Journalism Quarterly

DEVOTED TO RESEARCH STUDIES IN THE FIELD OF MASS COMMUNICATIONS

FALL 1953

"Cloze Procedure": A New Tool For Measuring Readability

BY WILSON L. TAYLOR*

Here is the first comprehensive statement of a research method and its theory which were introduced briefly during a workshop at the 1953 AEJ convention. Included are findings from three pilot studies and two experiments in which "cloze procedure" results are compared with those of two readability formulas.

"CLOZE PROCEDURE" IS A NEW PSYchological tool for measuring the effectiveness of communication. The method is straightforward; the data are easily quantifiable; the findings seem to stand up.

At the outset, this tool was looked on as a new approach to "readability." It was so used in three pilot studies and two experiments, the main findings of which are reported here.

First, the results of the new method were repeatedly shown to conform with the results of the Flesch and Dale-Chall devices for estimating readability. Then the scope broadened, and cloze procedure was pitted against those standard formulas.

If future research substantiates the results so far, this tool seems likely to have a variety of applications, both theoretical and practical, in other fields involving communication functions.

THE "CLOZE UNIT"

At the heart of the procedure is a functional unit of measurement tentatively dubbed a "cloze." It is pronounced like the verb "close" and is derived from "closure." The last term is one gestalt psychology applies to the human tendency to complete a familiar but not-quite-finished pattern—to "see" a broken circle as a whole one, for example, by mentally closing up the gaps.

*The writer is particularly obligated to Prof. Charles E. Osgood, University of Illinois, and Melvin R. Marks, Personnel Research Section, A.G.O., Department of the Army, for instigating and assisting in the series of efforts that yielded the notion of "cloze procedure." Both are experimental psychologists. Among others who have advised, encouraged or otherwise aided are these of the University of Illinois: Prof. Lee J. Cronbach, educational psychologist and statistician; Dean Wilbur Schramm, Division of Communications; Prof. Charles E. Swanson, Institute of Communications Research, and George R. Klare, psychologist, both of whom have authored articles on readability; and several journalism teachers who lent their classes. Kalmer E. Stordahl and Clifford M. Christensen, until recently research associates of the Institute, also contributed.

One can complete the broken circle because its shape or pattern is so familiar that, although much of it actually is missing, it can be recognized anyway.

The same principle applies to language. Given "Chickens cackle and ——— quack," almost anyone can instantly supply "ducks." If that word really is the same as the one omitted, the person scores one cloze unit for correctly closing the gap in the language pattern.

> Note that the sentence pattern is a complex one made up of many subpatterns. One must know not only the meanings (i.e., patterns of symbol-meaning relationships) and forms (patterns of letters) of all the five words, but also the meanings of given combinations of them—plus the fact that the sentence structure seems to demand a term parallel to "cackle" but associated with ducks instead of chickens. In other words, one must guess what the mutilated sentence means as a whole, then complete its pattern to fit that whole meaning.

A cloze unit may be defined as: Any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product) by deciding, from the context that remains, what the missing part should be.

Cloze procedure may be defined as: A method of intercepting a message from a "transmitter" (writer or speaker), mutilating its language patterns by deleting parts, and so administering it to "receivers" (readers or listeners) that their attempts to make the patterns whole again potentially yield a considerable number of cloze units.

HOW THE METHOD WORKS

As defined, the concept of cloze procedure involves both oral and written communication and does not specify any particular kind of "part" for deletion. The research on which this report is based, however, employed only reading materials and deleted only words.

In practice, the readabilities of two or more passages of about equal length were contrasted, for any given population, by:

1. Deleting an equal number of words from each passage by some essentially random counting-out system. Such a system was based on a table of random numbers or else it simply counted out every *n*th word (every fifth one, for example) without any regard for the functions or meanings of specific words.
2. Reproducing each mutilated passage with a blank of some standard length (so the length would not influence the guessing) in place of every missing word.
3. Giving copies of all reproduced passages to all subjects—or to equal numbers of randomly selected subjects —in a sample group representative of the population in question.
4. Asking all subjects to try to fill in all blanks by guessing, from the context of remaining words, what the missing words should be.
5. Totaling for each passage separately the number of times original words were correctly replaced, and considering these totals as readability scores.
6. Contrasting the cloze totals of the various passages. The passage with the highest score was considered "most readable," the one with the second-highest score next-most readable, etc.— pending the outcome of statistical tests of the significance of the differences observed.

SOME DISTINCTIONS

Cloze procedure is neither just another readability formula nor just another form of the familiar sentence-completion test.

Not a Formula

The cloze method is not a formula at all.

Neither in theory nor practice does it resemble current "element counting"

devices (Flesch, Dale-Chall, etc.) which assume a high correlation between ease of comprehension and the frequency of occurrence of selected kinds of language elements—short or common words, short or simple sentences, certain parts of speech, the active voice, "concrete" terms and such.

Cloze procedure counts no such elements. It seems, however, to measure whatever effects elements actually may have on readability. And it does so at the same time that it is also taking account of the influences of many other factors readability formulas ignore.

Typically, the formulas are insensitive to a particular population's previous knowledge of the topic being discussed. They cannot allow for the effects of non-idiomatic uses of common words, nonsense combinations of words, awkward and confusing sentence structure or pronouns without definite antecedents. And the basic assumptions of formulas may be directly contradicted.

> "Respectability," despite its six syllables and high level of abstraction, is much easier for the average reader than "erg."
>
> "He came in smiling and sat down" is *not* approximately two or three times as difficult as "He came in. He was smiling. He sat down."
>
> "I came like Water, and like Wind I go" from the Rubaiyat makes no sense to second-graders even if they do know all the words.

One can think of cloze procedure as throwing all potential readability influences in a pot, letting them interact, then sampling the result.

The procedure also might be likened to a polling method with experimental controls. It asks members of a population sample to *demonstrate* how well they understand the meaning of a mutilated version of what some writer wrote by having them "vote" on what the missing words should be. The passage whose deleted words are most often "written in" on the "ballot" is elected most readable.

More precisely, the cloze method seems to deal with more-or-less parallel sets of *meaning-pattern relationships.* Different persons may express the same meaning in somewhat differing ways, and the same language patterns may have differing meanings for different people. Cloze procedure takes a measure of the likeness between the patterns a writer has used and the patterns the reader is anticipating while he is reading.

Not a Sentence-Completion Test

Obviously, cloze procedure is something like this familiar form of examination. It is similar in that the subject is presented with incomplete sentences and there are blanks to be filled in from context.

But the typical sentence-completion test is for gauging a person's knowledge of specific and more or less independent points of information, hence the words to be deleted are pre-evaluated and selected accordingly. And for every new topic, some well-versed person must construct and try out a new test based on another set of information.

For one thing, cloze procedure deals with contextually interrelated series of blanks, not isolated ones.

For another, the cloze method does not deal directly with specific meaning. Instead, it repeatedly samples the extent of likeness between the language patterns used by the writer to express what he meant and those possibly different patterns which represent readers' guesses at what they *think* the writer meant.

However, because it counts instances of language-usage correspondence rath-

er than meanings themselves, the cloze unit seems to classify as a *common denominator* of communication success; and with it the readabilities of materials on totally different topics can be compared directly.

For this sort of contrast, an essentially random deletion of words seems required. And this makes the task of actually picking out what words to delete purely clerical—and so simple that anyone who can count to ten can do it for any sort of material, regardless of its topic or difficulty.

SOME THEORETICAL CONSIDERATIONS

The main contributions to the notion of cloze procedure have come from the concepts of "total language context," Osgood's "dispositional mechanisms" and statistical random sampling.

1. Total Language Context

For more than half a century, experimenters have been reporting findings that may be interpreted as showing that language behavior depends on "total context."

> The results indicate that the ability to identify, learn, recognize, remember or produce any language "symbol" (element or pattern) depends heavily on the variable degrees to which it is associated with everything else by larger and meaningful (familiar) overall combinations.[1]

The total context of any language behavior includes everything that tends to motivate, guide, assist or hinder that behavior. It includes verbal factors—grammatical skills and multitudes of symbols—and non-verbal ones such as fears, desires, past experience and intelligence.

> "I heard a —— bark" is likely to elicit "dog" both because that word is habitually associated with "bark" *and* because it fits in with past experience with noisy dogs. If the verbal context is enlarged to "For the first time, I heard a —— bark," the impulse to supply "dog" may be reduced by commonsense; the subject may ask himself: "Who is this guy that has never heard a dog? Could he be referring to some other animal?" And if the preceding sentence has mentioned a voyage to the Pribilof Islands, the reader may draw on past knowledge to produce "seal."

Quite recently, Marks and Taylor reported an experiment in which the influences of varying intensities of both verbal and non-verbal contextual factors on the generation of language elements were shown to be measurable by quantitative methods.[2]

2. Dispositional Mechanisms

The notion of cloze procedure was "sparked" by implications of Osgood's learning theory of communication. He relates the "redundancies" and "transitional probabilities" of language to the development of "dispositional mechanisms" that play a large part in both transmitting and receiving messages.[3]

> *Redundancy*—"Man coming" means the same as "A man is coming this way now." The latter, which is more like ordinary English, is redundant; it indicates the singular number of the subject three times (by "a," "man," and "is"), the present tense twice ("is coming" and "now"), and the direction of action twice ("coming" and "this way"). Such repetitions of meaning, such internal ties between words, make it possible to replace "is," "this," "way," or "now," should any one of them be missed.

[1] For a comprehensive summary of these contributions to communication theory, the reader is referred to George A. Miller's *Language and Communication* (New York: McGraw-Hill, 1951).

[2] M. R. Marks and Wilson L. Taylor, "The Effect of Goal and Contextual Constraints Upon Meaningfulness of Language," paper presented by Marks at annual meeting of American Psychological Association, Chicago, 1951; summary in *American Psychologist*, 6:325 (1951).

[3] For a fuller explanation of this topic, the reader is referred to Charles E. Osgood: (a) "The Nature and Measurement of Meaning," *Psychological Bulletin*, 49:197–237 (May 1952); (b) *A Theory of Language Behavior* (tentative title), monograph in preparation, Institute of Communications Research, University of Illinois; (c) *Method and Theory in Experimental Psychology* (New York: Oxford, 1953).

Transitional Probabilities — Some words are more likely than others to appear in certain patterns or sequences. "Merry Christmas" is a more probable combination than "Merry birthday." "Please pass the ———" is more often completed by "salt" than by "sodium chloride" or "blowtorch." Some transitions from one word to the next are, therefore, more probable than others.

Dispositional Language Habits—In learning "to think in" a language, an individual develops an enormous number of complex verbal skill patterns—"bundles of skill sequences"—which stand for innumerable kinds and shades of meaning and tend to become so automatic that they "run themselves off" in pertinent situations. These habits reflect the redundancies and transitional probabilities of the language patterns these skills involve.

Out of his personal experiences and circumstances, each human develops his own set of these habits. To the extent that his set corresponds to the sets of others in his culture, he can communicate easily; he and they have learned similar meaning-language relationships —the same patterns of symbols go with the same meanings. But any two sets of language mechanisms can differ considerably within the same culture; one man becomes more disposed to run off particular sequences than another man does. To the same extent, the related sets of redundancies and transitional probabilities can differ also.

Habits of expression take over most of the work of translating an individual's meaning into an organized series of language symbols for transmission to others. Likewise, his habits of reading or listening cause him to anticipate words, almost automatically, when he is receiving messages. When he sees the start of a phrase that looks familiar, he immediately tends to complete it in his own way even when the written phrase actually ends differently.

When words come in sequences that best fit the existing receiving habits of a reader, he understands with little effort. When the symbols appear in less familiar sequences, comprehension is slower and less sure. And sufficiently improbable patterns seem like nonsense; they do not stand for anything in his experience.

3. *Random Deletion*

A random deletion method (or an every-*n*th equivalent) which ignores the differences between specific words appears to be not only defensible but rationally inescapable when cloze procedure is used for contrasting readabilities.

The main reasons for this view relate to two questions most often asked by those who have seen the data of this report.

Question 1: How can a random system play fair when some words are easier to replace than others?

> Obviously, one is more likely to be able to supply "an" in "A quinidine is —— alkaloid isomeric . . ." than to guess "$6,425" in "The city council voted ———— for a new swimming pool." Yet the former example is far more difficult reading.

The answer is that if *enough* words are struck out at random, the blanks will come to represent proportionately all kinds of words to the extent that they occur. The matter boils down to "How many blanks are enough?"—a problem to be settled by experiment.

> Somewhat the same principle is involved in the substitution of a more convenient every-*n*th system for a random one. For several blanks, an every-*n*th system might tend to fall in with the "rhythm" of an author's style and take out mostly nouns, or mostly articles. . . . The answer is that rhythms break, and *n*th deletions, if continued long enough, start taking out other parts of speech and, eventually, yield the equivalent of random deletion. Again, the practical question is "How many are enough?"

Question 2: Wouldn't a deletion system be more sensitive and more reliable if it dealt only with words classified, say, by their "importance to meaning" or their familiarity as gauged by tabled frequencies of use?

The answer seems to be "No."

For one thing, specified words or kinds of words may not occur equally often in different materials. That fact itself may be a readability factor, and its effect can be measured only by a method that operates independently.

> An attempt to restrict a counting-out system to "important" words (nouns and verbs, for example, as against articles and conjunctions) may find that one of two equally long passages contains twice as many "importants" as the other! What then?
>
> Because the effect of such a difference needs to be included in—not excluded from—the results, it seems necessary to let the occurrences of presumably important words be represented proportionately in deletions.
>
> What has just been stated about deleting only "important" words applies with equal force to varying degrees of familiarity.

Also, it should be remembered that cloze procedure deals only with words as they actually occur in larger patterns which stand for particular meanings at the time they are transmitted or received. The result is that infrequently used words may not be hard to replace at all; and supposedly unimportant words may become extremely so.

> Most Americans can effortlessly supply "tipped" and "lady" in "The polite old gentleman always ——— his hat when he met a ———." The ability to do so has very little to do with the frequency with which those words, considered individually, occur in the language; "the," which is hundreds of times more common, simply doesn't fit in either blank. The kind of frequency that most matters here is frequency-in-context.
>
> An article can be more important to meaning than any other word in its sentence, and harder to replace than the words in the previous example. "You want to know what the wolf did to the sheep? He killed —— sheep." Note that "sheep," a noun and the object of the verb, matters hardly at all in its second appearance. Also, if the missing word is "a," it would be quite difficult to guess correctly—because "the," "some," "every," "many," "no" or some finite number could fit too.

"SCORES"—NOT FREQUENCIES

For the purpose of statistical analysis, cloze data are treated as "true scores" throughout this report.

This is in conformance with the opinion of Lee J. Cronbach. He said the nature of cloze results satisfies the assumptions for scores, but not those for chi-square frequency tests because successive cloze units cannot be considered independent.

If, in "Then he took off his hat," "he" and "his" both were blanked out, getting the first right would probably mean getting "his" right too; just as "she" would go with "her."

At first sight, cloze results appeared to be frequencies (the mere number of times missing words were correctly replaced). But "correctly" implies an underlying qualitative continuum of relative rightness ranging from a completely inappropriate word, through poor, medium and good synonyms, to the exact word left out. Whether or not only precise matches are counted does not affect the existence of such a continuum.

GENERAL PROCEDURE; PLAN OF REPORT

For clarity's sake, it seems best to present the report of Experiment 1 completely—specific purposes, procedure and results—before starting on Experiment 2. However, some procedural aspects were common to all research designs used.

For deletion purposes, a "word" was defined by the white spaces with which an author had separated it from other words.

Contractions like "don't"; abbreviations like "Mr." or "U.N."; figures like "1,006"; hyphenated combinations like "self-conscious"—all these were considered as single words. But "self conscious," written without the hyphen, counted as two.

All passages were 175-plus words in length; the sentence in which the 175th word occurred was reproduced in full.

Only "mechanical" deletion systems, random or every-*n*th, were employed.

After mutilation, each version of a passage was reproduced on separate typewriter-size sheets with an underlined blank 10 spaces long in place of each missing word.

The sheets for any one subject were assembled in a predetermined order and stapled together with an explanatory foreword. Such assemblies were randomly assigned to subjects.

Pilot studies used small groups of subjects, a dozen or less, chosen on a casual "handy and willing" basis. The subjects were not supervised and no time limits were set.

The experiments, however, used larger groups made up exclusively of juniors and seniors in University of Illinois journalism courses. The administrator read the foreword aloud and answered questions regarding it before subjects began work. Each subject was allowed from 10 to 15 minutes for filling in the blanks of each mutilated passage he received.

EXPERIMENT 1

PURPOSES AND PROCEDURE

Three passages borrowed directly from Flesch's *How to Test Readability*[4] were used in Experiment 1 and the two pilot studies which preceded it. Listed in a footnote[5] are the original sources of the passages, the pages on which they are found in the Flesch book, the "reading ease" scores given by that author, and the Dale-Chall scores computed by this experimenter.[6]

The main purpose here was to see if cloze scores also would rank these passages as the formulas do. If the ranks were maintained, the differences among the cloze scores would be tested for statistical significance.

The research designs of this experiment and its pilot studies also sought answers to several methodological questions. It was asked whether significant differences in relative cloze scores would accompany variations in:

(1) Word-deletion systems—
  a. Random deletion vs. the more convenient system of counting out every *n*th word.
  b. Counting out every fifth word, for example, vs. every tenth.
  c. Counting out "few" (16) words per passage vs. "many" (35) words.
(2) Presentation orders—
  Presenting mutilated passages in one order vs. presenting them in another.
(3) Scoring methods—
  Scoring as correct only those fill-ins

[4] Rudolf Flesch, *How to Test Readability* (New York: Harper, 1951).

[5] *Source—Author & Work*	*Page*	*R.E. Score; Interpreted*	*Dale-Chall "raw"*	*Dale-Chall Gr. level*
James Boswell, *Life of Johnson*	12	89—"easy"	6.4	7th-8th
Julian Huxley, *Man Stands Alone*	16	68—"standard"	7.1	9th-10th
Henry James, *The Ambassadors*	22	47—"difficult"	9.2	13th-15th

(NOTE: Dale-Chall scores yield lower—instead of higher—values for "easier" and are interpreted in terms of school-grade levels.)

[6] Method described in pamphlet: Edgar Dale and Jeanne S. Chall, *A Formula for Predicting Readability* (Ohio State University, Bureau of Educational Research); reprints two articles from *Educational Research Bulletin*, 27:11–20, 37–54 (Jan., Feb. 1948).

that precisely matched original words vs. the more tedious process of judging synonyms and allowing half for each "good enough" one.

To explore the word-deletion questions, each passage was mutilated by three different systems: "Every fifth," "every tenth" and "random 10 percent."

In Pilot Study 1, the initial word, then every fifth word thereafter, was knocked out until 35 words, about 20 percent of the total number of original ones, were missing.

Both Pilot Study 2 and Experiment 1 contrasted random and every-*n*th systems of deleting about 10 percent of total wordage.

One system was based on a table of random numbers; the other counted out every tenth word. Each system deleted 16 original words from each passage.

(Almost entirely different sets of words were taken out by the different deletion methods. There was no overlap at all between the every-*n*th systems. The random 10 percent system took out only two of the same words deleted by each of the every-*n*th ones.)

To discover possible order effects, all six orders of three passages (abc, acb, bac, bca, cab, and cba) were equally represented in each pilot study and in the experiment.

The scoring-method question was disposed of after data of Pilot Study 2 were evaluated in both ways and the results were compared.

In summary, these hypotheses were developed and tested by Experiment 1 and its pilot studies:

1. Cloze scores would rank the three passages in the same order as the two standard formulas.

2. For any condition of data gathering, the overall difference among the scores of the three passages would yield a significant value of F, as computed by analysis of variance, double classification.

Because cloze data for the three passages would be correlated, such computation also would yield an F for the overall difference among subjects.

3. The relationship between the cloze scores for the three passages would remain essentially the same despite the use of different word-deletion systems or the specific words they counted out, despite different presentation orders, and despite different scoring methods.

Pilot Studies 1 and 2 used 6 and 12 subjects, respectively. Experiment 1 used 24 subjects. In every case, a subject received all three passages.

RESULTS

This summary concerns itself mainly with the findings of Experiment 1, which tested hypotheses inferred from the empirical findings of the preceding pilot studies. However, certain pilot-study results also are presented.

Some findings of both studies appear in Table 1 with reference to the "existence" and discriminatory powers of cloze scores as related to various methods of gathering them. Further, Pilot Study 2 was itself the experimental test of the hypothesis, drawn from the first study, that the evaluation and scoring of synonyms would *not* be profitable (See Table 2).

1. "Existence" of Cloze Measure of Readability

Entries in Table 1 show how consistently cloze scores ranked the three selected passages in the same order of readability as do the Flesch and Dale-Chall formulas. One could assume that cloze procedure and the formulas were measuring the same thing.

2. Power of Discrimination

For every experimental condition entered in Table 1, analysis of variance of the overall difference among the

cloze scores of the three passages yielded an F which was very highly significant, to above the 0.1 percent level of confidence.

All figures in the first five columns of Table 1 (this excludes validation entries from Experiment 2) involve correlated data; that is, all subjects were given all three passages, hence the type of analysis of variance called for ("double classification") discriminated among subjects as well as passages.

All but one of the between-subject F's were significant to the 5 percent level or higher.

3. *Methodological Findings*

Results tending to answer questions about methodology appear in both Table 1 and Table 2.

TABLE 1

Maintenance of Ranks of Three Passages; Significance of Overall Difference among Clozes

	Pilot Study		*Experiment 1*			*Experiment 2*
	1	*2*	*R-10%*	*Ev-10*	*Total*	*(validation)*
CONDITIONS OF DATA GATHERING						
S's per passage:	6	12	12	12	24	18
Blanks per passage:	35	16	16	16	32	25
Deletion frequency:	20%	10%	10%	10%	10%	14.3%
TOTAL CLOZE SCORES PER PASSAGE						
Ranked by Formulas						
1. Bos-J:	136	118	113	87	200	186
2. J. Hux:	87	97	80	71	151	155
3. H. Jas:	63	72	63	53	116	135
OVERALL DIFFERENCE WITHIN EACH GROUP OF THREE*						
Between Passages—						
Computed F:	57.855	15.447	60.455	9.867	42.151	9.418**
Deg. Freedom:	2&10	2&22	2&22	2&22	2&46	2&69
P less than:	.001	.001	.001	.001	.001	.001
Between Subjects—						
Computed F:	12.159	2.531	5.735	1.872	3.327	....
Deg. Freedom:	5&10	11&22	11&22	11&22	23&46	....
P less than:	.001	.05	.001	(not)***	.001	....

*Significance of overall difference among scores in each condition computed by analysis of variance. . . . **Experiment 2's data for each passage were based on relatively separate groups of subjects, hence data considered uncorrelated and computation was by "simple classification"—one-dimensional—analysis of variance, for passages only. . . . ***F of 1.872 fails to reach the 5 percent level of confidence.

This table shows (1) that the readability rank order given three passages by both the Flesch and Dale-Chall formulas was maintained by cloze scores obtained under a variety of conditions; (2) that the overall difference between the scores of the different passages was significant in all conditions to above the 0.1 percent level of confidence; (3) that the two findings just stated were verified by subsequent data from Experiment 2. Also, it was discovered that most conditions yielded F's which also discriminated among subjects. In the breakdown of Experiment 1 results, the fact that the every-tenth deletion pattern did not yield a significant between-subjects F is evidence that it was a less efficient pattern than its corresponding random 10 percent one, but the findings of both patterns were qualitatively similar, hence they were considered combinable for further analysis.

a. *Deletion Systems*—Table 1 shows that variations in totals of blanks per passage, amounts of deletion ranging from 10 percent to 20 percent, random and every-*n*th systems, and almost entirely different sets of specific words knocked out—all were accompanied by results that consistently upheld both the "existence" of cloze procedure as a readability measure and its power to discriminate between passages.

Qualitatively, then, all conditions yielded the same results. There appeared, however, some quantitative differences—that is, some conditions were more "efficient" than others, particularly with regard to the discovery that the analysis also discriminated among subjects.

The 35 blanks and every-fifth deletion of Pilot Study 1 discriminated better between its six subjects than did Pilot Study 2, based on only 16 blanks, between its 12 subjects. (Perhaps the former group was simply more heterogeneous, but it also may be that the degree of significance found depends on plenty of blanks or on greater intensities of deletion.)

TABLE 2
Cloze Data on Two Methodological Problems

(A)

WHEN SYNONYMS ARE CONSIDERED

From Pilot Study 2 Data
(12 Subjects, 10 percent deletion, 16 blanks per passage)

		Passages			
		Bos-J	*J. Hux*	*H. Jas.*	*Total*
Precise Matches Only:	Score:	118	97	72	287
	p:*	.41	.34	.25	1.00
Matches Plus Synonyms:	Score:	139.5	122.5	91.5	353.5
(½-count allowed for each of 133 judged as "good enough")	*p*:	.39	.35	.26	1.00

*Proportions of total score associated with each passage.

This shows that allowing half-scores for "good enough" synonyms and adding those counts to the precise-match scores raised the scores somewhat but did not improve discrimination between passages. The proportions are almost identical.

(B)

COMPARISON OF PRESENTATION ORDERS

From Data in Experiment 1
(24 Subjects, 10 percent deletion, 16 blanks per passage;
Random 10 percent and Every-10th results combined)

	Total Cloze Scores for Every Place in Presentation Order			
	1st	*2nd*	*3rd*	*Total*
Passages				
Bos-J:	65	66	69	200
J. Hux:	48	52	51	151
H. Jas:	41	36	39	116
Totals:	154	154	159	467

For any one passage, i.e., any one row, the subtotals for each of the three places differ only slightly. Order effects were assumed to be unimportant.

In the breakdown of Experiment 1 results between random and every-*n*th systems, the random 10 percent yielded a highly significant between-subject *F*, but the every-tenth data, while qualitatively similar, failed to reach the 5 percent level. (It seems that the contrast should be interpreted as meaning that random and every-*n*th systems will give more nearly equivalent results if more than 16 blanks are deleted per passage.)

b. *Synonyms*—Table 2 entries indicate that the more tedious method of judging synonyms as "good enough" to be allocated half-counts yielded slightly larger total scores for the passages, but the degree of differentiation was virtually identical to scoring only precise matches.

Note the similarity in the proportions of the two kinds of totals broken down between passages.

c. *Presentation Orders*—Table 2 also shows that the orders in which the three passages were presented had virtually no effect on their scores. (It was assumed, therefore, that order effects are unimportant.)

EXPERIMENT 2

PURPOSES AND PROCEDURE

The list of passages for this experiment was expanded to eight by adding five more to the three already used.

Two considerations governed the choice of additional passages. Desired were (1) a list of passages that seemed to be distributed fairly evenly over a long hard-to-easy range and (2) the inclusion of materials which, it was thought, would show that cloze procedure could "handle" passages which neither of the standard formulas could.

Both considerations were served by the choice of passages from:

Erskine Caldwell, *Georgia Boy*[7]

Gertrude Stein, *Geography and Plays*[8]

James Joyce, *Finnegans Wake*[9]

In the experimenter's opinion, the first of these is by far the easiest to read of all the passages used, and the other two are next-hardest and hardest, respectively. It was expected that what were subjectively assumed to be the "true" readabilities of all three selections would be under-rated or over-rated by either or both formulas.

> The Stein selection, although it seemed comparatively unintelligible, was characterized by words both short and familiar and by short sentences. . . . It was thought that both formulas would over-rate its readability inasmuch as the Flesch device counts only the number of syllables per word and the number of words per sentence, and the Dale-Chall method counts only the number of words per sentence and the number of words not on a list of familiar words.
>
> The other two passages were expected to "fool" the Flesch formula more than the Dale-Chall one.
>
> It was thought that the Caldwell selection would be under-rated by the Flesch method because it contains rather long sentences consisting largely of independent clauses connected by "and," and because this fact would not be offset by consideration of the familiarity of the vocabulary.
>
> Flesch's formula was expected to over-rate the Joyce selection's readability. That passage is made up of relatively short words and sentences, but many of the words appear in no dictionary.

The two selections still needed to bring the total to eight were chosen after trying out four other passages in Flesch's book on readability testing. Those four, together with the places where they are reprinted, were from

---

[7] From short story, "My Old Man and Pretty Sooky," passage beginning "My old man picked up one morning . . ." p. 80 (New York: Avon, 1946).

[8] Passage beginning "So the main is seen and the green is green," pp. 84-85 (Boston: Four Seas, 1922).

[9] Passage beginning "But who comes yond with pire on poletop?" p. 244 (New York: Viking, 1943).

Swift's *Gulliver's Travels,* p. 14; Charles Dickens' *Bleak House,* p. 15; William James' *Psychology,* p. 19; and Article I, Section 10, U. S. Constitution, p. 28.

Altogether, 10 passages, these four and the six already earmarked for Experiment 2, were used in Pilot Study 3. Each of the six subjects received copies of all and was allowed to take them home and complete them at leisure.

All passages used in this pilot study and the final experiment were mutilated by still another deletion system; every seventh word was counted out until each passage had 25 blanks.

Analysis of the pilot study's results showed that the Swift and Dickens selections tended to distribute themselves conveniently with respect to the six passages already chosen, hence the ones from William James and the Constitution were discarded.

The eight passages thus chosen for Experiment 2 then were ranked by their median cloze scores, and the order in which they fell was considered a prediction of how larger groups would rank them.

Additional Flesch and Dale-Chall scores were computed, and the ranks in which they fell also were considered as predictions.

These were the chief hypotheses tested by Experiment 2:

1. The main results of Experiment 1 would be validated.

2. The readability order of eight passages, as empirically determined by Pilot Study 3, would dependably predict the rank order of the cloze scores given them by other and larger samples of subjects.

3. Cloze procedure would handle materials which either or both of the standard formulas could not.

4. Cloze predictions would be more successful than those of the two formulas in guessing the ranks of final cloze scores.

5. The overall difference among the cloze scores of the various passages would be found significant when submitted to analysis of variance.

Because only half an hour was available for the administration of Experiment 2, each subject received only two of the eight passages. Each of the selections Pilot Study 3 indicated as among the four easier ones was paired with each of those which seemed to be among the four harder, the easier always preceding the harder in assembly.

The 16 kinds of pairs were represented four times each in the final data, to which a total of 72 subjects contributed. Each passage was administered to 18 subjects.

The final experiment's data also were subjected to certain exploratory attacks:

1. To see how finely the cloze method seemed to discriminate between readability levels, passages adjacent in final rank were taken two at a time, and the significance of the difference between their mean scores was computed by the *t*-test for small samples.

2. What may be called the "internal consistency" of cloze results was attacked from the standpoints both of passages and of subjects by "fractionation" methods.

> The total cloze score for each passage was broken down into subtotals showing the aggregate score for that passage for the first five blanks, the second five blanks, etc., until all fifths of 25 blanks were represented. The passages then were ranked by these subtotals for each five-blank series. Then the degree of overall rank correlation for all passages for all five of these fifths of total blanks was computed.
>
> Then the question was asked: Did subjects who were higher in total cloze

scores also tend to make consistently higher scores for fractions of the 25-blank series? The subjects for each passage were divided into thirds, and the performances of those in the upper third—as judged by total scores—were compared with the performances of those in the lower third.

RESULTS

Data pertinent to all findings under this heading, except for the "validation" one disposed of in the next paragraph, are shown in Tables 3 and 4.

1. *Validation of Experiment 1's Findings*

Entered in Table 1 are certain of Experiment 2's results that verify the "existence" and power-of-discrimination findings for the three passages used in Experiment 1 and its pilot studies.

Note that Experiment 2 used still another deletion system on the passages by Boswell, Huxley and Henry James. Its every-seventh method took out a set of 25 words that overlapped the 10 percent systems, random and every-tenth, only two words each; and it struck out only five of the 35 words which the every-fifth system deleted.

2. *Predictive Success of Pilot Study 3*

Table 3 shows that Pilot Study 3's median scores for six subjects ranked the eight passages in almost exactly the same way as did the performances of the 18-subject groups used in Experiment 2.

The rank difference coefficient of .98 exceeds the 0.1 percent level of confidence. The Boswell and Huxley passages in the middle of the distribution changed places, but Table 4 shows that those two passages did not receive significantly different cloze totals anyway.

Note that although the same six subjects participated in the predictions for all eight passages (hence Pilot Study 3's data were correlated), in the experiment that followed, each passage was rated by the performance of a relatively separate and independent group of subjects. Experimental conditions allowed only two passages to be administered to any one subject, and the experiment's design was such as to reduce correlation between the results of any two passages to the minimum of 28 percent; that is, only five of the 18 subjects who rated any passage also were among the 18 who rated any other passage. Experiment 2's scores for different passages were, therefore, assumed to be independent.

3. *Handling of Certain Materials*

It seems indisputable that cloze procedure came closer than either of the standard formulas to ranking properly the relative readability levels of certain passages.

The Stein passage, for example, is ranked first by both formulas. The Flesch method concludes it is "very easy" reading, and the Dale-Chall score rates it as within the comprehension level of fourth or fifth grade school children!

Otherwise, the two formulas did not perform very similarly. The Dale-Chall ranks for the Caldwell and Joyce selections appear to make much more sense than the Flesch ratings of them.

4. *Cloze Predictions Better for Population Used*

It follows from the findings just set down that previous cloze results were more successful than those of the two standard formulas in predicting the ranks of future cloze results for the population used.

Although the rank correlation, .70, between the Flesch and Dale-Chall predictions is moderately significant (to above the 5 percent level of confidence), neither of those sets of predictions correlate significantly either with cloze predictions or with final cloze results.

The lower part of Table 3 demonstrates how the removal from consideration of the data from any one—or any two—of the three handpicked passages (Stein, Joyce and Caldwell) would change the measures of the predictive accuracies of the formulas.

5. *Significance of Difference Among Passages*

Table 4 shows that analysis of variance, simple classification, yielded an *F* for the overall difference among the scores of various passages which is significant far above the 0.1 percent level.

Because each passage was administered to a relatively distinct group of 18 subjects, Bartlett's homogeneity of variance test was used. Its non-significant finding justifies the assumption that the differences between passage scores need not be attributed to mere differences between the groups.

TABLE 3

Cloze, Flesch & Dale-Chall Predictions Compared with Respect to Experiment 2 Cloze Results

(Every-7th; 25 Blanks Per Passage; 6 S's Predict for Groups of 18.)

RANK-ORDER PREDICTIONS

Passages	Cloze P. Study 3 Mdn.	Cloze P. Study 3 RO	Standard Formulas Flesch R.E.	Standard Formulas Flesch RO	Standard Formulas Dale-Chall "Raw"	Standard Formulas Dale-Chall RO	Exp. 2 Results Cloze Totals	Exp. 2 Results Final RO
Caldw	19.5	1.0	79	4.5	5.6	2.0	336	1.0
Dickn	16	2.0	69*	6.0	6.7	4.0	263	2.0
Bos-J	12	3.0	89*	2.0	6.4	3.0	186	3.0
J. Hux	11	4.0	68*	7.0	7.1	6.0	155	5.0
Swift	10.5	5.0	80*	3.0	7.0	5.0	170	4.0
H. Jas	9	6.0	47*	8.0	9.2	7.5	135	6.0
Stein	8.5	7.0	96	1.0	5.0	1.0	123	7.0
Joyce	3	8.0	79	4.5	9.2	7.5	49	8.0

RANK-DIFFERENCE CORRELATION COEFFICIENTS

Between Predictions		*Success of Predictions*	
Cloze vs. Flesch	—.12	Cloze vs. Final	.98**
Cloze vs. Dale-Chall	.44	Flesch vs. Final	—.02
Flesch vs. Dale-Chall	.70	Dale-Chall vs. Final	.46

Effects of Including Joyce, Caldwell, and Stein Passages

(As determined by making removals indicated, re-ranking remaining selections, and recomputing coefficients.)

Predictions vs. Final:	w/o Joyce	w/o Caldw	w/o Stein	w/o J.&C.	w/o J.&S.	w/o C.&S.
Cloze	.96***	.96	.96	.94***	.94	.94
Flesch	—.04	—.04	.33	—.03	.64	.71
Dale-Chall	.21	.40	.92	.09	.94	.93

*These five R.E.'s reported by Flesch himself; remaining R.E.'s and all Dale-Chall scores computed by experimenter.

**Significant; *P* less than .001.

***Slight reduction in cloze prediction coefficients reflects shortening series of passages—from 8 to 7, then to 6.

The ranks of Pilot Study 3 scores from six subjects are shown to have predicted almost perfectly how larger groups of subjects would rank eight passages. The cloze predictions suffered only one reversal; it occurred in mid-scale and involved two passages whose scores were not significantly different anyway (see Table 4). The two standard formulas stumbled in their predictions. Both rated the Stein passage as most readable. The Flesch formula had a bad time with the Caldwell and Joyce passages; it rated them equally as "fairly easy."

TABLE 4

Significance of Differences between Means of Passages Adjacent in Rank

Data from Experiment 2
(18 Subjects and 25 Blanks Per Passage)

Passages in Final Rank O.	*Total Cloze Score*	*p* of 450*	*(All N's = 18)*		*Signif. Diff. Betw. Adjac. Means*	
			Means	*S.D.*	*"t"*	*P*
Caldw	336	.747	18.67	1.544		
					7.068	.001
Dickn	263	.584	14.61	1.800		
					6.719	.001
Bos-J	186**	.413	10.33	1.915		
					1.179	.25
Swift	170**	.378	9.44	2.454		
					1.035	.31
J. Hux	155**	.344	8.61	2.214		
					1.364	.18
H. Jas	135**	.300	7.50	2.522		
					.853	.40
Stein	123**	.273	6.83	2.034		
					6.990	.001
Joyce	49	.108	2.72	1.325		
					8.474	.001
(Zero)	(0)	(.000)	(0.00)	....		

Bartlett's Homog. of Var.: Chi-square = 13.582 w/7 df, P .10 to .05; not significant; group heterogeneity rejected.

Analysis of Var., "Simple": F = 100.183 w/7&136 df, P less than .001; difference among passages highly significant.

*Proportion observed cloze score was of total possible perfect score: 18 x 25 = 450.

**Comparison of alternate pairs of passages yielded these additional results:

	"t"	P
Bos-J vs. J. Hux	2.421	.03
Swift vs. H. Jas	2.274	.03
J. Hux vs. Stein	2.442	.02

Each passage was administered to a relatively separate group of 18 subjects, hence analysis assumed the means to be non-correlated. Actually, because each subject rated two passages, there was a fractional correlation between the subjects involved with any pair of passages; five subjects would be in both groups of 18. It was assumed legitimate to ignore this 28 percent overlap. . . . All P's shown are "two-tailed"; under justifying experimental conditions, such P's as these could be halved and significance levels raised.

EXPLORATORY FINDINGS

1. *Differences between Adjacent Passages*

Table 4 shows the results of t-test analysis of the differences between the mean scores of passages adjacent in final rank.

The sizes of the P's indicate that the Caldwell passage scored significantly higher than the Dickens one, the Dickens one significantly higher than the middle five (Boswell, Swift, Huxley, Henry James and Stein), and that whole group higher than the Joyce selection. Also, the last was significantly above hypothetical zero.

TABLE 5

Clozes, Net and Accumulated, for Fifths of 25 Blanks

	1st 5	*2nd 5*	*3rd 5*	*45th 5*	*5th 5*
High S's:	127 (127)	119 (246)	102 (348)	118 (466)	104 (570)
Low S's:	78 (78)	76 (154)	69 (223)	85 (308)	60 (368)

Within the group of five, no passage scored significantly higher than the one adjacent in rank, but those in alternate ranks were associated with moderately significantly different scores. (See footnotes of Table 4.)

The *P*'s in the table are all "two-tailed." That means that when experimental conditions justify the process—which passage is the more readable is already known, therefore the only question is how much—such *P*'s could be halved and the level of significance more easily reached.

2. *Internal Consistency*

The overall rank correlation between the performances of all eight passages, as expressed by their score subtotals on five-blank fifths of 25 blanks, was computed by a method described by Kendall.[10] It yielded a "W," or "coefficient of concordance," of .56, which is significant to above the 1 percent level of confidence.

This suggests that the passages would have been ranked in about the same way even if only, say, half as many words had been deleted in each passage.

This "*W*" of .56 is the equivalent of an "average rho" of .455, which could otherwise be computed by finding the ten possible rank difference coefficients between five sets of rankings, taken two at a time, and calculating the mean.

Further investigation into passage consistency considered larger fractions of 25-blank totals. When data associated with the 13th blank were discarded and results for the remaining 24 blanks were divided into thirds, the three ranks of subtotals yielded a higher *W*, of .68. And when all data were divided into dovetailed halves corresponding to odd- and even-numbered blanks (with the data of the 25th blank split between them) the two sets of rankings yielded a rank correlation coefficient of .95.

To investigate subject consistency, the total scores of those six subjects who scored highest on every passage were broken down to correspond to five-blank fifths of 25 blanks. Then the subtotals for the different passages were combined. The same was done for the six who scored lowest on every passage. The contrasting pairs of subtotals are shown in Table 5.

Pushing this line of inquiry farther, the aggregate clozes of the "highs" and "lows" for each passage were examined blank by blank as they accumulated. In the case of every passage, the cumulative score of the high scorers quickly separated from that of the low group—usually within the first five blanks—after which the groups continued to diverge.

SUMMARY AND CONCLUSIONS

It was assumed that the relative readabilities of two or more written passages of about equal length could be contrasted by mechanically deleting an equal number of words in each, then totaling, for each passage, the number of times subjects succeeded in reproducing missing words.

[10] Maurice G. Kendall, *The Advanced Theory of Statistics*, Vol. 1 (London: Griffin & Co., 1948).

1. *Ranked Relative Readability*

Under a variety of data-gathering conditions, cloze scores consistently ranked three selected passages in the same way that the Flesch and Dale-Chall readability formulas do. If the formulas accurately ranked them, cloze procedure did too.

2. *Statistically Useful Contrasts*

Also found was evidence that the cloze method could discriminate effectively between different levels of readability—that is, was sufficiently sensitive to yield statistically contrastable scores.

The overall difference among the scores of the three passages most used was found to be highly significant under all data-gathering conditions.

Likewise, the overall difference among the scores of eight passages used in the final experiment was highly significant.

Further, exploratory analysis of the final experiment's data showed that some of the individual differences between the scores of passages adjacent in rank, taken two at a time, were highly significant. And all differences between those alternate in rank were at least moderately significant.

3. *Reliability and Prediction*

The mere repetitive ranking of three passages in the same way was rough evidence of the reliability of cloze results. More dramatic was the finding that the median scores of six subjects on eight passages ranked those passages in almost precisely the same way as did subsequent scores of relatively independent groups of 18 subjects, one group to a passage.

Further evidence of the cloze method's reliability as a way to compare passages was the exploratory "internal consistency" finding that a significant overall rank correlation existed between the ways in which passages ranked on each five-blank fifth of Experiment 2's 25-blank series.

4. *Range of Applicability*

Cloze procedure, in both a pilot study and the final experiment, assessed the assumed "true" readabilities of passages by Caldwell, Gertrude Stein and James Joyce more adequately than either or both of the standard formulas. Both formulas, for example, seemed to err badly in ranking the Stein passage as most readable of the eight employed.

However, the point here is *not* that the formulas did poorly, for that was preordained. Those three passages were handpicked because their characteristics violated certain common assumptions about what sorts of language elements correlate highly with readability. The real point is that cloze procedure appeared to handle those passages adequately.

5. *Methodology*

Analyses of cloze scores gathered by contrasting methods yielded qualitatively similar and quantitatively significant differences for passages despite:

a. Different mutilation systems—random vs. every *n*th; deletion densities varying from 10 percent to 20 percent of original words; totals of blanks varying from 16 to 35 per passage; and almost entirely different sets of specific words blanked out.

b. The orders in which two or more passages were administered to the same subject.

c. Ignoring vs. allowing for synonyms in final scoring.

However, some methods seemed to be relatively more efficient quantitatively. More needs to be determined about the comparative advantages of random vs. every-*n*th systems; various intensities of deletion; minimum sizes of population samples, etc.

6. *Between Subjects*

The possibility of using cloze procedure for contrasting the reading abilities of different individuals—as opposed to

the readabilities of different passages—was clearly suggested by the between-subject *F*'s associated with finding between-passage *F*'s in the correlated data of Pilot Studies 1 and 2 and Experiment 1.

This suggestion also is supported by the results of the "internal consistency" exploration of the fractionated performances of high- and low-scoring individuals in Experiment 2.

DISCUSSION

Evaluation

At present, the more outstanding characteristics of cloze procedure appear to be: The fact that it worked, its "mechanical" simplicity, the range of influences it involves, and its apparent possibilities for future use.

All hypotheses developed and tested seemed to stand up. Of course, the conclusions do not yet bear much generalizing—they apply only to the specific passages employed—but they were consistent.

To avoid misunderstanding, it is emphasized that the observed contrasts in the readability of passages are not to be extended either to the works from which the selections were drawn, or to the styles of the authors themselves. To make conclusions about the works or the authors would require the use of adequately representative samples of materials.

The practical advantages of a procedure that is simple and straightforward, that does not involve "experts" or difficult administrative judgments, are obvious.

The discovery that presentation orders had no important effects on the score of a passage tends toward simplification of experimental designs.

The finding that the trouble of evaluating and scoring synonyms was unprofitable was most welcome. It appeared that the problem of "coder reliability" that so plagues "content analysis" could be avoided.

Because various kinds of deletion systems yielded corresponding results, most of the questions concerning them seem to boil down to using *enough* blanks and determining, by further experiment, whether an every-fifth system is more or less efficient than, for example, an every-tenth or every-15th one. Or how many every-*n*th blanks are necessary to yield a dependable equivalent to a random system that takes out enough words.

Potentially important, it seems, is the fact that a cloze score appears to be a measure of the aggregate influences of *all factors* which interact to affect the degree of correspondence between the language patterns of transmitter and receiver. As such, its potential usefulness is by no means confined either to readability or to the reading abilities of individuals.

It seems that the effect of any manipulable factor, or any combination of two or three of them, might be measured by properly controlled experiments which contrast the cloze performances of groups equated or homogeneous with respect to other factors.

As it now stands, cloze procedure not only needs research to make it as efficient as possible but it needs to be better validated and the range of its possible applications explored.

More research is under way. Some of it is directed at the problem of validating cloze scores as measures of comprehension—that is what the term "readability" seems to imply. Tentative results indicate that cloze scores correlate highly with the scores of tests designed to measure comprehension and general intelligence.

Regarding "Formulas"

It is not the purpose of this report to disparage the Flesch, Dale-Chall and other formulas. The two named were

chosen for experiment because they are among the more prominent in the field.

Also, it must be admitted that these and most other formulas have certain advantages over cloze procedure.

They are easier and quicker to apply. Their use does not require word deletion, the reproduction of materials, experimental controls, and representative population samples. (They do, of course, have similar material-sampling problems.)

Also, for what may be called "standard" materials, these formulas seem reasonably accurate—the occurrences of the elements they choose to count usually *do* correlate better than chance with such criteria of validity as comprehension test scores and lists of graded readings.

And they are "reliable." With relatively little training, different users of the same formula get virtually identical results for the same materials. Also, the results of different formulas have often been shown to correlate significantly.

Their disadvantages, however, are too important to be overlooked.

There seems no positive way to identify in advance which materials are "standard" enough to be handled by a particular formula.

"Reliability" isn't everything; a formula can be *reliably wrong.* Anyone using either the Flesch or Dale-Chall formula on the eight passages in Experiment 2 would have to conclude that the Stein selection is the "easiest."

Also, it seems that a readability gauge might well be flexible enough to apply not only to a generalized "average American," for example, but to particular populations too. It is a little unreasonable that a single readability score for an article on cattle breeding should apply alike to residents of Texas "cow country" and metropolitan Brooklyn. In such cases, it appears that the user of a formula might employ cloze procedures to check up on his results.

"The editorial page [should be] the instrument through which a newspaper seeks to influence public opinion in what its publisher and editors, its reporters and deskmen consider to be the right direction.

"To do this effectively, a newspaper must work at the task steadily and consistently, not be forever jumping from one crusade to another. After all, a newspaper, like an individual, only has so many basic ideals and principles. But by applying them to various situations as they arise, a newspaper can indoctrinate its readers with its own beliefs and obtain their acceptance in the community.

"We do not like it if the people of Toledo ask what stand the Blade *is going to take on a sharp, given issue. If we have done our work well and made our principles known, they should know how we will apply them to any particular problem.*

"And so when a storm of controversy breaks out over the admission of Negroes to public housing projects where segregation had been practiced, it is not so much what we say then that will calm the furor and lead to the right solution. It is what we have been saying on our editorial page for years which counts when such a crisis comes."—MICHAEL BRADSHAW, editor, the Toledo *Blade,* in address at 1953 AEJ convention.