DOCUMENT RESUME

ED 095 900 IR 001 082

AUTHOR Kauffman, Dan; And Others
TITLE The Empirical Derivation of Equations for Predicting Subjective Textual Information. Final Report.
INSTITUTION Air Force Human Resources Lab., Williams AFB, Ariz. Flying Training Div.
REPORT NO AFHRL-TR-74-55
PUB DATE Jul 74
NOTE 40p.

EDRS PRICE MF-$0.75 HC-$1.85 PLUS POSTAGE
DESCRIPTORS *Education; *Information Theory; *Instructional Technology; *Psychology; *Training

ABSTRACT

A study was made to derive an equation for predicting the "subjective" textual information contained in a text of material written in the English language. Specifically, this investigation describes, by a mathematical equation, the relationship between the "subjective" information content of written textual material and the relative number of errors committed by a learner when asked to predict, letter by letter, the content of given textual material. The relationship shows that the subjective information of a given text for a specific learner is directly proportional to the number of wrongly-guessed signs made by that learner. This is expressed mathematically by: I=3.1E; Where: I=Information in Bits; E=Number of wrongly guessed signs. The application of Shannon's guessing procedure (1951) in this study permits the measurement of the "subjective" information of a given text for a specific learner. The derived equation permits the measurement of information in terms of a value that is dependent not only on the inherent qualities of the subject matter, but also on the internal state of the learner. (Author/WCM)

AFHRL-TR-74-55

AIR FORCE

HUMAN RESOURCES

THE EMPIRICAL DERIVATION OF EQUATIONS FOR PREDICTING SUBJECTIVE TEXTUAL INFORMATION

By

Dan Kauffman
Mike Johnson
Gene Knight
Arizona State University
Tempe, Arizona 85281

FLYING TRAINING DIVISION
Williams Air Force Base, Arizona 85224

July 1974

LABORATORY

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When US Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Arizona State University, Tempe, Arizona 85281, under contract F41609-71-C-0027, project 1123, with the Flying Training Division, Air Force Human Resources Laboratory (AFSC), Williams Air Force Base, Arizona 85224. Captain Gary B Reid, Flying Training Division, was the contract monitor.

This report has been reviewed and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

This technical report has been reviewed and is approved.

WILLIAM V. HAGIN, Technical Director
Flying Training Division

Approved for publication.

HAROLD E. FISCHER, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER AFHRL-TR-74-55	2. GOVT ACCESSION NO.
3. RECIPIENT'S CATALOG NUMBER	
4. TITLE (and Subtitle) THE EMPIRICAL DERIVATION OF EQUATIONS FOR PREDICTING SUBJECTIVE TEXTUAL INFORMATION	5. TYPE OF REPORT & PERIOD COVERED Final
	6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Dan Kauffman Mike Johnson Gene Knight	8. CONTRACT OR GRANT NUMBER(s) F41609-71-C-0027
9. PERFORMING ORGANIZATION NAME AND ADDRESS Arizona State University Tempe, Arizona 85281	10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 11230205
11. CONTROLLING OFFICE NAME AND ADDRESS Hq Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235	12. REPORT DATE July 1974
	13. NUMBER OF PAGES 40
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office) Flying Training Division Air Force Human Resources Laboratory Williams Air Force Base, Arizona 85224	15. SECURITY CLASS. (of this report) Unclassified
	15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

This report was approved (74-58, 9 April 1974) for publication in the Instructional Science, an international journal.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

information theory
psychology
instructional technology
education
training

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Parameters pertaining to information processing by human beings have, in the past, been determined by learning and memory experiments with nonsense syllables, number sequences, etc. However, in the real world we are concerned with the processing of "meaningful information" not senseless texts. Therefore, the determination of *subjective* information directly from meaningful material becomes extremely important in the instruction-learning process.

This study derives an equation for predicting the *subjective* textual information contained in a text of material written in the English language. Specifically, this investigation describes, by a mathematical equation, the relationship between the *subjective* information content of written textual material and the relative number of errors

DD FORM 1473 1 JAN 73 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 (continued)

committed by a learner when asked to predict, letter by letter, the content of given textual material. This relationship shows that the subjective information of a given text for a specific learner is directly proportional to the number of wrongly-guessed signs made by that learner. This is expressed mathematically by

$$I \sim 3.1\,E$$

Where: I = Information in Bits

E = Number of wrongly guessed signs

The application of Shannon's guessing procedure (1951) in this study permits the measurement of the *subjective* information of a given text for a specific learner. Unlike senseless texts, the *subjective* information of a meaningful text varies from learner to learner. Therefore, the derived equation permits the measurement of information in terms of a value that is dependent not only upon the inherent qualities of the subject matter, but also upon the internal state of the learner.

The derived equation for the English language is then compared to an equation derived by Weltner (1967) for the German language and found to be remarkably similar.

PREFACE

This study was conducted under Project 1123, USAF Flying Training Development, Task 112302, Instructional Innovations in USAF Flying Training.

The research was carried out under the provisions of contract F41609-71-C0027 by the Educational Technology Department of Arizona State University, Tempe, Arizona 85281. Contract monitor was Captain Gary B. Reid.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

THE EMPIRICAL DERIVATION OF EQUATIONS FOR PREDICTING SUBJECTIVE TEXTUAL INFORMATION

Introduction

Objective

This study was designed to empirically derive an equation for predicting the subjective textual information contained in a text of material written in the English language. Specifically, this investigation describes, by a mathematical equation, the relationship between the subjective information content of written textual material and the relative number of errors committed by a learner when asked to predict, letter by letter, the content of the textual material.

Rationale

Parameters pertaining to information processing by human beings have, in the past, been determined by learning and memory experiments with nonsense syllables, number sequences, etc. Because of the relative simplicity of these types of experiments, the flow of information and the subsequent information processing of these senseless texts could be measured and varied according to exact prescriptions. However, in the real world we are concerned with the processing of "meaningful information," not senseless texts. When the parameters are determined from senseless texts uncertainty is contained in the extrapolation to meaningful materials. Therefore, if we could determine the subjective information directly from meaningful material, we would most certainly reduce these uncertainties. As a result, one would be able to answer questions of the following type:

1. What is the amount of information contained in the laws of thermodynamics for a specific learner?

2. How much information does a learner gain, in a certain period of instruction, under a given set of instructional conditions?

3. How great is the flow of information through a particular instruction-learning channel? Is it too much? Is it too little?

4. How much does the learner already know?

5. How much does the learner need to know?

Instruction Learning Process

Instruction and learning are two vital aspects of the educational process that depend upon the manipulation of three activities: the input of meaningful information; the processing of this information; and the output of meaningful information. These three activities form the basis for the science of information theory.

The science of information theory, since it deals with the fundamental processes involved in the instruction-learning process, provides the quantitative instruments needed to describe this process.

When we discuss the instruction-learning process we are essentially describing a communication system. Typically, a communication system can be described in terms of the following (see Figure 1):

Information source: The mechanism that selects a desired message out of a set of possible messages.

Transmitter: Encodes the message into a signal.

Channel: The medium through which the signal is transmitted.

Receiver: Accepts and decodes the transmitted signal into the message.

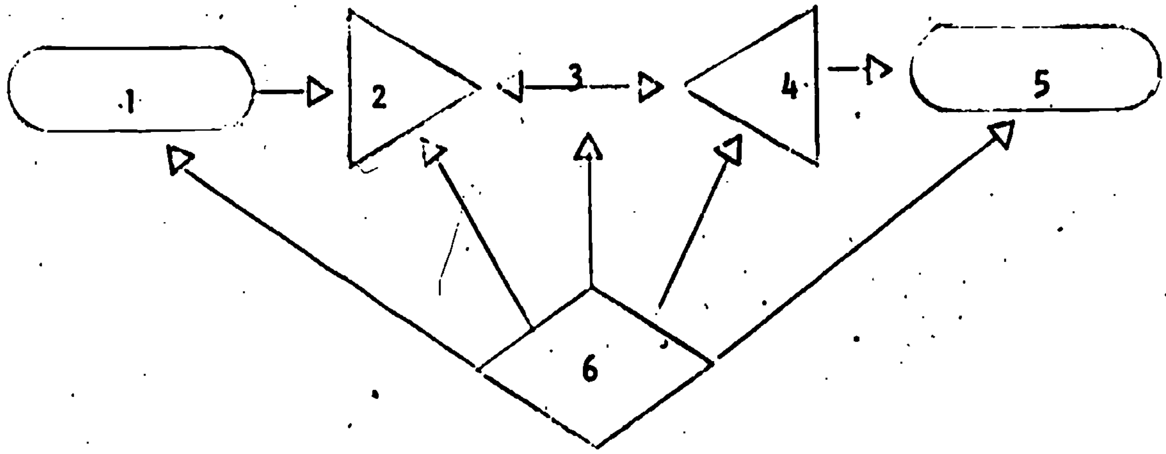
Destination: Interprets the message.

Noise: Unwanted additions imposed on the message. These additions can be either from an external source, internal source or combinations of the two. Noise changes the intended message.

Quantifying Information

Since instruction is concerned with the transference of information and learning is the act of (or the result of) the information transfer, it is obvious that we are not only describing a communication system, but also an instruction-learning system. The information source and transmitter being the teacher, textbook, computer, etc., while the receiver is the learner. Because our instruction-learning system is an information processing system, we are now faced with several questions:

1. What is information?
2. How can we measure information?
3. What are the characteristics of an efficient coding process?
4. How could we measure the capacity of a channel?



1 - Information Source
2 - Transmitter
3 - Channel
4 - Receiver
5 - Destination
6 - Noise

Figure 1.--Simplified View of Communication System.

5. What are the general characteristics of noise?
6. What are the effects of noise on a message?
7. How do we increase or reduce the effects of noise?

All of the questions are intriguing and deserve attention. However, this investigation addresses itself only to the first two questions. The other questions provide a basis for additional studies along with extending the results of this investigation.

Many definitions of information are given in the literature (Ash, 1965; Edwards, 1964; Frank, 1962; Fuchs, 1968; Guilbaud, 1959; Khinchin, 1957; Singh, 1966; Weltner, 1970; & others). But one appropriate for this investigation is the definition given by Weaver (1969). He says, "Information is a measure of one's freedom of choice when one selects a message." This conveniently ties the definition of information and its measurement directly together.

To take the simplest case of two alternative messages (yes or no), we will say that the information associated with this ensemble of two messages is unity. The concept of information, then, applies not to the individual messages but to the ensemble as a whole.

To be more precise, we can define the amount of information by the logarithm of the number of available choices (messages) contained in the ensemble. Thus, in our case of two messages (yes or no), the information is proportional to the logarithm of 2 to the base 2 such that: $\log_2 2 = 1$ which is unity. This unit of information is called a "bit" from the two words binary digit. To generalize then, if we have 8 alternative messages, among which we are equally free to choose, we have $\log_2 8 = 3$ giving us an ensemble consisting of 3 bits of information.

This situation is completely adequate if we are only dealing with messages that have an equally probable chance of being selected. The probability of this occurring in language structure is extremely small and is zero for the English language. In this study we are concerned with the information content of the English language which is composed of messages (letters, signs) arranged into stochastic ensembles. These ensembles are arranged according to certain probabilities in which the probabilities depend on previous events.

Because of the probabilistic dependencies of the message (letters), it becomes exceedingly more difficult to measure the amount of information in the English language. However, Shannon (1951) derived a "guessing procedure" for analyzing the syntactical information content of meaningful texts. His guessing procedure enables one to calculate the subjective information of a text for a specific learner. The learner attempts to guess the ensemble, letter by letter. After each guess he is told whether his guess was correct or not. If incorrect, he guesses again. If correct, he guesses the next letter or message of the ensemble until the entire ensemble has been derived. This procedure is based on the assumption that at each guess the learner will name the letter with the highest subjective probability. Thus, we have a measure of the amount of information a particular ensemble has for a particular learner. Shannon's equation forms the basis for this investigation and is presented later. An example of the guessing procedure and subsequent analysis follows.

Example of Guessing Procedure

The guessing procedure enables one to empirically estimate the subjective information of a text for a particular subject. The subjective information can be calculated according to several techniques. However, each technique assumes that the subject guesses the sign with the highest subjective probability. Consider the following hypothetical example:

THE COW JUMPED O.....

612 1 421 1 811111 1 3.....

Under each letter of the text is the number of guesses the subject made until the correct sign (letter) was guessed. From these numbers it is relatively easy to calculate the information content of the text by adding the information value of the signs. The first letter (T) contains $\log_2 6$ bits of information, the second letter (H) $\log_2 1$ bits of information. Therefore, the first letter (T) contains 2.58 bits ($\log_2 6 = 2.58$) whereas the second letter (H) contains 0 bits ($\log_2 1 = 0$). While the first letter did contain information for the subject (2.58 bits), the second was perhaps expected with certainty and was, therefore, devoid of any information (0 bits). The <u>total</u> information of the text can now be measured in bits as:

$$
\begin{aligned}
H(\text{text}) &= \log_2 6 + \log_2 1 + \log_2 2 + \log_2 1 + \log_2 4 + \log_2 2 \\
&\quad + \log_2 1 + \log_2 1 + \log_2 8 + \log_2 1 + \log_2 1 + \log_2 1 \\
&\quad + \log_2 1 + \log_2 1 + \log_2 1 + \log_2 3 \\
&= \log_2 8 + \log_2 6 + \log_2 4 + \log_2 3 + 2 \log_2 2 + 10 \log_2 1 \\
&= 3.00 + 2.58 + 2.00 + 1.58 + 2.00 + 0 \\
&= 11.16 \text{ BITS}
\end{aligned}
$$

Dividing this by the length of the text N = 16 we arrive at .698 Bits/Sign for the information content of our sample text.

Although this method is relatively simple, it does not provide the best estimate of the information content. Shannon (1951) derived equation (1) which according to Frank (1962) is not only a better measure but also is closer to the actual figure than other calculations would indicate.

$$H(\text{text}) = \Sigma N_r \cdot (r \cdot ld(r) - (r-1) \cdot ld(r-1)) \quad (1)$$

Where: N = amount of information measured in bits

N_r = absolute frequency with which the number r occurs in the guessing sequence of N signs

ld = logarithm to the base 2 ($\log_2 r$)

Applying Shannon's equation to our example, we obtain 18.248 bits as the information of the text segment or 1.405 bits/sign which is a better estimate than our previous figure of .698 bits/sign.

Procedures and Results

This study consists of five procedural phases of operation. They are:

Phase I: The design and development of computer programs to present and process the experimental data.

Phase II: The selection and presentation of textual information to selected subjects and the collection of their responses.

Phase III: Calculation of subjective information content contained in the textual material for selected subjects.

Phase IV: Derivation of equations to fit experimental data.

Phase V: Selection of "best-fitting" equation, and deletion of outlier point.

PHASE I

Computer Program Development

Two computer programs were written specifically for this investigation. Two other programs were used in the synthesis of the data. The programs, written in BASIC, were designed to operate in a time-shared environment.

Program CYBER4

Program CYBER4 was designed as a general purpose program to present selected textual passages to the subjects. CYBER4 can present textual information by letters, words, phrases, sentences, paragraphs, or any combination thereof. The content presentation can be predetermined by the investigator or the program can be adapted to function under learner control. In addition, it can also be adapted to display information at any intellectual level.

The program initially outputs an extended phrase or set of phrases. The length of the phrase is pre-programmed by the investigator and, for the most part, depends on the complexity of the material and the relative sophistication of the subject with regard to the academic content.

After the initial phrase is printed by the computer terminal, the subject is required to guess, letter by letter, the subsequent message. If the subject is incorrect, his response is placed in a "response file;" a zero is recorded in a "score file;" and the subject is required to guess again. This process continues until his response is correct.

If the subject's guess is correct, his response is placed in the "response file;" a one is recorded in the "score file;" his correct answer is acknowledged by the computer; and the next letter in the guessing sequence is presented. When the subject has responded to the entire text, he is thanked for his effort and the time of completion is recorded by the computer.

Program CYBER2

Program CYBER2 was developed to calculate the amount of information for each subject, using equation (1) derived by Shannon (1951).

As can be seen in the sample output listed in Table 1, CYBER2 outputs the percentage of incorrect responses (X_ss) per subject and the amount of information per subject (Y_ss) contained in the text for that particular subject. The measurement of information is in units of bits/sign, the sign being the message-unit letter. Each pair of numbers represent the X and Y coordinates used in fitting an equation to the data. In addition to calculating the X and Y values, CYBER2 also orders the results from the lowest to highest on the percentage of incorrect responses (X_ss

TABLE 1

X AND Y COORDINATES FOR HYPOTHETICAL SAMPLE

Percent Incorrect (X_ss)	Bits/Sign (Y_ss)
X(1) = 22.2222	Y(1) = .9148
X(2) = 25.0000	Y(2) = .5629
X(3) = 26.0000	Y(3) = 1.3686
X(4) = 60.0000	Y(4) = 1.8500

Program EQUFIT

Program EQUFIT is a commercially developed program available on the HG-255 time-sharing computer system at Arizona State University. EQUFIT applies curve fitting techniques to determine which of six general types of curves best fits the supplied data. The six general types of curves range from a linear function to a hyperbolic function.

The X and Y coordinates generated by Program CYBER2 are used by EQUFIT to determine the equations that have the "best fit" for the supplied data. EQUFIT outputs the following information for each equation specified by the user:

1. Index of determination for each equation

2. Coefficients for each equation

3. X and Y values as supplied by user

4. Calculated values of Y for each specified equation

5. The percentage difference between the actual Y value and the calculated Y value for each specified equation.

Program PTAFIT

PTAFIT is a modified and improved version of EQUFIT developed by Plan-Test Associates of Phoenix, Arizona.* PTAFIT greatly extends the capabilities of EQUFIT. In addition to the information supplied by EQUFIT, PTAFIT supplies the following:

1. The level of significance at which the equation fits the supplied data.

2. The value of each coefficient, the standard deviation associated with that coefficient and confidence limits for the coefficient. The confidence limits are chosen by the user and can be at the 90%, 95%, and/or the 99% level of significance.

3. Confidence bands about each (X, Y) coordinate. The user may specify the 90%, 95% and/or 99% prediction limits.

4. Evaluation of how significant is the apparent differences between equations.

PHASE II

Selection and Presentation of Textual Information

The textual information for this investigation was selected from two areas. The first area was concerned with academic facts in the areas of information science. All participants had either prior experience in the academic field or were enrolled in an information science course. The second area was concerned with current events. The most topical of current events, at the time of the investigation, was the Watergate affair. It was being presented daily on television and in the news media. It was thought that all subjects would be knowledgeable in this area; however, this was not the case. A greater percentage of incorrect responses was made on the Watergate textual material than on the information science textual material.

Experimental Subjects

A total of 239 subjects were available from the population of graduate students in the College of Education. An analysis of the

*Plan-Test Associates, Financial Center, Suite 609, 3443 North Central Avenue, Phoenix, Arizona 85012.

subjects' academic background indicated that all participants were functioning within the normal range of intelligence. All subjects used in this investigation were enrolled in College of Education graduate classes in an uncontrolled manner typical of college enrollment. It is assumed that all subjects used in this investigation are representative, in terms of language skills, of the population of college graduates whose primary language is English.

Out of the 239 subjects available to participate in this investigation 118 were selected. These 118 subjects were divided into three groups based upon their academic background.

Group 1: 25 subjects
A test group to validate the computer programs and selection of the textual information. The data derived from these subjects was not used in the final determination of the equation.

Group 2: 59 subjects
An experimental group of subjects who were exposed to the two types of textual materials. Group 2 subjects were of varied academic background and therefore considered to be less reliable in terms of their usefulness as a homogeneous group. The textual material presented to this group consisted of both academic materials (Information Science) and current material (Watergate affair). The data generated by Group 2 was suspect because of numerous computer failures throughout the experimental period and was therefore rejected.

Group 3: 34 subjects
An experimental group of subjects who were equivalent in terms of academic background in the field of information science. All of the Group 3 subjects were concurrently enrolled in an information science course at the time of the study. The textual material presented to this group was directly concerned with previously learned material and material they were about to study in their class. This group provided the most reliable data for the final determination of the derived equation.

BEST COPY AVAILABLE

PHASE III

Calculation of Subjective Information

After all subjects had completed the guessing sequences, Program CYBER2 processed the "score file," calculated the percentage of incorrect responses and computed the subjective, textual information for each subject according to equation (1):

$$H(text) = \Sigma N_r \cdot (r \cdot ld(r) - (r-1) \cdot ld(r-1)) \qquad (1)$$

Table 2 presents the results derived from test Group 3 as calculated by Program CYBER2. This table shows the percentage of incorrect responses and the corresponding information content for each of the subjects. The table indicates that, in general, as the percentage of incorrect responses increases the amount of information increases. This tends to confirm the theoretical predictions.

PHASE IV

Determination of Equations

After all subjects had completed the guessing sequences, program CYBER2 processed the "Score File," calculated the percentage of incorrect responses and computed the subjective textual information for each subject according to equation (1). This provided the necessary information for EQUFIT to process.

Program EQUFIT provides a least squares curve fit for the data generated in Phase III by program CYBER2. The six curve types and results are displayed in Table 3. The index of determination indicates that equations 6, 3 and 1, respectively, provide a fit that is "quite good."

Examination of Table 3 would tend to indicate that equation (6) is the best fit. Although this equation shows the highest Index of Determination, the other two equations are so close that it is difficult to judge which is the best fitting equation using only the Index of Determination.

One would prefer the linear function because of its ease of usage, but only if it proved to be "as good a fit" as the other functions. The equation derived by Weltner (1967) for the German language

Table 2

X AND Y COORDINATES FOR GROUP 3

Percent Incorrect	Bits/Sign
X(1) = 7.6923	Y(1) = .2281
X(2) = 10.0000	Y(2) = .2664
X(3) = 10.0000	Y(3) = .2897
X(4) = 12.5000	Y(4) = .3255
X(5) = 15.0000	Y(5) = .4942
X(6) = 15.0000	Y(6) = .5328
X(7) = 15.3846	Y(7) = .4238
X(8) = 15.3846	Y(8) = .3077
X(9) = 17.5000	Y(9) = .4821
X(10) = 17.5000	Y(10) = .4724
X(11) = 20.0000	Y(11) = .5407
X(12) = 20.0000	Y(12) = .6020
X(13) = 20.0000	Y(13) = .5264
X(14) = 20.0000	Y(14) = .5286
X(15) = 20.0000	Y(15) = .5627
X(16) = 20.0000	Y(16) = .5698
X(17) = 21.4286	Y(17) = .5364
X(18) = 22.5000	Y(18) = .6406
X(19) = 27.2727	Y(19) = .6918
X(20) = 27.5000	Y(20) = .9640
X(21) = 27.5000	Y(21) = 1.0123
X(22) = 27.5000	Y(22) = .9593
X(23) = 27.5000	Y(23) = .8112
X(24) = 32.5000	Y(24) = .9116
X(25) = 35.0000	Y(25) = 1.2194
X(26) = 35.0000	Y(26) = 1.2905
X(27) = 35.0000	Y(27) = 1.2210
X(28) = 37.5000	Y(28) = .7500
X(29) = 37.5000	Y(29) = 1.2191
X(30) = 38.4615	Y(30) = 1.1554
X(31) = 40.0000	Y(31) = 1.2366
X(32) = 42.5000	Y(32) = 1.3555
X(33) = 45.4545	Y(33) = 1.2501
X(34) = 56.2500	Y(34) = 1.5010

BEST COPY AVAILABLE

TABLE 3

LEAST SQUARES CURVE FIT

EQUFIT PROGRAM

Curve Type	Index of Determination	A	B
1. Y=A+(B*X)	.892389	-1.49193E-02	3.02491E-02
2. Y=A*EXP(B*X)	.851611	.229436	4.19008E-02
3. Y=A*(X+B)	.922853	2.41865E-02	1.05825
4. Y=A+(B/X)	.693833	1.35166	-12.2042
5. Y=1/(A+B*X)	.711047	3.48035	- 6.92674E-02
6. Y=X/(A+B*X)	.930395	36.2543	- 5.09608E-02

is of a linear nature (see Figure 2). This would lead us to suspect that the English language equation might also exhibit similar linear properties. Therefore, it was decided, at this point, to further analyze the Group 3 data. This is done in Phase V.

PHASE V

Selection of "Best-Fitting" Equation

Phase V, in part, utilized the program PTAFIT to further analyze the equations to determine which would be the most appropriate to describe the empirical data.

Table 4 shows the least-curves fit generated by program PTAFIT. All equations are shown to be significant at the 99% level of confidence.

TABLE 4

PTAFIT LEAST SQUARES CURVE FIT

Curve Type	Index of Determination	A	B	Significance 90%	95%	99%
1. Y=A+(B*X)	.8924	-1.494E-2	3.025E-2	*	*	*
2 Y=A*EXP(B*X)	.8517	.2294	.0419	*	*	*
3 Y=A*(X↑B)	.9229	2.418E-2	1.058	*	*	*
4 Y=A+(B/X)	.6938	1.353	-12.2	*	*	*
5 Y=1/(A+B*X)	.7111	3.48	- 6.927E-2	*	*	*
6 Y=X/(A+B*X)	.9304	36.26	- 5.097E-2	*	*	*

Number of Observations: 34

Equations 6, 3 and 1 again exhibit the highest Index of Determination and therefore are prime candidates for further examination. Tables 5 and 6 show the standard deviation and confidence limits for the coefficients of equations 6 and 1.

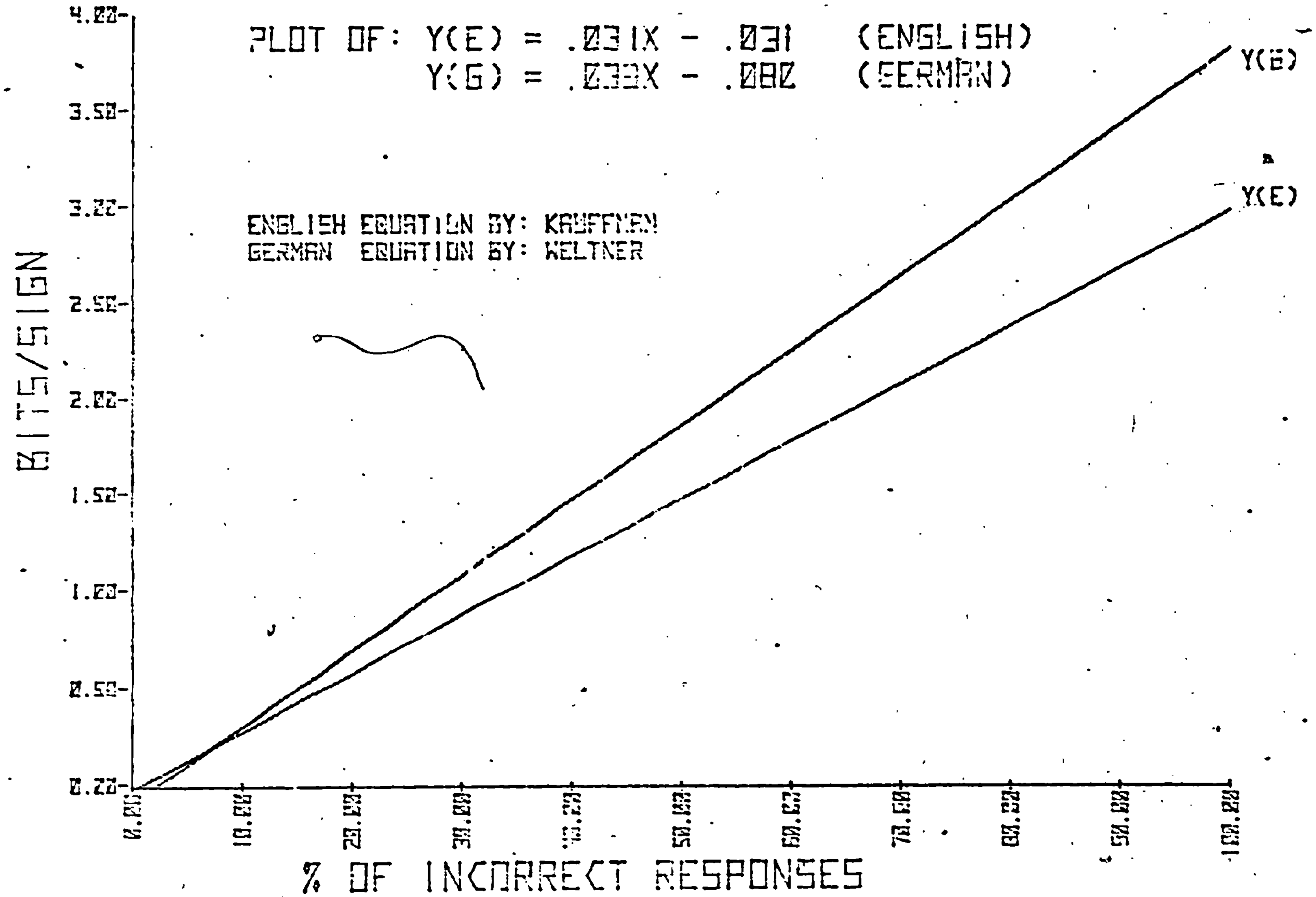


Figure 2.--Plot of English and German equations.

TABLE 5

EQUATION TYPE 6

Y=X/(A + B*X)

Coefficient	Value	Standard Deviation	95% Confidence Limits	
A	36.26	1.75	32.68	39.83
B	- 5.09E-2	9.52E-2	- .24	.14

TABLE 6

EQUATION TYPE 1

Y=A+(B*X)

Coefficient	Value	Standard Deviation	95% Confidence Limits	
A	-1.49E-2	5.20E-2	- .12	9.11E-2
B	3.02E-2	1.85E-3	2.64E-2	3.40E-2

Table 7 shows the PTAFIT output comparing the fit of all pairs of equations. It shows that equation 6 does not fit significantly better than equation 1. Therefore, since equation 1 represents a straight line it will be given special consideration in view of its convenience of interpretation.

TABLE 7

COMPARISON OF INDICES OF DETERMINATION

	95% CONFIDENCE						99% CONFIDENCE					
	CURVE NO.						CURVE NO.					
	1	2	3	4	5	6	1	2	3	4	5	6
6	*	*	*			*	*	*	*			*
3	*	*	*			*	*	*	*			*
1	*	*	*			*	*	*	*	*	*	*
2	*	*	*	*	*	*	*	*	*	*	*	*
5		*		*	*		*	*		*	*	
4		*		*	*		*	*		*	*	

* ASTERISKS INDICATE PAIRS OF CURVES ARE THE 'SAME'. (NOT SHOWN TO HAVE SIGNIFICANTLY DIFFERENT INDICES OF DETERMINATION AT THE STATED CONFIDENCE LEVELS).

Tables 8 and 9 show further output from the PTAFIT program. The point with the coordinates of 37.5 and 0.75 appeared suspiciously deviant and an outlier test run on the residuals confirms this suspicion. The outlier test was run using a Plan-Test Associates program based on a method reported by Chazal (1967). The graphical and tabular output from this program are shown in Tables 10 and 11, respectively.

In view of this, PTAFIT was rerun omitting the outlier. The resulting outputs are shown in Tables 11 through 17. We now find that equation 3 fits better than equation 6 (see Table 12), but still neither is significantly better than equation 1 (see Table 15).

TABLE 8

EQUATION TYPE 1

BEST COPY AVAILABLE

Y=A+(B*X)

X	Y	Y-CALC	% DIF	95 % PREDICTION LIMITS	
% INC	H				
7.692	.2231	.2177	4.768	-4.476E-2	.4802
10	.2664	.2875	-7.352	2.718E-2	.5479
10	.2897	.2875	.7509	2.718E-2	.5479
12.5	.3255	.3632	-10.37	.1048	.6215
15	.4942	.4388	12.63	.1821	.6955
15	.5328	.4388	21.42	.1821	.6955
15.35	.4238	.4494	-5.694	.1929	.7059
15.38	.3077	.4503	-31.67	.1938	.7068
17.5	.4821	.5144	-6.284	.259	.7698
17.5	.4724	.5144	-8.17	.259	.7698
20	.5627	.5901	-4.636	.3356	.8445
20	.5698	.5901	-3.433	.3356	.8445
20	.5407	.5901	-8.365	.3356	.8445
20	.602	.5901	2.024	.3356	.8445
20	.5264	.5901	-10.79	.3356	.8445
20	.5286	.5901	-10.42	.3356	.8445
21.43	.5364	.6333	-15.3	.3793	.8874
22.5	.6406	.6657	-3.769	.4119	.9195
27.27	.6918	.81	-14.59	.5564	1.064
27.5	.9593	.8169	17.43	.5633	1.071
27.5	.8112	.8169	-.7032	.5633	1.071
27.5	.964	.8169	18.	.5633	1.071
27.5	1.012	.8169	23.88	.5633	1.071
32.5	.9116	.9682	-5.846	.7134	1.223
35	1.291	1.044	23.68	.7878	1.3
35	1.221	1.044	16.97	.7878	1.3
35	1.219	1.044	16.78	.7878	1.3
37.5	.75	1.119	-33.	.862	1.377
37.5	1.219	1.119	8.892	.862	1.377
38.46	1.155	1.149	.5657	.8904	1.407
40	1.237	1.195	3.507	.9358	1.454
42.5	1.356	1.271	6.711	1.009	1.532
45.45	[illegible]	1.36	-8.086	1.096	1.624
56.25	1.501	1.687	-11.01	1.408	1.965

TABLE 9

EQUATION TYPE 6

$Y=X/(A+B*X)$

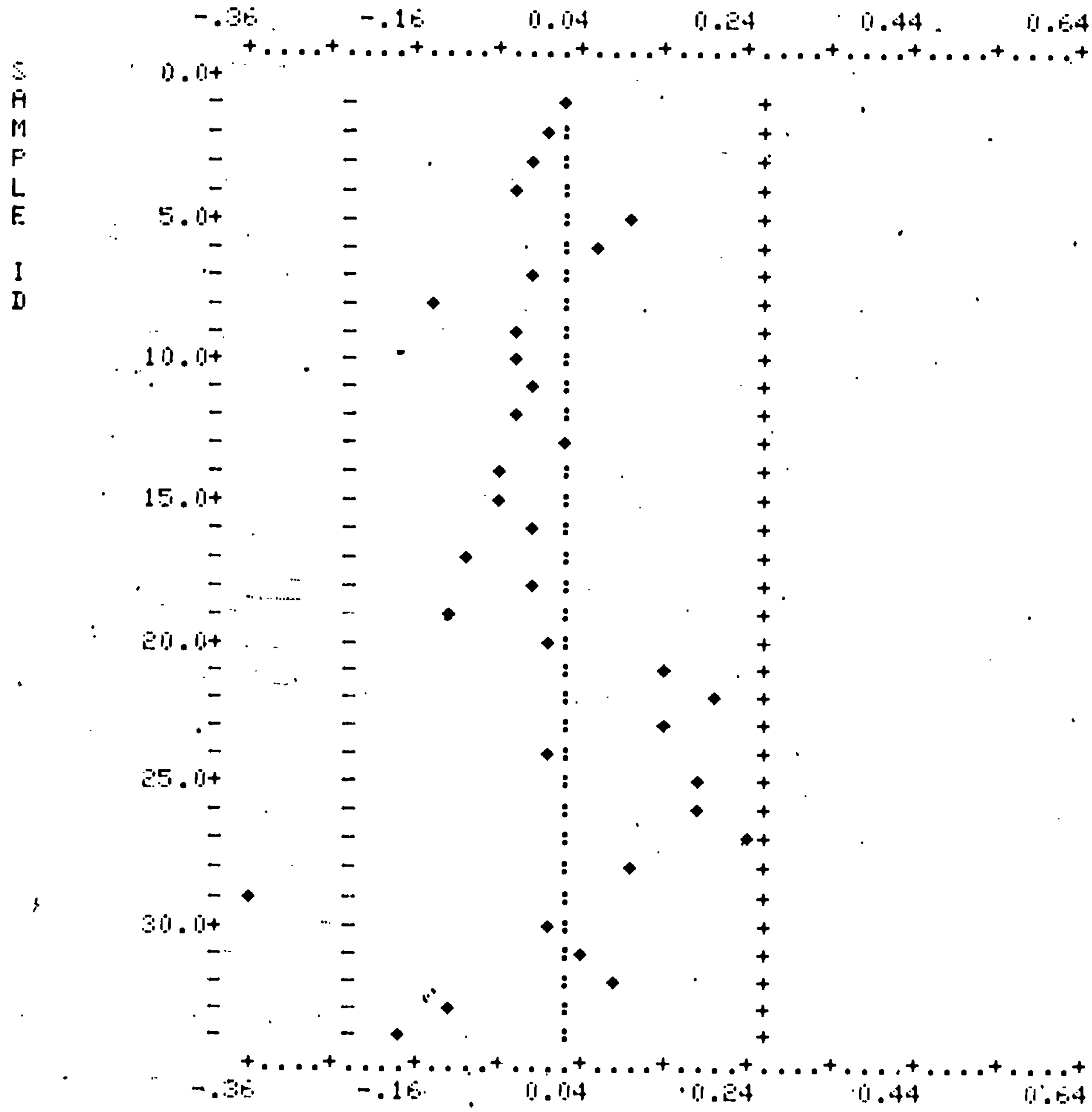
X	Y	Y-CALC	% DIF	95 % PREDICTION LIMITS	
% INC	H				
7.692	.2281	.2145	6.347	.1901	.2461
10	.2664	.2798	-4.776	.2422	.3311
10	.2897	.2798	3.553	.2422	.3311
12.5	.3255	.351	-7.252	.2955	.4321
15	.4942	.4227	16.93	.3457	.5438
15	.5328	.4227	26.06	.3457	.5438
15.35	.4238	.4327	-2.065	.3525	.5604
15.38	.3077	.4336	-29.04	.353	.5618
17.5	.4821	.4949	-2.581	.393	.6681
17.5	.4724	.4949	-4.542	.393	.6681
20	.5627	.5676	-.8663	.4377	.8072
20	.5698	.5676	.3846	.4377	.8072
20	.5407	.5676	-4.742	.4377	.8072
20	.602	.5676	6.057	.4377	.8072
20	.5264	.5676	-7.261	.4377	.8072
20	.5286	.5676	-6.874	.4377	.8072
21.43	.5364	.6095	-11.99	.4622	.8946
22.5	.6406	.6409	-4.483E-2	.48	.9641
27.27	.6918	.7822	-11.55	.5546	1.326
27.5	.9593	.789	21.58	.559	1.346
27.5	.8112	.789	2.809	.558	1.946
27.5	.964	.789	22.17	.558	1.346
27.5	1.012	.789	28.26	.558	1.346
32.5	.9116	.9394	-2.956	.6284	1.859
35	1.291	1.015	27.15	.6611	2.187
35	1.221	1.015	20.25	.6611	2.187
35	1.219	1.015	20.06	.6611	2.187
37.5	.75	1.092	-31.31	.6923	2.583
37.5	1.219	1.092	11.64	.6923	2.593
38.46	1.155	1.121	2.989	.7039	2.757
40	1.237	1.169	5.811	.7221	3.069
42.5	1.356	1.247	8.761	.7505	3.691
45.45	1.25	1.339	-6.662	.7825	4.643
56.25	1.501	1.685	-10.91	.8867	16.85

TABLE 10

OUTLIER TEST USING CHAZAL'S METHOD

SYMBOL DEFINITION:

◆ = MEAN OF SAMPLE DATA
: = GRAND MEAN OF DATA (EXCLUDING OUTLIERS)
- = LOWER OUTLIER LIMIT
+ = UPPER OUTLIER LIMIT



INCREMENT OF HORIZONTAL SCALE: 2.00E-02
INCREMENT OF VERTICAL SCALE: 1.00

NO. OF POINTS PLOTTED ON VERTICAL SCALE: 34

BEST COPY AVAILABLE

TABLE 11

OUTLIER TEST USING CHAZAL''S METHOD

MEAN	OUTLIER	LIMITS
0.010	-.237	0.262
0.002	-.237	0.262
-.021	-.237	0.262
-.038	-.237	0.262
0.094	-.237	0.262
0.055	-.237	0.262
-.026	-.237	0.262
-.143	-.237	0.262
-.042	-.237	0.262
-.032	-.237	0.262
-.020	-.237	0.262
-.049	-.237	0.262
0.012	-.237	0.262
-.064	-.237	0.262
-.061	-.237	0.262
-.027	-.237	0.262
-.097	-.237	0.262
-.025	-.237	0.262
-.118	-.237	0.262
-.006	-.237	0.262
0.147	-.237	0.262
0.195	-.237	0.262
0.142	-.237	0.262
-.009	-.237	0.262
0.177	-.237	0.262
0.175	-.237	0.262
0.247	-.237	0.262
0.100	-.237	0.262
◆ -.369	-.237	0.262
0.006	-.237	0.262
0.042	-.237	0.262
0.085	-.237	0.262
-.110	-.237	0.262
-.186	-.237	0.262

GRAND MEAN (EXCLUDING OUTLIERS) = 0.013
BETWEEN SD (EXCLUDING OUTLIERS) = 0.10260
WITHIN SD (EXCLUDING OUTLIERS) = 0.00000

◆ = EXCLUDED ON PASS NO. 1

BEST COPY AVAILABLE

TABLE 12

LEAST SQUARES CURVE FIT WITH OUTLIER REMOVED

CURVE TYPE	INDEX OF DETERMINATION	A	B	SIGNIFICANCE 90%	95%	99%
1 Y=A+(B*X)	.9249	-3.081E-2	3.132E-2	*	*	*
2 Y=A*EXP(B*X)	.8692	.2256	4.302E-2	*	*	*
3 Y=A*(X↑B)	.9418	2.244E-2	1.086	*	*	*
4 Y=A+(B/X)	.7118	1.375	-12.51	*	*	*
5 Y=1/(A+B*X)	.7171	3.5	-7.057E-2	*	*	*
6 Y=X/(A+B*X)	.9364	36.71	-8.614E-2	*	*	*

NUMBER OF OBSERVATIONS: 33

BEST COPY AVAILABLE

TABLE 13

EQUATION TYPE 3

Y=A*(X↑B)

LEAST SQUARES CURVE FIT WITH OUTLIER REMOVED

COEFFICIENT	VALUE	STANDARD DEVIATION	95 % CONFIDENCE LIMITS	
A:	2.244E-2	1.166	1.642E-2	3.067E-2
B:	1.086	4.847E-2	.9872	1.185

BEST COPY AVAILABLE

TABLE 14

EQUATION TYPE 1

$Y=A+(B*X)$

LEAST SQUARES CURVE FIT WITH OUTLIER REMOVED

COEFFICIENT	VALUE	STANDARD DEVIATION	95 % CONFIDENCE LIMITS	
A:	-3.081E-2	4.442E-2	-.1214	5.978E-2
B:	3.132E-2	1.603E-3	2.805E-2	3.459E-2

ESTIMATED POPULATION STANDARD DEVIATION
ABOUT REGRESSION LINE = .1041

BEST COPY AVAILABLE

TABLE 15

COMPARISON OF INDICES OF DETERMINATION

WITH OUTLIER REMOVED

	95% CONFIDENCE							99% CONFIDENCE					
	CURVE NO.							CURVE NO.					
	1	2	3	4	5	6		1	2	3	4	5	6
3	◆	◆	◆			◆	3	◆	◆	◆			◆
6	◆	◆	◆			◆	6	◆	◆	◆			◆
1	◆	◆	◆			◆	1	◆	◆	◆			◆
2	◆	◆	◆	◆	◆	◆	2	◆	◆	◆	◆	◆	◆
5		◆		◆	◆		5		◆		◆	◆	
4		◆		◆	◆		4		◆		◆	◆	

◆ ASTERISKS INDICATE PAIRS OF CURVES ARE THE 'SAME'. (NOT SHOWN TO HAVE SIGNIFICANTLY DIFFERENT INDICES OF DETERMINATION AT THE STATED CONFIDENCE LEVELS).

BEST COPY AVAILABLE

TABLE 16

EQUATION TYPE 1

Y=A+(B*X)

SUMMARY TABLE WITH OUTLIER REMOVED

X	Y	Y-CALC	% DIF	95 % PREDICTION LIMITS	
% INC	H				
7.692	.2281	.2101	8.564	-1.302E-2	.4332
10	.2664	.2824	-5.663	.0611	.5037
10	.2897	.2824	2.587	.0611	.5037
12.5	.3255	.3607	-9.757	.1411	.5803
15	.4942	.439	12.58	.2208	.6572
15	.5328	.439	21.37	.2208	.6572
15.35	.4238	.45	-5.813	.2319	.668
15.38	.3077	.4509	-31.76	.2329	.6689
17.5	.4821	.5173	-6.803	.3002	.7344
17.5	.4724	.5173	-8.679	.3002	.7344
20	.5627	.5956	-5.523	.3793	.8119
20	.5698	.5956	-4.331	.3793	.8119
20	.5407	.5956	-9.217	.3793	.8119
20	.602	.5956	1.076	.3793	.8119
20	.5264	.5956	-11.62	.3793	.8119
20	.5286	.5956	-11.25	.3793	.8119
21.42	.5364	.6404	-16.24	.4244	.8563
22.5	.6406	.6739	-4.941	.4581	.8897
27.27	.6918	.8233	-15.97	.6076	1.039
27.5	.9593	.8305	15.51	.6148	1.046
27.5	.8112	.8305	-2.323	.6148	1.046
27.5	.964	.8305	16.08	.6148	1.046
27.5	1.012	.8305	21.86	.6148	1.046
32.5	.9116	.9871	-7.648	.7702	1.204
35	1.291	1.065	21.18	.8475	1.283
35	1.221	1.065	14.61	.8475	1.283
35	1.219	1.065	14.42	.8475	1.283
37.5	1.219	1.144	6.584	.9245	1.363
38.46	1.155	1.174	-1.599	.9539	1.394
40	1.237	1.222	1.228	1.001	1.443
42.5	1.356	1.3	4.284	1.078	1.523
45.45	1.25	1.393	-10.25	1.167	1.618
56.25	1.501	1.731	-13.28	1.493	1.969

BEST COPY AVAILABLE

TABLE 17

EQUATION TYPE 3

Y=A*(X↑B)

SUMMARY TABLE WITH OUTLIER REMOVED

X	Y	Y-CALC	% DIF	95 % PREDICTION LIMITS	
% INC	H				
7.692	.2281	.2058	10.86	.154	.2748
10	.2664	.2736	-2.635	.2066	.3623
10	.2897	.2736	5.88	.2066	.3623
12.5	.3255	.3486	-6.639	.2648	.4591
15	.4942	.425	16.28	.3238	.5578
15	.5328	.425	25.37	.3238	.5578
15.35	.4238	.4358	-2.748	.3321	.5718
15.38	.3077	.4367	-29.54	.3328	.573
17.5	.4821	.5025	-4.051	.3835	.6582
17.5	.4724	.5025	-5.981	.3835	.6582
20	.5627	.5809	-3.128	.4438	.7602
20	.5698	.5809	-1.906	.4438	.7602
20	.5407	.5809	-6.915	.4438	.7602
20	.602	.5809	3.638	.4438	.7602
20	.5264	.5809	-9.377	.4438	.7602
20	.5286	.5809	-8.999	.4438	.7602
21.43	.5364	.6261	-14.33	.4785	.8192
22.5	.6406	.6601	-2.96	.5046	.8637
27.27	.6918	.8134	-14.95	.6214	1.065
27.5	.9593	.8209	16.86	.627	1.075
27.5	.8112	.8209	-1.181	.627	1.075
27.5	.964	.8209	17.43	.627	1.075
27.5	1.012	.8209	23.28	.627	1.075
32.5	.9116	.9842	-7.377	.7505	1.291
35	1.291	1.067	21.03	.8126	1.4
35	1.221	1.067	14.47	.8126	1.4
35	1.219	1.067	14.28	.8126	1.4
37.5	1.219	1.15	6.028	.8748	1.511
38.46	1.155	1.182	-2.259	.8988	1.554
40	1.237	1.233	.3102	.9372	1.623
42.5	1.356	1.317	2.953	.9997	1.735
45.45	1.25	1.417	-11.77	1.074	1.869
56.25	1.501	1.786	-15.95	1.345	2.37

BEST COPY AVAILABLE

The English Equation

Our straight line equation, rounded to three decimal places, from Table 14 becomes

$$H(\text{text})_E = .031X - .031 \qquad (2)$$

Where: $H(\text{text})_E$ = subjective textual information per text--English

And: X = percentage of incorrect responses

It is interesting to note, however, that the -.031 intercept could well be zero for the population since its 95% confidence band spans that value (see Table 14). This zero value is to be expected since the theoretical information content should be zero for zero incorrect responses. In other words, if a person does not make a response error he is predicting the text with certainty and, therefore, the text contains zero information for him. The English equation closely approximates this. Also the slope could (at the 95% level of confidence) span from 0.028 to .0345.

The German Equation

The equation derived by Weltner for the German language is:

$$H(\text{text})_G = .039X - .080 \qquad (3)$$

Where: $H(\text{text})_G$ = subjective textual information per text--German

And: X = percentage of incorrect responses

We have no measure of the confidence bands on the German data, but the agreement between the equations is quite remarkable. See the graphs in Figure 2 (page 21.

One additional point in regard to the residuals is the extreme runs observed which might suggest a somewhat better fit with higher degree polynomials. This has not been pursued in this study.

It is interesting to note that the equation can be put into a more useful form. The derivation is as follows:

$$H(\text{text})_E = .031X - .031 \qquad (2)$$

Since X = percentage of incorrect responses

which can be represented by $E/N \cdot 100$

Where E = number of wrongly guessed signs

And N = text length

Then

$$H(\text{text}) = .031[E/N \cdot 100] - .031$$

$$H(\text{text}) = 3.1E/N - .031$$

$$N \cdot H(\text{text}) = 3.1E - .031N$$

Setting $N \cdot H(\text{text}) = I$, where I = information in bits we have

$$I = 3.1E - .031N \qquad (4)$$

With the equation in this form we can now directly evaluate the information content of a text of length N for a specific subject. In view of our earlier evaluation of the confidence on the .031 coefficient which may well be zero, it may be concluded that our equation is independent of the text length N such that equation (4) at the 95% confidence level becomes

$$I = 3.1E \qquad (5)$$

Conclusion

The application of Shannon's guessing procedure permits us to measure the subjective information of a given text for a specific learner. Unlike senseless texts, the subjective information of a meaningful text can vary from learner to learner. The derivation of equation (2), equation (4) and equation (5) in this study now allows a direct evaluation of the "meaningfulness" of a specific English text for a specific learner.

Educational theory tells us that the interactive events between a learner and his environment have a direct influence upon his learning. Since these interactive events can now be specified in terms of information theory, the specification of information-theory-based criteria for the selection and completion of education goals is possible.

These equations now permit us to measure information in terms of a value that is dependent upon the internal state of the learner. We have always known that the internal state of the learner is directly related to his ability to process information in a meaningful manner (learning). We have also known that subject matter has an inherent difficulty factor. The degree of this difficulty factor plus the internal state of the learner are direct variables that influence the effectiveness of an instructional system. The equations derived, in this study, now provide a quantitative measure of these factors. This, in turn, permits us to specify attainable goals for both the instructional system and the learner. In practice, this can provide us with the means to specify precisely the instructions needed by the learners.

To the educator this means that he cannot only derive a quantitative measure of information contained in an instruction sequence for each learner, but he can use this measure to evaluate the effectiveness of his instruction.

REFERENCES

Ash, R. Information theory. New York: Interscience, 1965.

Chazal, M. How to discard invalid measurements. Chemical Engineering, February, 1967. Pp. 182-184.

Edwards, E. Information transmission. London: Chapman & Hall, 1964.

Frank, H. Kybernetische grundlagen der padagogik. Baden-Baden: Agis Verlag, 1962.

Fuchs, W. Cybernetics for the modern mind. New York: Macmillan, 1971.

Guilbaud, G. What is cybernetics? London: William Heineman, 1959.

Khinchin, A. I. Mathematical foundations of information theory. New York: Dover, 1957.

Minsky, M. Semantic information processing. Cambridge: MIT Press, 1968.

Singh, J. Great ideas in information theory, language and cybernetics. New York: Dover, 1966.

Shannon, C. The mathematical theory of communications. Bell System Technical Journal, 1948, 27, 279-425/623-656.

Shannon, C. Prediction and entropy of printed English. Bell System Technical Journal, January, 1951.

Shannon, C., & Weaver, W. The mathematical theory of communication. Chicago: University of Illinois Press, 1969.

Weltner, K. Zum ratetest nach Shannon. In H. Schmidt (Ed.), Grundlagenstudien aus kybernetik und geisteswissenschaft. Band 6. Quickborn Bei Hamburg: Schnelle Verlag, 1965.

Weltner, K. Der Shannonsche ratetest in der praxis der programmierten instruktion. Lehrmaschinen in kybernetischer und padagogischer sicht 4. Stuttgart: Ernst Klett Verlag, 1966. Pp. 40-54.

Weltner, K. Zur bestimmung der subjektiven information durch ratetests. Praxis und perspektiven des programmierten unterricts. Band II. Bevensee, Quickborn, Germany: Verlag Schnelle, 1967. Pp. 69-74.

Weltner, K. Informationstheorie und erziehungswissenschaft. Quickborn: Verlag Schnelle, 1970.