A readability formula for Vietnamese

Nguyen is assistant director of the National Center for Materials and Curriculum Development at the University of Iowa, Iowa City. He was formerly Vice-Minister of Education for the Republic of Vietnam. Henkin teaches in the College of Education and is involved in instructional design, materials development, linguistics, and studies of refugee transition and adaptation, also at the University of Iowa.

Liem Thanh Nguyen
Alan B. Henkin

International concern about the readability of prose has led to the development of a plethora of readability formulas over the last several decades (Klare, 1980). Most readability formulas were developed for use with English language materials. A relatively small number have been designed for languages other than English, and fewer have been practically applied.

Although Vietnamese writers and educators are concerned about the readability of printed materials (Le, 1950; Ky, 1961), no objective method to assess the readability level of Vietnamese prose has been available. A Vietnamese readability formula is of particular interest to writers, editors, educators, and social service providers who develop Vietnamese materials for populations of Vietnamese refugees in other countries.

Readability formulas are scientific procedures intended to provide quantitative estimates of the reading ease of written material (Klare, 1963). Most formulas are in the form of linear equations in which reading ease is a function of a weighted combination of measurable factors of a reading passage. These factors are the internal elements of the language assumed

to contribute most to the readability level of a passage. These internal elements include vocabulary difficulty measured in terms of word length, number of syllables, or word familiarity, and syntactical difficulty measured in terms of sentence length. Scores obtained through the use of readability formulas usually are expressed as corresponding to grade levels (Harris and Jacobsen, 1979) or reading levels of the passage.

Developing a formula for Vietnamese

To develop a readability formula, one should measure certain factors: (1) the readability level (RL) of a passage judged by experts or based on test results, (2) vocabulary or word difficulty (WD), (3) average sentence length, and (4) average word length. The obtained measures may be used to establish the formula by means of multiple regression techniques. In the development of this readability formula for Vietnamese, these were the procedures followed.

A sample of 20 passages of approximately 300 words each was selected from Vietnamese novels, magazines, and textbooks from grades 4 to college. The judged readability level (RL), the percentage of difficult words in the passage (WD), the average sentence length (SL), and the average word length (WL) were computed using the following procedures.

1. Judged readability level (RL). Passages were reviewed by 10 former Vietnamese teachers with specializations in Vietnamese language arts. Four were elementary school teachers, five were secondary school teachers, and one was a college instructor. All were graduates of either the Saigon Normal School or the Faculty of Pedagogy of Saigon University. All had five or more years of teaching experience. They all held certification in Vietnamese philology, Vietnamese literature, or Sino-Vietnamese literature.

The teachers read the passages carefully and rated them according to the grade levels of the Vietnamese educational system (Nguyen and Henkin, 1981a) as follows: elementary—grades 1-5; secondary (first cycle)—grades 6-9; secondary (second cycle)—grades 10-12; college—first year, 13, second year, 14, third year and above, 15. There were few differences in the scores of the raters on most passages. A difference of 3 points on one passage was the largest found. The scores were averaged to obtain the RL scores of the passages. The mean RL of the 20 passages was 7.775 with a standard deviation of 3.971.

2. Percentage of difficult words in a passage (WD). Nguyen and Henkin (1981b) compiled a list of basic words in Vietnamese containing 1,355 high frequency words. This list constitutes a basic vocabulary for a Vietnamese student in third or fourth grade, and distinguishes common or easy words from difficult words. Difficult words (not on the list) in each passage were counted. The percentage of difficult words is rendered by the formula:

$$WD = \frac{\text{\# of difficult words}}{\text{total \# of words}}$$

Theoretically, the WD may range from 0 to 1, with 0 being easiest and 1 being most difficult.

3. Average sentence length (SL). The average sentence length of each passage was obtained by dividing the total number of sentences into the total number of words in that passage:

$$SL = \frac{\text{total \# of words}}{\text{total \# of sentences}}$$

Table 1
Correlations among variables (N = 20)

	Readability level	Word difficulty	Sentence length	Word length
Readability level	1.00			
Word difficulty	.678	1.00		
Sentence length	.661	.554	1.00	
Word length	.776	.856	.419	1.00

Table 2
Multiple regression equation

Variable	R	R^2	B	Beta
Word length	.776	.602	2.884	.832
Sentence length	.860	.739	0.214	.478
Word difficulty	.871	.759	-8.793	.299
Constant			-7.685	

Analysis of variance

Source	DF	SS	MS	F
Regression	3	127.244	42.415	16.759*
Residual	16	40.494	2.531	

*$p < .05$

4. Average word length (WL). Vietnamese is substantially monosyllabic, but some polysyllabic compound words are formed by combining monosyllables. For example, *bắc-sĩ* and *hồng-thập-tự*. Compound words are counted as one word.

Each word is composed of letters with or without tonal and word marks. For use with the formula, each tonal or word mark is counted as one letter. Thus, the word *an* has two letters, but the word *ăn* is considered as having three letters. For a compound word, then, count all letters, tonal marks, word marks, and hyphens.

Counting may be done conveniently by computer, using coded Vietnamese protocols or computerized Vietnamese (Nguyen and Henkin, 1981b).

The average word length of a passage is obtained by dividing the total number of words into the total number of letters in the passage:

$$WL = \frac{\text{total \# of letters}}{\text{total \# of words}}$$

Analysis of the formula

Statistical analysis began with a correlational study among variables. Table 1 contains the correlation coefficients. It shows a strong relationship between the dependent variable and WL (r = .776). The next independent variable that correlates highly

Table 3
Readability table

Average sentence length (number of words)	Average word length (number of letters and marks)									
	3.0	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9
5	1	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8
6	1.2	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3
7	1.4	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2
8	1.6	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4
9	1.8	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6
10	2	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8
11	2.2	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4
12	2.4	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2
13	2.6	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4
14	2.8	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6
15	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
16	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
17	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
18	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
19	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
20	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
21	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
22	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
23	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4
24	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6
25	5	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8
26	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8	7
27	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8	7	7.2
28	5.6	5.8	6	6.2	6.4	6.6	6.8	7	7.2	7.4
29	5.8	6	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6
30	6	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8
31	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8	8
32	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8	8	8.2
33	6.6	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4
34	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6
35	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8
36	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9
37	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2
38	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4
39	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6
40	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8

with RL is WD (r = .678). However, WL and WD also correlate highly (r = .856), and there is redundancy in terms of contribution to the variance of the criterion when both WL and WD enter the multiple regression equation. Thus, we decided to enter WL in the multiple regression prior to SL and WD.

The next step in the statistical analysis was to perform a series of multiple regression equations with RL as the criterion, and WL, SL, and WD as the predictors. The stepwise procedure was used in that analysis. The general model is: $RL = B_o + B_1WL + B_2SL + B_3WD + E$. The results of the found multiple regression equation are reported in Table 2. Thus, the final equation is RL = 2.88 WL + .21 SL - 8.79 WD - 7.69 (1).

The three predictors together account for about 76% of the variance of the criterion. The standard error of estimate is 1.59.

However, the partial F test for WD is not significant after WL and SL have been entered in the equation. Thus,

Table 3 (continued)
Readability table

Average sentence length (number of words)	Average word length (number of letters and marks)									
	4.0	4.1	4.2	4.3	4.4	4.5	4.6	4.7	4.8	4.9
5	3	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8
6	3.2	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5
7	3.4	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2
8	3.6	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4
9	3.8	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6
10	4	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8
11	4.2	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6
12	4.4	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2
13	4.6	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4
14	4.8	5	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6
15	5	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8
16	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8	7
17	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8	7	7.2
18	5.6	5.8	6	6.2	6.4	6.6	6.8	7	7.2	7.4
19	5.8	6	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6
20	6	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8
21	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8	8
22	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8	8	8.2
23	6.6	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4
24	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6
25	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8
26	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9
27	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2
28	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4
29	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6
30	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8
31	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8	10
32	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8	10	10.2
33	8.6	8.8	9	9.2	9.4	9.6	9.8	10	10.2	10.4
34	8.8	9	9.2	9.4	9.6	9.8	10	10.2	10.4	10.6
35	9	9.2	9.4	9.6	9.8	10	10.2	10.4	10.6	10.8
36	9.2	9.4	9.6	9.8	10	10.2	10.4	10.6	10.8	11
37	9.4	9.6	9.8	10	10.2	10.4	10.6	10.8	11	11.2
38	9.6	9.8	10	10.2	10.4	10.6	10.8	11	11.2	11.4
39	9.8	10	10.2	10.4	10.6	10.8	11	11.2	11.4	11.6
40	10	10.2	10.4	10.6	10.8	11	11.2	11.4	11.6	11.8

the following formula is adopted as the final form: RL = 2 WL + .2 SL - 6 (2).

The two predictors together account for about 74% of the variance of the criterion, and the standard error of estimate is 1.6. The predicted RL will yield the readability level of a piece of writing corresponding to the grade level of the reader. For example, passage 1 with 3.5 word length and 11.16 average sentence length is predicted as RL = 2(3.5) + .2(11.16) - 6 = 3.23, which is estimated to fall between grades 3 and 4.

How to use the formula

To compute the predicted readability level (RL), one must compute the average word length (WL) and the average sentence length (SL), then substitute the results for WL and SL in the formula RL = 2 WL + .2 SL - 6. The value of RL indicates the grade level of the passage.

We have developed the Readability Table and Scale (Tables 3 and 4) to facilitate the computation of RL and the interpretation of results. The Readability Table gives the values of

Table 3 (continued)
Readability table

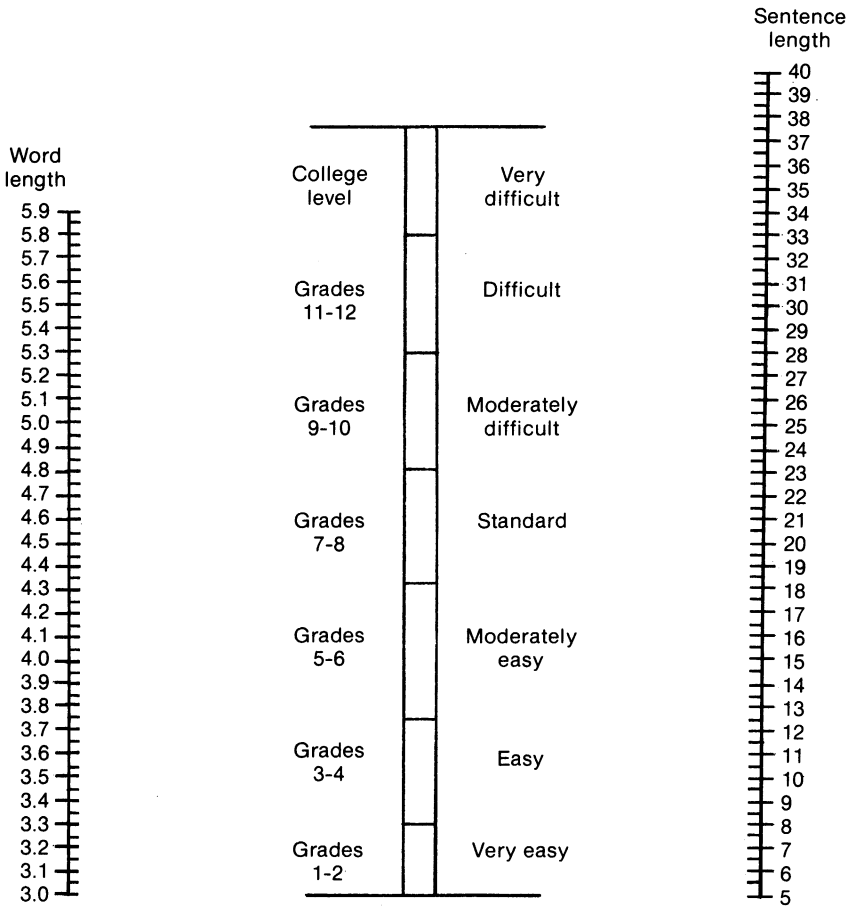
Average sentence length (number of words)	Average word length (number of letters and marks)									
	5.0	5.1	5.2	5.3	5.4	5.5	5.6	5.7	5.8	5.9
5	5	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8
6	5.2	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8	7
7	5.4	5.6	5.8	6	6.2	6.4	6.6	6.8	7	7.2
8	5.6	5.8	6	6.2	6.4	6.6	6.8	7	7.2	7.4
9	5.8	6	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6
10	6	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8
11	6.2	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8	8
12	6.4	6.6	6.8	7	7.2	7.4	7.6	7.8	8	8.2
13	6.6	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4
14	6.8	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6
15	7	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8
16	7.2	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9
17	7.4	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2
18	7.6	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4
19	7.8	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6
20	8	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8
21	8.2	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8	10
22	8.4	8.6	8.8	9	9.2	9.4	9.6	9.8	10	10.2
23	8.6	8.8	9	9.2	9.4	9.6	9.8	10	10.2	10.4
24	8.8	9	9.2	9.4	9.6	9.8	10	10.2	10.4	10.6
25	9	9.2	9.4	9.6	9.8	10	10.2	10.4	10.6	10.8
26	9.2	9.4	9.6	9.8	10	10.2	10.4	10.6	10.8	11
27	9.4	9.6	9.8	10	10.2	10.4	10.6	10.8	11	11.2
28	9.6	9.8	10	10.2	10.4	10.6	10.8	11	11.2	11.4
29	9.8	10	10.2	10.4	10.6	10.8	11	11.2	11.4	11.6
30	10	10.2	10.4	10.6	10.8	11	11.2	11.4	11.6	11.8
31	10.2	10.4	10.6	10.8	11	11.2	11.4	11.6	11.8	12
32	10.4	10.6	10.8	11	11.2	11.4	11.6	11.8	12	12.2
33	10.6	10.8	11	11.2	11.4	11.6	11.8	12	12.2	12.4
34	10.8	11	11.2	11.4	11.6	11.8	12	12.2	12.4	12.6
35	11	11.2	11.4	11.6	11.8	12	12.2	12.4	12.6	12.8
36	11.2	11.4	11.6	11.8	12	12.2	12.4	12.6	12.8	13
37	11.4	11.6	11.8	12	12.2	12.4	12.6	12.8	13	13.2
38	11.6	11.8	12	12.2	12.4	12.6	12.8	13	13.2	13.4
39	11.8	12	12.2	12.4	12.6	12.8	13	13.2	13.4	13.6
40	12	12.2	12.4	12.6	12.8	13	13.2	13.4	13.6	13.8

RL from 1 to 13.8 when WL varies from 3.0 to 5.9 and SL varies from 5 to 40, according to the prediction equation (RL = 2 WL + .2 SL - 6).

Values on the Readability Table (1-13.8) which correspond to grade 1 through the college level respectively are divided arbitrarily into seven categories with each category corresponding to two grades. Further, to facilitate teacher assessment of the readability level of a passage, the categories are labeled from easiest to most difficult, as follows:

Category 1, grades 1-2, RL 1.0-2.0, Very Easy;
Category 2, grades 3-4, RL 2.2-4.0, Easy;
Category 3, grades 5-6, RL 4.2-6.0, Moderately Easy;
Category 4, grades 7-8, RL 6.2-8.0, Standard;
Category 5, grades 9-10, RL 8.2-10.0, Moderately Difficult;
Category 6, grades 11-12, RL 10.2-12.0, Difficult;
Category 7, college level, RL 12.2, Very Difficult.

Table 4
Readability scale



These instructions will assist the user of the formula.

1. Selection of the passage. The first task is to select randomly a number of passages of approximately 100 words from the study material. Nine passages chosen at random from segments of a book or a long article, for example, will usually suffice. Start at the beginning of a sentence and stop at the end of a sentence.

2. Sentence count (total number of sentences). For each passage, count the number of sentences. Exclamation marks, periods, and colons are considered as marking the end of sentences.

3. Word count (total number of words). For each passage, count the words. Compound words should be considered as one word. For example, *học-sinh* is counted as 1, *hồng-thập-tự* is also counted as 1.

4. Letter count (total number of letters). For each passage, count the number of letters. Each tonal mark, word mark, and hyphen (in compound

words) should be counted as one letter. For example, *âu-trî* is counted as 9 (5 letters, 1 word mark ^, 2 tonal marks ´, and 1 hyphen) and *bệnh-viện* is counted as 13 (8 letters, 2 word marks ^^, 2 tonal marks ..., and 1 hyphen).

5. Computation of WL (average word length). For each passage, compute WL by dividing the total number of words into the total number of letters. Round the results to the nearest tenth.

6. Computation of SL (average sentence length). For each passage, compute the SL by dividing the total number of sentences into the total number of words in the passage. Round the results to the nearest one.

7. Finding and interpreting the RL (readability level). Use the values of WL and SL to find the value of RL (readability level) in the Readability Table (Table 3). The value of RL yields an estimation of the corresponding grade level of the passage. For example, if WL = 5.6 and SL = 15, first locate column 5.6 in the Readability Table, then read down the column to row 15. The given value of RL is 8.2. Thus, the grade level of the passage is between grades 8 and 9.

8. Readability Scale. To facilitate the practical interpretation of the readability level of a passage, use the Readability Scale (Table 4). Draw a line connecting the value of WL to the value of SL. The intersection of the line with the center line shows the readability level of the passage.

Approaches to future research

This Vietnamese readability formula will be the subject of future research focusing on its reliability and validity. Both the coefficient of stability and the coefficient of equivalence may be investigated with regard to the reliability of the formula. A sample of about 20 passages of several hundred words each will be selected from Vietnamese newspapers, magazines, and books in the course of such a study. Two Vietnamese teachers or educators will use the formula to assess the readability level of these passages. After a time interval of two weeks, the same sample of passages will be given to the teachers, who will once again assess the readability levels. Correlations between the results of the first and the second administrations, computed separately for each teacher, will provide the coefficient of stability. The correlation between the results provided by the teachers on one assessment will yield the coefficient of equivalence.

Study of formula validity may encounter difficulty if the researcher endeavors to correlate the results yielded by the formula with the scores on a reading comprehension test administered to Vietnamese students. Problems stem from the likelihood that reading capacity levels in Vietnamese may deteriorate as the process of acculturation accelerates; hence, the questionable utility of reading comprehension tests.

One alternative is to correlate formula scores and rating scores of experts or experienced teachers with specializations in Vietnamese. A sample of 20 to 30 passages should be sufficient for that purpose.

Formula validation will require repetitive studies over the long term. The intent is to incorporate results of such work in revisions and adjustments of the formula in a quest for optimal accuracy.

References

Harris, Albert J., and Milton D. Jacobsen. "A Framework for Readability Research: Moving beyond Herbert Spencer." *Journal of Reading*, vol. 22 (January 1979),

pp. 390-98.
Klare, George R. *A Manual for Readable Writing*. Glen Burnie, MD.: Rem Co., 1980.
Klare, George R. *The Measurement of Readability*. Ames, Iowa: Iowa State University Press, 1963.
Ky, Vu. *Nghe-Thuat Viet Van, Doc Van* (The Art of Writing and Reading). Saigon, Vietnam: Khai Tri, 1961.
Le, Nguyen Hein. *Nghe-Thuat Viet Van* (The Art of Writing). Saigon, Vietnam: Nguyen Hein Le publisher.
1950.
Nguyen, Liem T., and Alan B. Henkin. *Vietnamese Refugee Students: Legacies of an Educational Past*. Iowa City, Iowa: National Center for Materials and Curriculum Development, 1981a.
Nguyen, Liem T., and Alan B. Henkin. *Fundamental Vietnamese Vocabulary*. Iowa City, Iowa: National Center for Materials and Curriculum Development, 1981b.

College reading improvement journal

The College Reading Improvement Special Interest Group of the International Reading Association seeks papers for its refereed journal, *Forum for Reading*. Articles should relate to some aspect of reading improvement for college and other postsecondary students. Guidelines available from Rona Flippo, Coeditor, *Forum for Reading*, Georgia Department of Education, Division of Staff Development, Twin Towers East-Suite 1858, Atlanta, Georgia 30334, USA. For subscription (includes membership in the special interest group), send a check for US$8.00 made out to CRI/SIG of IRA, to Archie Davis, Illinois Central College, Route 8, 203 Fenestra Lane, East Peoria, Illinois 61611, USA.

New information service

The Far West Laboratory for Educational Research and Development announces a new service called Technology Information for Educators (TIE). The TIE staff answers questions on all aspects of educational technology, including computer literacy, instructional software, electronic record keeping, interactive video, satellite TV, hardware, and related topics.

For each question, TIE compiles a packet of up-to-date references, referrals, and materials by using computerized databases (such as ERIC and Microcomputer Index) as well as TIE's extensive collection of technology related journals, news articles, catalogs, reviews, and reports. The packet is sent within 72 hours after the TIE staff receives the question.

TIE can be used as a conventional mail and telephone service or as an electronic service. Using a computer terminal or microcomputer and a telephone modem, the user can transmit questions to TIE and receive TIS's answers electronically.

For further information, including fees, contact Jean Marzone, TIE Director, Far West Laboratory, 1855 Folsom Street, San Francisco, California 94103, USA, telephone (415) 565-3035.