hand computations Formula 2 is recommended for middle-grade material.

For computerized computation and scoring the three-variable formulas are preferable to the two-variable formulas. Adding a fourth variable does not provide any further improvement in validity.

The reliabilities of the formulas are also shown in Table 2. They are all .92 or better, indicating very satisfactory reliability. We recommend taking five samples from a book, or three samples from a short selection. The average readability score obtained in either case should be very reliable. These reliabilities were obtained by separating the samples into random halves, getting the correlations between the two sets of formula scores, and applying the Spearman-Brown Formula.

Variable 5, the Spelling Patterns variable, is not included in Table 2; its correlations with the criterion are shown on the bottom line of Table 1. The 12 best spelling patterns for grades 1-3 correlate .93 with the criterion, higher than any of the correlations in Table 2. The 12 best patterns for grades 1-6 correlate .913 with the criterion, equal to Formula 4 and better than the other four formulas. Scoring and computing this variable by hand would be prohibitively time-consuming. We plan to try to simplify this variable further. If it can be reduced to four or five components, hand scoring may become feasible, although it will still be quite laborious. Meanwhile use of this variable requires employment of the special computer programs developed by Jacobson. Adding any two of the other four variables does not further improve the correlation of spelling patterns with basal reader levels.

The reliabilities of the five formulas are shown in the right-hand column of Table 2, and range from .916 to .947. Since the validity coefficients are almost as high as the reliabilities, it would be necessary to raise the reliabilities still higher to achieve further increases in validity.

We have also tried correlating our variables with the 50 per cent and 70 per cent comprehension scores of the McCall, Crabbs Standard Test Lessons in Reading, the criterion used by Dale and Chall. The correlations were slightly but consistently higher with the 50 per cent criterion. A combination of four variables including Spelling Patterns gives a multiple correlation with McCall, Crabbs of .74, which is about 8 per cent better than the correlation reported by Dale and Chall for their formula. However, this readability formula requires the use of special computer programs.

It is our impression, after working with the McCall, Crabbs exercises, that the grade scores for many of them are inaccurate; the exercises should be re-standardized to provide a better readability criterion.

Corrected Grade Equivalents. Whenever predicted scores are obtained from a regression equation, the predicted scores are less variable than the criterion scores. The low scores are not as low, and the high scores are not as high. Dale and Chall found it necessary to provide a correction table for their formula scores. For example, a Dale-Chall obtained score between 7.0 and 7.9 is interpreted as indicating ninth to tenth grade reading difficulty.

We have found it necessary to provide corrections also, although our corrections are not as large. The corrections for the five H-J formulas are shown in Table 3. To use this table, first locate the column for the formula used. Next, locate the score interval into which the obtained formula score falls. Finally, read horizontally to the left to find the readability level corresponding to that score interval. For example, a Formula 5 score of 4.10 falls in the interval 3.77 to 4.23, which corresponds to high third readability level.

Practical Utility

We have confirmed the well-established finding that the most valid single indicator of readability is a good measure of vocabulary difficulty. Our best measure of vocabulary difficulty is Spelling Patterns, but that variable is at present scorable only with the use of special computer programs. Of the other measures of vocabulary we have tried, the per cent of words not found in the H-J Short List is the most valid, with per cent of long words only slightly behind.

Since our formulas provide scores which are not greatly different in validity from those obtained with the Spache and Dale-Chall Formulas, the question of practical utility becomes important. The H-J Short List has slightly fewer root words than the revised Spache List. In checking whether a word is familiar or unfamiliar the V1 score is substantially faster and easier to obtain than the corresponding Spache score, because a word either is or is not in the list, while the Spache List requires the application of 11 rules. V3, the per cent of long words, is even faster to obtain and entails only a slight loss in validity. Similarly, the Short List is only 30 to 35 per cent as long as the Dale List, which requires the application of 22 rules.

Further Research Needed

The formulas presented here, like the other readability formulas in common use, are based on characteristics of samples of graded reading material. It is assumed that the ability of children to understand the material is directly and closely related to these characteristics. This assumption needs to be tested.

We are planning a testing program in which sample McCall, Crabbs exercises will be re-standardized, and the average comprehension scores of children on them will provide another and perhaps better criterion for validating or improving our readability formulas.

Full directions for using the Harris-Jacobson Readability Formulas 1 and 2, including a copy of the Short List used in V1, are given in a book to be published in February, 1975 (Harris and Sipay, 1975). Those who may be interested in using any of the H-J computerized readability formulas are invited to get in touch with Dr. Jacobson.

REFERENCES

Dale, Edgar, and Chall, Jeanne S. A formula for predicting readability. Educational Research Bulletin (Ohio State University), Jan. 21 and Feb. 17, 1948, 27, 11-20, 37-54.

Hanna, Paul R., Hanna, Jean S., Hodges, Richard E., and Rudorf, Edwin H. Phoneme-grapheme correspondences as cues to spelling improvement. Washington, D.C.: Office of Education, Department of Health, Education, and Welfare, 1966.

Harris, Albert J., and Jacobson, Milton D. Basic Elementary Reading Vocabularies. New York: Macmillan Publishing Co., 1972.

Harris, Albert J., and Sipay, Edward R. How to increase reading ability, 6th Ed. New York: David McKay Co., 1975.

Jacobson, Milton D. Predicting reading difficulty from spelling. Spelling Progress Bulletin, Spring 1974, 14, 8-10.

Lorge, Irving. Predicting readability. Teachers College Record, March 1944, 45, 404-419.

Spache, George D. Good reading for poor readers, Rev. 1974. Champaign, Ill.: Garrard Publishing Co., 1974.

Table 1. Correlation coefficients of readability variables with basal reader levels

Variable	Grades 1-3	Grades 1-6
V1. % of unique words in H-J Short List	.797	.868
V2. mean number of words per sentence	.831	.794
V3. % of words having more than 5 letters	.814	.795
V4. mean number of letters per word	.736	.739
V5. spelling patterns	.930	.919

Table 2. Validity and reliability of five Harris-Jacobson Readability Formulas

	Validity			Reliability
	R	R^2	SEest	(Spearman-Brown)
Formula 1	.898	.807	.384	.934
Formula 2	.904	.817	.714	.944
Formula 3	.888	.788	.402	.916
Formula 4	.918	.843	.347	.941
Formula 5	.909	.826	.698	.947

Table 3. Corrected grade equivalents for predicted scores on five H-J Readability formulas

Readability Level	Predicted Score		
	Formula 1	Formula 3	Formula 4
Preprimer (1.0 - 1.34)	1.0 - 1.53	1.0 - 1.48	1.0 - 1.49
Primer (1.35 - 1.64)	1.54 - 1.74	1.49 - 1.80	1.50 - 1.74
First reader (1.65 - 1.99)	1.75 - 1.98	1.81 - 2.15	1.75 - 2.04
Low second (2.00 - 2.49)	1.99 - 2.37	2.16 - 2.57	2.05 - 2.47
High second (2.50 - 2.99)	2.38 - 2.84	2.58 - 2.90	2.48 - 2.89
Low third (3.00 - 3.49)	2.85 - 3.30	2.91 - 3.18	2.90 - 3.30
High third (3.50 - 3.99)	3.31 - 3.74	3.19 - 3.40	3.31 - 3.74
Fourth and up (4.00 +)	3.75 and up	3.41 and up	3.75 and up

Readability Level	Predicted Score	
	Formula 2	Formula 5
Preprimor (1.0 - 1.34)	1.0 - 1.63	1.0 - 1.57
Primor (1.35 - 1.64)	1.64 - 1.83	1.58 - 1.80
First reador (1.65 - 1.99)	1.84 - 2. 07	1.81 - 2. 08
Low second (2.00 - 2.49)	2.08 - 2.42	2.09 - 2.50
High second (2.50 - 2.99)	2.43 - 2.98	2.51 - 3.07
Low third (3.00 - 3.49)	2.99 - 3.70	3.08 - 3.76
High third (3.50 - 3.99)	3.71 - 4.21	3.77 - 4.23
Fourth (4.00 - 4.99)	4.22 - 4.80	4.24 - 4.81
Fifth (5.00 - 5.99)	4.81 - 5.28	4.24 - 5.30
Sixth (6.00 - 6.99)	5.29 - 5.67	5.31 - 5.73
Seventh (7.00 - 7.99)	5.68 - 6.05	5.74 - 6.08
Eighth and up (8.00 +)	6.06 and up	6.09 and up

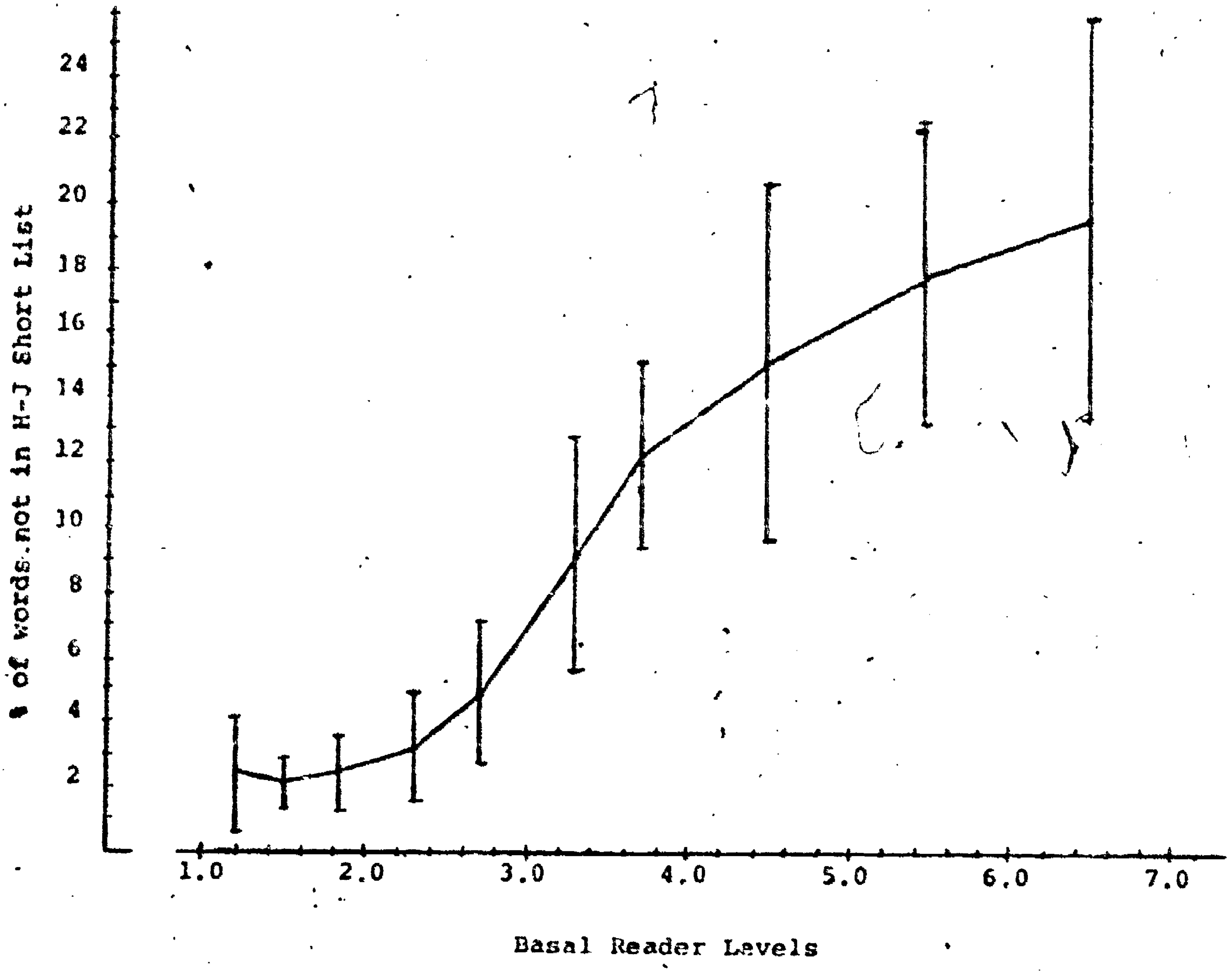

Fig. 1. Variable 1. Means and standard deviations for percent of words not in the H-J Short List

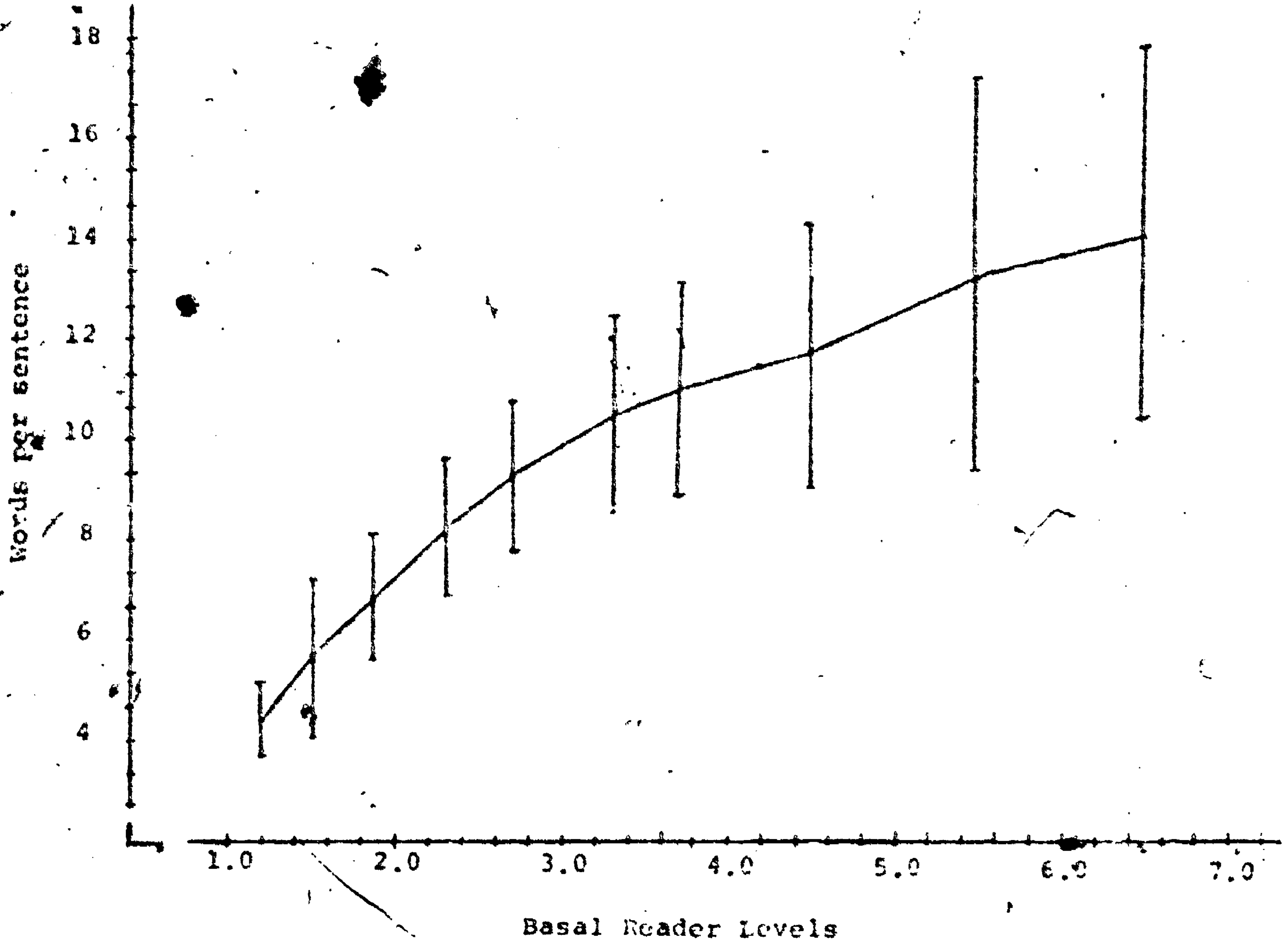

Fig. 2. Variable 2. Means and standard deviations for mean number of words per sentence.

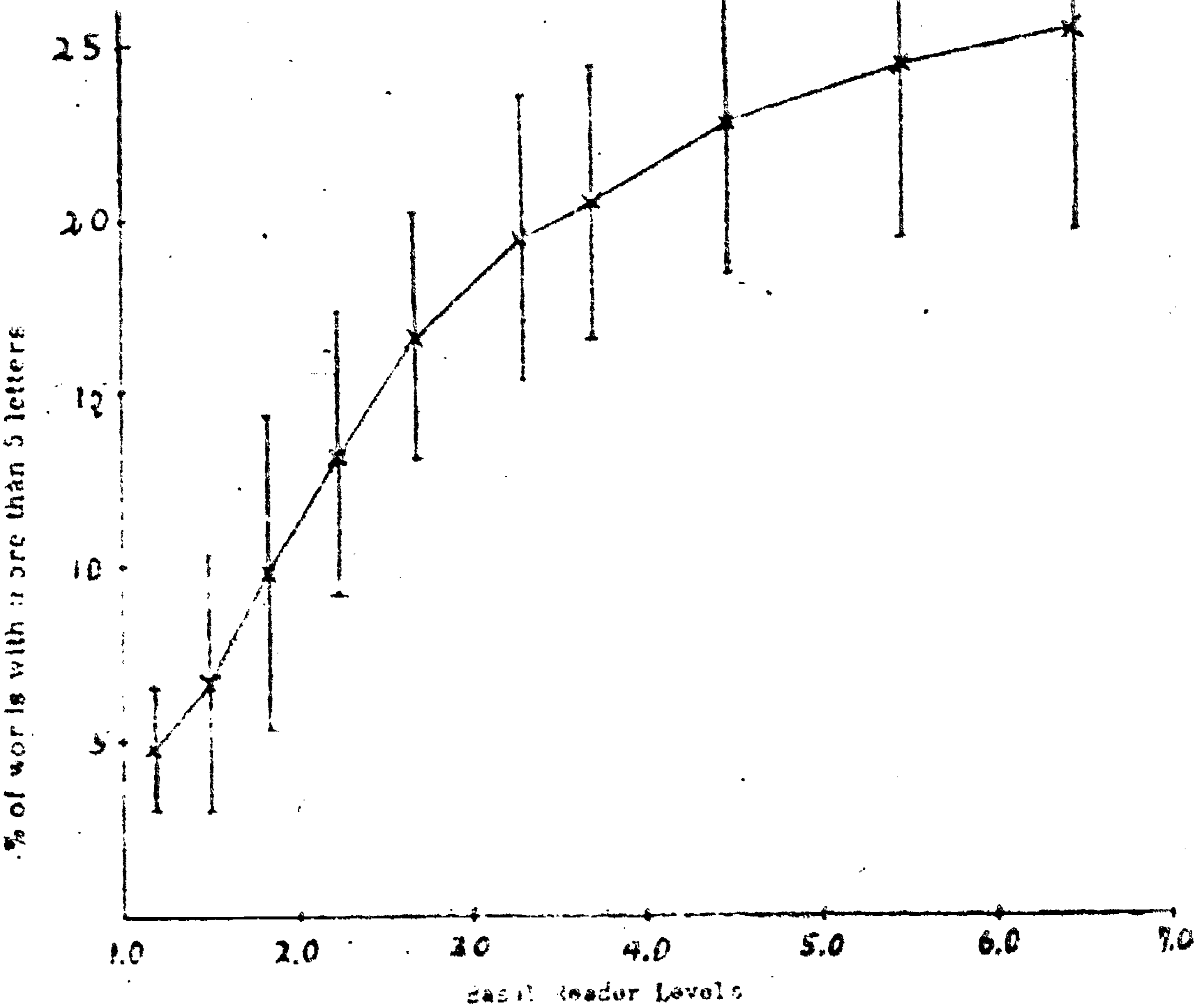

Fig.3. Variable 3. Means and standard deviations for per cent of words with more than five letters.

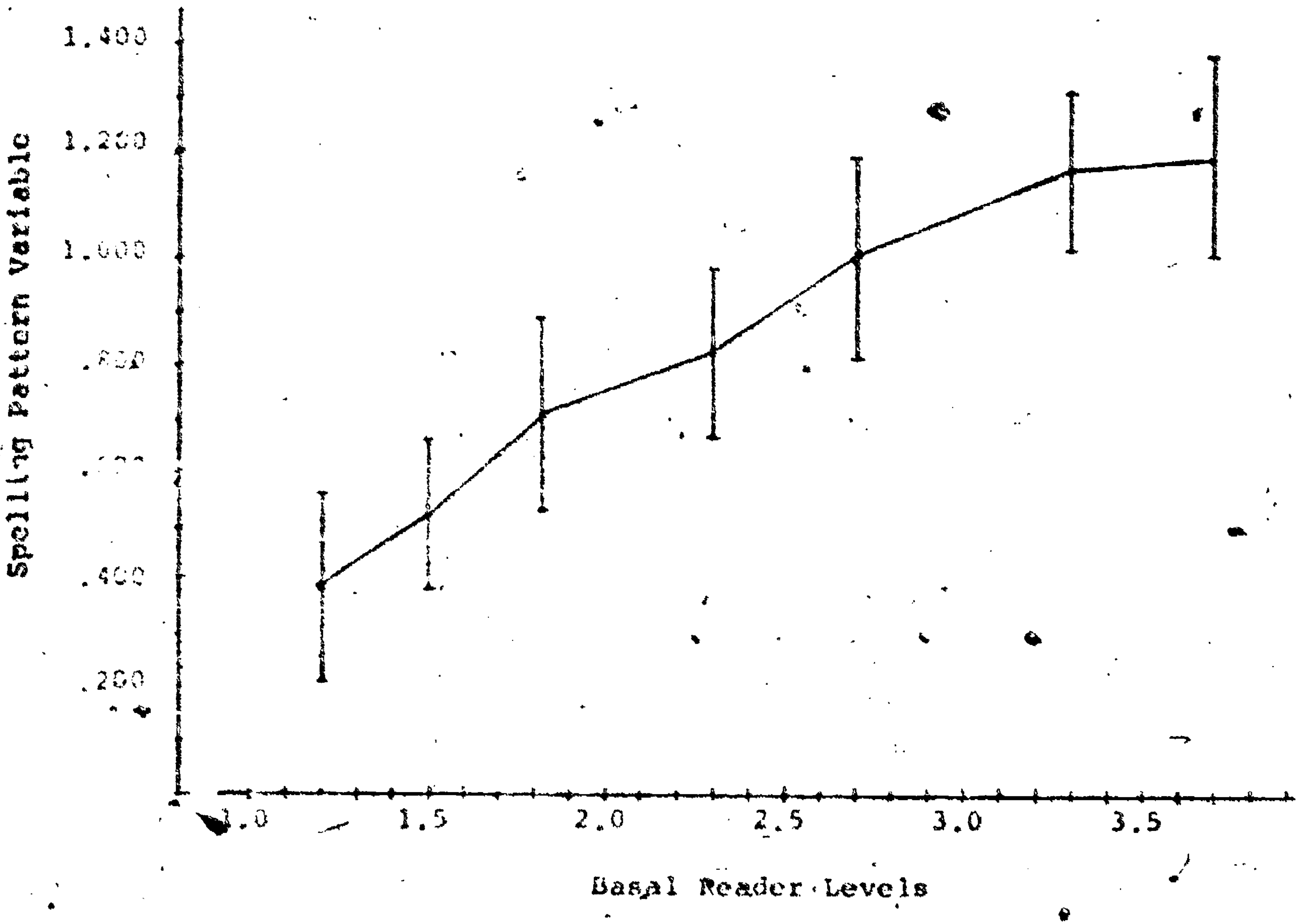

Fig. 4. Variable 5. Means and standard deviations for Spelling Pattern Variable, Primary 1-3

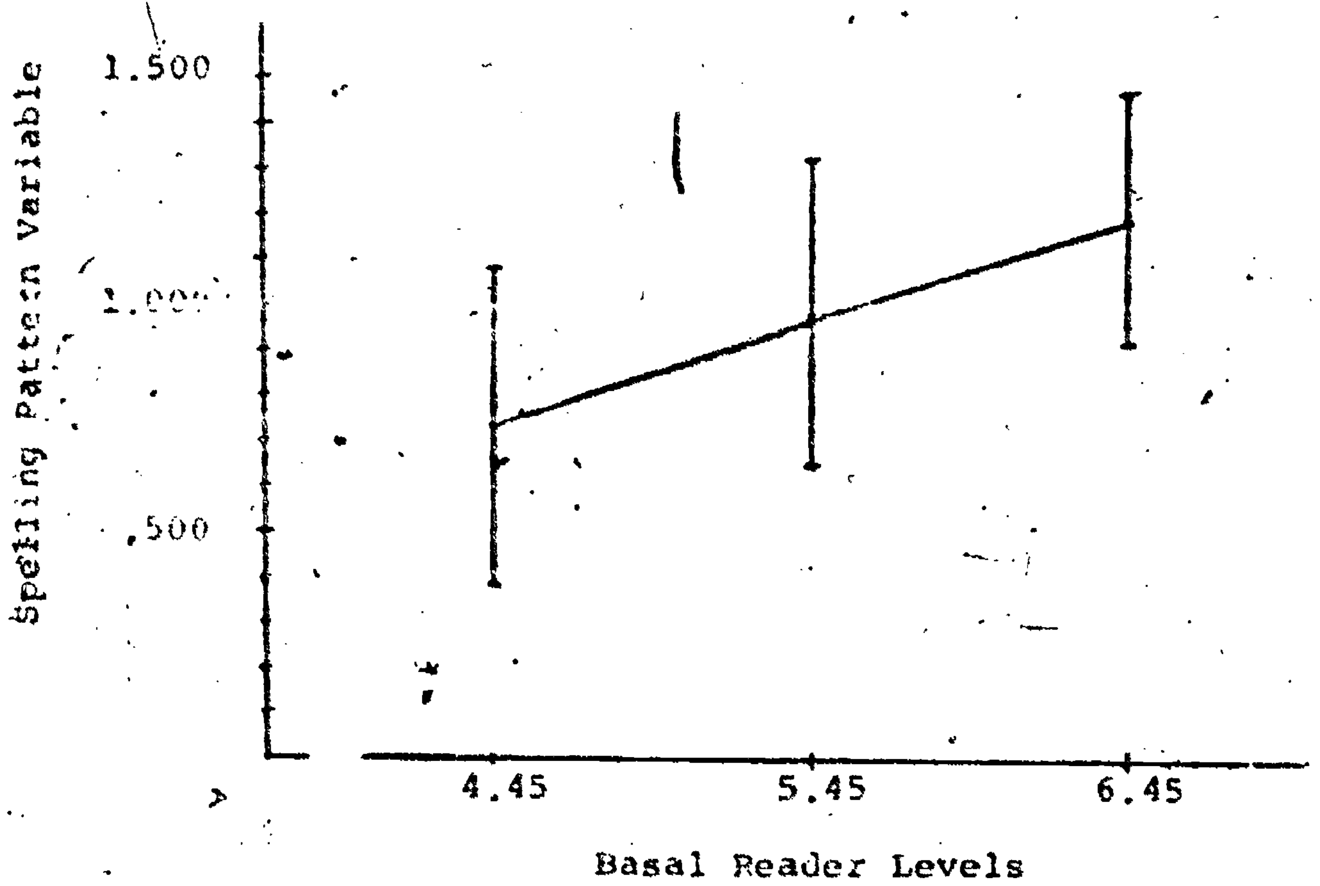

Fig. 5. Variable 5. Means and standard deviations for Spelling Pattern Variable, Levels 4-6

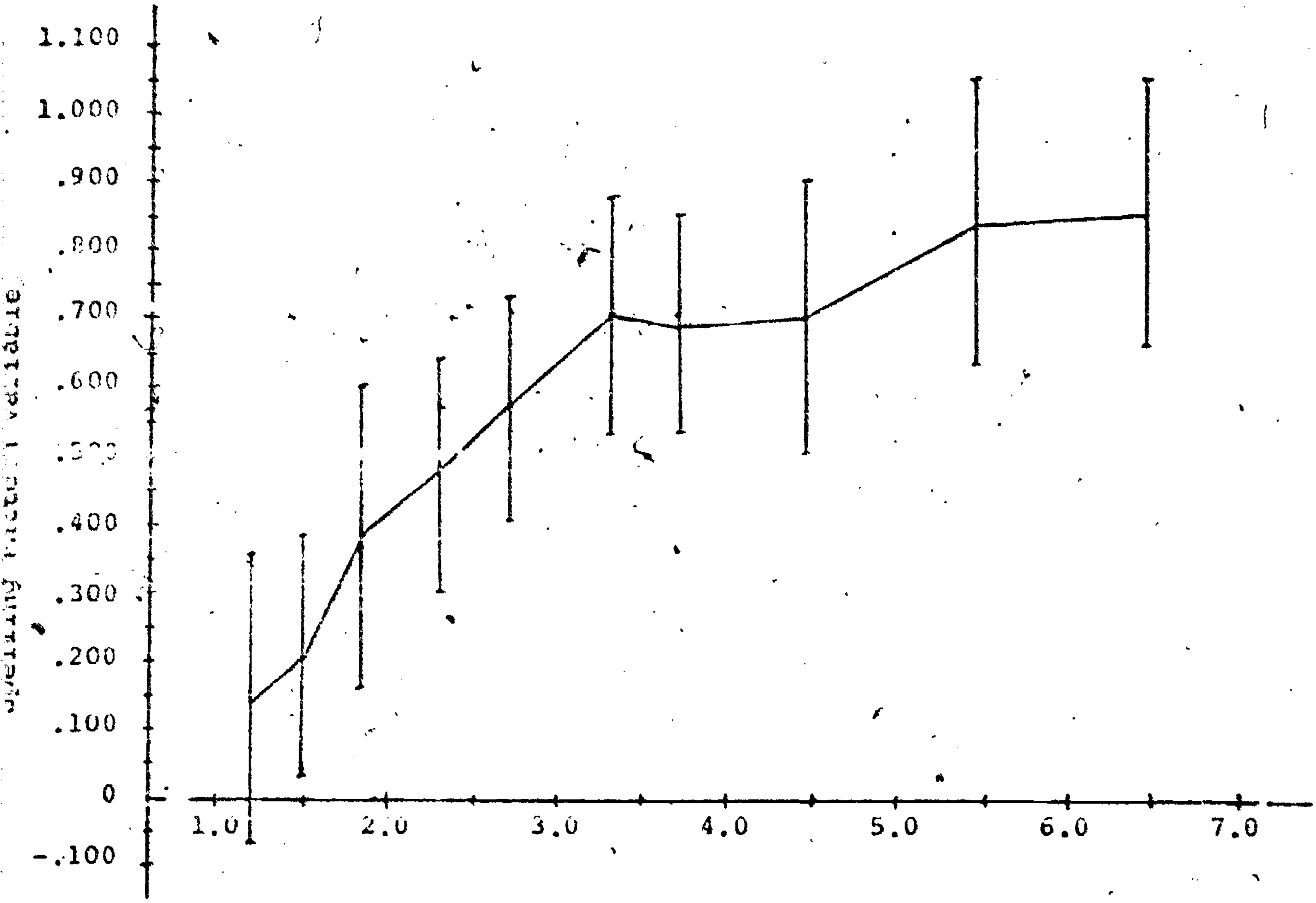

Fig. 6. Variable 5. Means and standard deviations for Spelling Pattern Variables, Levels 1-6