TEACHING WRITING:

Methods, Materials, & Measurement

Department of English
University of Delaware

Volume 9, Number 1

The Prediction of Freshman Composition Grades at a Community College: A Correlational Study Based on a Noncomputational Readability Scale

Bill H. Lamb

One of the main problems community college composition instructors face is the diversity of writing ability among entering freshmen students. Research has attempted to deal with this problem on many occasions by the use of various tests that measure English achievement. These tests often include some designed specifically to measure English ability such as the Test for Standard Written English, but many times tests of a more general nature such as the ACT or SAT are used. The purpose of this study was to take a more unique approach to this problem by testing the effectiveness of a noncomputational readability scale applied to a sample Freshman student essay in order to predict success in English composition.

It was initially believed that if an appropriate index could be found to measure writing ability, then the measurement derived could differentiate skill levels between entering composition students. If the correlation between a readability index and the end course grade could be found to be significant, then those who in all likelihood would fail a standard composition course could be identified at the beginning of the semester and channeled into a remedial or developmental writing section. Thus, the prediction made prior to the second week of the semester could give some idea as to whether the student is likely to pass or fail English Composition I. This report will present the findings from this study designed to test the effectiveness of a noncomputational readability scale as the lone predictor for composition and course grades.

THE PILOT STUDY

The premise on which this research was based was identified through the use of a standardized readability scale applied to textbook materials in order to determine reading material suitable for Freshman level students. Initially it was hypothesized that if a readability scale can differentiate between published manuscripts, then the same concept could be applied to a student writing sample and very possibly differentiate levels of writing ability among students.

There are basically two types of readability scales in use today: computational scales and non-computational scales. Examples of computational scales would include the SMOG and Fry Readability scales; noncomputational scales include the SEER technique and the Rauding Scale of Prose Difficulty. The purpose of all is the same in that each will evaluate reading material and produce a grade level range representing how difficult the material is to read and at what grade level a particular piece of reading material may be assumed to be appropriate. Instructors, librarians, and publishers have used these scales for years to help aid in the identification of appropriate reading material for students. Obviously, an instructor at the grade school level

would not wish to require a textbook with a college reading level, nor would the reverse be likely to occur. By using these scales, teachers can select books appropriate for the students' abilities and promote more effective use of the materials for learning purposes. It was the intent of this research to attempt to use the same philosophy in the identification of students who do and do not have the writing ability suitable for a first-semester college composition course.

To test the effectiveness of a readability scale applied to a student writing sample, a pilot study was performed using the Edward Fry Readability Technique--a fairly well-known computational readability scale. In general terms, the Fry scale establishes grade levels for written material based on an average syllable count and an average sentence length per 100 written words. This is the standard method used by what are called computational readability scales. The assumption is made that the higher the syllable count, the higher the vocabulary: the fewer number of sentences per 100 written words, the more complex the sentence style.

As an example of how this method works, consider the following quotation randomly selected from Tom Wolfe's essay "The Right Stuff":

> In the Navy, in addition to the stages that Air Force trainees went through, the neophyte always had waiting for him, out in the ocean, a certain grim gray slab; namely, the deck of an aircraft carrier; and with it perhaps the most difficult routine in military flying, carrier landings. He was shown films about it, he heard lectures about it, and he knew that carrier landings were hazardous. He first practiced touching down on the shape of a flight deck painted on an airfield. He was instructed to touch down and gun right off. This was safe enough--the . . . (Wolfe, 20).

To establish the readability level for the selection, the rater would first count the number of syllables contained in the 100 words selected (the total from this selection being 141). The rater then counts the number of sentences contained in the 100 words. Partial sentences are turned into decimals representing approximately how much of the sentence is contained in the 100 words used. Therefore, if you have half a sentence included, it would count as .5. In this example, the last line represents the first five words in a twenty-five word sentence; therefore, it would be rated as .2 or one-fifth of the whole sentence. The total sentence count for this example would be 4.2.

The next step is to align the syllable count and the sentence count using Fry's prepared evaluation slide (available through Jamestown Publishers). The slide has the syllable count listed at the top and the sentence count perpendicular to this. An open box allows for the identification of the grade level once the slide has been moved to locate the appropriate syllable and sentence count. In this case, a total of 141 syllables matched to a sentence count of 4.2 yields a grade level of 9. An important note is that to insure reliability, the syllable and sentence count should be based on the average of three randomly selected

100 word samples in lieu of the one here used as an example of the procedure.

When this method was applied to sample student essays collected for the pilot study, the correlation between the Fry grade levels and the actual and course grade was r=.124 which was not significant for the sample tested. This indicated that in this group there was no correlation between the two variables. By observation, two flaws were discovered with this method: 1) The system could not identify sentence flaws such as fragments or run-ons. Thus, a student could incorrectly fuse together several sentences and still receive a high rating. 2) Sentences that included several multiple syllable words would also rate high even though the words might be awkward, misspelled, or misused in context. For example, a 100 word selection such as:

> I gone to community college cause community college is very good for someone like me cause I was brung up in a bery small community that woudn't have a community college and if it had a community college that wood of been a little different to me. This is real cool and all i play games and stuff without much hep and stuff cause when i have all ways had a kinda special talent for being good without my hands that makes me specey in some ways. This is true. Because when I think I can fund real great.

By rating this 100 word selection using the Fry method, one would have to use 135 as the syllable count and 4.0 for sentence number which would yield a grade level of 9. Thus, both this selection and the sample taken from Wolfe would be identified as the same grade level, and obviously the above sample cannot be considered equal to published material with flawless grammar and sentence structure. Because of this finding, a noncomputational readability sacle was identified as a better testing variable, and the study was revamped to apply the Rauding Scale of Prose Difficulty (Carver, 1974) in place of the Fry Readability Scale.

A noncomputational readability scale compares a piece of writing with an unknown readability level to a piece with an established grade level. The analysis compares the sentence writing style, vocabulary, punctuation, and content. The advantage of this measure applied to a writing sample is that the rater can allow for such flaws as run-together sentences or faulty vocabulary usage which might rate high on a computational scale. Carver's Rauding Scale was selected because of its grade level range (1 to 18) and its demonstrated validity and reliability with published materials (Carver, 1975-76, Duffelmeyer, 1982).

THE METHOD

To test the effectiveness of this noncomputational readability scale applied to a student writing sample, a study was conducted during the 1982 fall semester at a small, midwestern community college. The sample population consisted of 224 entering Freshmen composition students who were randomly divided into two groups (Group 1 = 110 subjects and Group 2 = 114 subjects). Each group was given an essay topic on which to write a 300 to 500 word

theme during a 30 minute class session. The writing topics were selected from those previously used at the institution, and the testing sections were assigned topics randomly so that section topics would not be carried over. Two sample topics used for this experiment included: "What are three main goals you now have that you would like to accomplish in the future?" and "What are three main reasons you decided to attend college?" In all likelihood, these sound very similar to most English writing instructors. The intention of asking for "three reasons" was to suggest both organization and development to the writer, and the subject area was really not significant to the study. Although the student was not told such, the ideal product would be a standard five-paragraph essay with an introduction, three body points, and a conclusion, all grammatically correct with proper development and centered around a central thesis statement. By not supplying this insight to the writer, it was easier to identify students who understood paragraphing and structure from those who might choose to respond with only a few combined sentences. The only key given to the writers through the instructions was that they should write in sentence rather than outline form.

The writing samples were collected by trained instructors on the second day of class and all necessary materials were supplied to the student. The training of those instructors administering the writing sample involved two meetings with the researcher. The first meeting involved an explanation of the purpose for the research, the need for careful administration of the survey, and the selection of sample writing topics. The second meeting was called the day prior to the first administration of the writing sample in order to stress precautions that must be followed to insure an accurate sample and to answer any questions the instructors might have. Each instructor was given a packet containing the necessary number of writing samples for each section, pencils, paper, and a list of instructions to be followed on the testing day (see figure 1). The most important caution to the administrators of the survey was to make sure they would in no way help, aid, or influence the student's own writing. This interference could negatively affect the outcome by causing the writing sample to not be a true representation of the student's knowledge and ability.

Once the writing samples were collected, three raters were selected based upon their successful completion of the Rauding Scale Qualification Test (Carver, 1974). The Qualification test was developed by Carver to better identify those who could differentiate between the grade level anchor passages and the piece of material to be rated. In his research, Carver found that some individuals were better at making this distinction than others. As he reports, there are no standard characteristics that can describe these individuals, but simply some seem to be more accurate than others. The test is used to select these individuals by asking them to rate sample reading passages using the Rauding Scale readability technique anchor passages. (These materials are all available through Dr. Carver.) In this research, ten individuals with quite diverse backgrounds were tested and four passed the qualification test with the necessary 100% proficiency. Of these four, three were selected to participate in the

experiment because of their strong interest and desire.

Each rater then rated the writing samples from Group 1 and identified a readability grade level index for each sample using the Rauding Scale anchor passages (Carver, 1974). The instrument was developed by Dr. Ronald Carver and through his research, the instrument was shown to be both valid and reliable. (See Carver, 1975-76 and Duffelmeyer, 1982).

THE PROCEDURE FOR RATING SAMPLES

The rating procedure followed exactly the Rauding Scale Readability Technique. As is described by Carver, "The Rauding Scale consists of an anchor set of 6 passages ostensibly representing grades 2, 5, 8, 11, 14, and 17. A qualified expert reads the passage to be rated and then decides where it fits on the Rauding scale, using the 6 passages as anchor passages. For example, if a qualified expert decided that a passage was much more difficult than the grade 5 anchor passage, but a little less difficult than the grade 8 anchor passage, the expert might assign the passage a grade 7 difficulty rating" (Carver, 1975-76). When rating writing samples, the same basic procedure is followed except in lieu of identifying the passage 'difficulty', the rater is more interested in the sample 'quality'. In other words, if the writing sample contains good content, organization, and a complex sentence style similar to anchor passage 8, but also have some faulty punctuation and several misspellings, the rater might select a grade level 7 for the sample. Obviously, the extreme ratings would apply to the paper that is well written, free of errors and neatly organized (the 15 to 17 range) with content being the deciding factor, or the paper that is full of mechanical errors and has little, if any, content (the 1 to 3 range). For the most part, papers will fall between the 4 to 14 range and receive their rating based primarily on word usage, sentence style, mechanics and content as they would correspond to the same characteristics present in the various anchor passages.

Each writing sample was given a three digit number to replace the student's name and any other information that might inadvertently describe the writer to the rater. Using a standard table of random numbers, each sample was placed in either Group 1 or Group 2. The first rater was then given Group 1 samples to rate using the Rauding Scale anchor passages. In addition, they were given a separate sheet of paper on which to record the paper number and subsequent grade level. It was important that they make no markings on the paper that might influence later raters.

After each rater finished a group of papers, the group was shuffled to avoid ordering effect--the possibility that each rater would read the same paper at the same position in the stack as had a previous rater. When all raters had finished rating Group 1 (a task which took approximately 90 to 120 minutes, the raters classified Group 2 following the same procedure. A small time lapse occurred to insure that the raters were not fatigued or less enthusiastic about the second group.

Once all the surveys were rated, the researcher recorded all three ratings on a separate sheet according to the student number and group. The level numbers were totaled and an average rating

for each paper was derived by dividing by 3. In most cases, the levels were all within one number of each other, which was also found to be true in Carver's research. However, in a few instances, a range of several grades would occur. This did not damage the research but simply reflected the need for three raters and the use of their average rather than individual rating. In these cases, it appeared on post evaluation that one rater may have judged the content of the paper above the grammatical errors; whereas, another rater would rate a paper lower because of weak mechanics. If only one rater would have been used, these error scores could have drastically affected the outcome of the study. This would seem to correspond with the idea that when instructors read through sample essays with the idea of learning how good the writers are, they tend to judge different characteristics in different ways, giving different weights to these various areas. By using this technique, the ratings become more standardized and one might be sure of a more comprehensive, accurate, and less biased evaluation.

After the mean average grade level was calculated for each sample, the mean figure was translated into whole grade levels, as is suggested by Carver, using his adjusted grade level table (Carver, 1974). The adjusted grade level table changes any fraction number (such as an average of 5.5) to a whole number such as 6. As was identified through the post evaluation of this study, there is very little need to make this transition when dealing with writing samples. With the availability of a computer, the calculations of fractions as opposed to whole numbers is not a significant problem. The second group ratings were then recorded and the same procedure followed.

Grades in Written Communications I for each student were collected at the conclusion of the semester. Group subjects who withdrew during the semester or received incompletes were withdrawn from the study as there was no other way to ascertain their true course grade. (This translated into a loss of five subjects in Group 1 and twelve in Group 2.) A Pearson Product Moment Correlation was then tabulated for the Rauding Scale Grade Level and the actual Written Communications I end course grade for Group 1. The correlation was shown to be r=.467 which was significant beyond the .001 level of confidence.

As the correlation was significant, a linear regression analysis was performed which generated the following prediction formula: $Y' = .20X + 1.031$. (See Table 1 for statistics related to Group 1). Group 2 was then analyzed and adjusted mean grade levels were applied to the prediction formula for cross validation purposes. Predicted grades were then correlated to the student's actual end course grade. The correlation was found to be r=.442 which was also significant beyond the .001 level. (See Table 2 for statistics related to Group 2). This cross validation indicated the degree of relationship or correlation between the predicted grade and the actual end course grade.

The purpose of cross validation is to allow the researcher to check to see how well the predicted grade corresponds to the actual grade the student received. As will be discussed further, predicting specific grades in a specific course is extremely

difficult was some "A" level writers may for reasons other than their writing ability receive a lower grade; similarly, some "C" level writers may work hard and receive a grade more representative of their growth during the semester than their actual writing ability. Therefore, a cross validation simply shows the researcher how well the prediction formula worked with a separate sample.

THE CONCLUSION

The significant correlation found to exist in Group 1 demonstrated that the technique used accounted for roughly 22% of the variance in student Written Communications I grades. When comparing this technique to other methods used for prediction purposes in composition, the results would indicate that the method is viable for screening students. The significance of the correlations and the amount of variance in course grades covered by this method surpasses many studies dealing with current tests used to identify students in need of remediation. The results are also comparable to studies using ACT scores, SAT scores, or high school grades as predictors.

Because of the high probability for error in the prediction formula, it was decided to use the Rauding Scale level 5 as the cutoff point for identification for remediation. This level included predicted grades of both D and F. When this level was applied to Group 2 subjects, it was found that roughly 75% of the students were correctly predicted to either pass or fail the course. In terms of error predictions, roughly 20% passed the course although they were predicted to fail, and only 5% failed who were predicted to pass.

This 5% figure is really quite low when considering the fact that many composition students fail because of reasons outside their particular writing ability. Some fail because they do not complete assignments, attend class or for a variety of other reasons. Likewise, the 20% who passed the class could have had a problem with the writing sample, worked much harder than average during the semester, or for several other reasons achieved a better grade than would have been expected.

It is important to note that there will always be these exceptions to the rule and some students will fail no matter how accurate their prognosis for success might be. Also, some students who were mis-rated could still be channeled into a regular composition class should the remedial instructor feel the student is in fact capable of that level of writing. The importance of this research is contained in the 75% who were correctly identified and some of those identified to be in need of remediation could very well benefit from some special, developmental writing instruction instead of simply being allowed to fail at the standard composition course level. In the past, these were the students who were in all likelihood cheated by being placed in a standard writing course where they lacked the basic ability necessary to achieve success.

Although significant, the cross validation correlation was low which can be attributed to the high probability for error in predicting specific course grades. When used on a Pass/Fail basis however, a more accurate prediction was evident. Pass/Fail

offers a more general prediction and will still identify the students in need of remediation as well as the actual prediction of the specific end course grade. (See Tables 3 and 4) The importance of this research is really to differentiate between those who might fail their first college English course rather than those who might receive a grade of A, B, or C.

Even though the method may prove effective for other institutions and should be considered universal in application, the formulas should be considered valid for only the institution under study. Any prediction formula will be influenced in some way by the population being tested. Obviously, students who choose to attend a small, midwestern community college are very different from students who may live on the East or West coast. However, once the prediction formula has been established for a particular institution, the formula will tend to remain valid for an extended period of time. Equally, through consistent cross validation testing and evaluation, the formula may be easily updated to include each new class of entering students prior to the next testing period. This updating will also continue to improve the accuracy of the predictions by refining the formula to better meet the changing needs of the student population.

The true value of this procedure can only be found through careful replication at a number of institutions. Nevertheless, this method does seem to hold several possibilities for future research in identifying English students in need of remediation.

Coffeyville Community College, Coffeyville, Kansas

Works Cited

ACT Assessment. Assessing Students on the Way to College: Vol. I & II. Iowa City, Iowa: The American Testing Program, 1972.

Aleumoni, L. M. and L. Oboler. "ACT vs. SAT in Predicting First Semester GPA." Educational and Psychological Measurement, (1978): 393-9.

Bailey, Roger. "The Test of Standard Written English: Another Look." Measurement and Evaluation in Guidance, 10(July 1977): 70-74.

Bloom, Benjamin and Frank Peters. The Use of Academic Prediction Scales for Counseling and Selecting College Entrants. New York: The Free Press of Glencoe, Inc., 1961.

Carver, Ronald P. Manual for the Rauding Scale Qualification Test. Kansas City, Mo.: Revrac Publications, 1974.

Carver, Ronald P. "Measuring Prose Difficulty Using the Rauding Scale." Reading Research Quarterly, 11(1975-76): 660-85.

Duffelmeyer, Frederick A. "A Comparison of Two Noncomputational Readability Techniques." The Reading Teacher, October, 1982: 4-7.

Fry, Edward B. Fry Readability Scale. Providence RI: Jamestown Publishers, 1978.

Hoyt, Donald P. "College Grades and Adult Achievement: A Review of Research." Educational Record, (Winter, 1966): 70-75.

Michael, William B. and Phyllis Shaffer. "A Comparison of the Validity of the Test of Standard Written English (TSWE) and

of the California State University and Colleges English Placement Test (CSUC-EPT) in the Prediction of Grades in Basic English Composition Courses and of Overall Freshman Year Grade Point Average." Educational and Psychological Measurement, 39(1979): 131-45.

Wolfe, Tom. The Right Stuff. New York: Bantam Books, 1979.

FIGURE 1

WRITING SAMPLE EXAMINER INSTRUCTIONS:

1. Hand out a copy of the survey to each student.
2. Place extra sheets of notebook paper on a desk and direct students to the location.
3. To the Class:
   A) "Read the instructions carefully, but do not begin writing until I tell you to begin."
   B) When everyone has finished reading the instructions: "You may begin writing."
   C) Write the starting time on the board, add 30 minutes, and write the ending time on the board.
   D) After the time is finished, call "Stop" and collect the papers.
4. Cautions:
   A) Do not read the instructions to the class.
   B) Do not answer questions pertaining to the question or the purpose of the sample--you may answer procedure questions, but do so only in a direct manner.
   C) Do not help a student get started.
   D) Do not supply grammatical information--i.e., spelling, punctuation, etc.
   E) Do not talk to students after the time to write has begun.
   F) If a student does ask a question pertaining to the survey, you are to answer: "Do the best you can."

THANK YOU FOR YOUR HELP AND CAREFUL ADMINISTRATION

TABLE 1

THE DESCRIPTIVE STATISTICS FOR MEASUREMENT X AND Y

Group 1

	Number N	Mean	Sum of Squares SS	Variance S^2	Standard S
RAUDING SCALE (X)	105	6.371	749.008	7.203	2.684
END COURSE GRADE (Y)	105	2.305	138.132	1.328	1.152

Pearson's Product Moment Correlation	Coefficient of Determination	Coefficient of Non-Determination	Standard Error of the Estimate
.467	.218 (21.8%)	.782 (78.2%)	1.024

TABLE 2

DESCRIPTIVE STATISTICS FOR THE MEASUREMENT OF THE CORRELATION CORRELATION BETWEEN PREDICTED GRADES AND ACTUAL COURSE GRADES

Group 2

	Number N	Mean	Sum of Squares SS	Variance S^2	Standard Deviation	Correlation r
PREDICTED GRADE (P)	102	1.814	31.318	.31	.557	
END COURSE GRADE (A)	102	2.225	140.036	1.386	1.177	.442

TABLE 3

SUCCESS AND FAILURE PREDICTIONS

	Predicted A, B, or C	Predicted D or F
Earned A, B, or C	67	19
Earned D or F	8	8

TABLE 4

PREDICTED GRADES

Number of Predictions that were EXACT	Number of Predictions that were OFF 1 GRADE	Number of Predictions that were OFF 2 GRADES	Number of Predictions that were OFF 3 GRADES
34	48	19	1
(33%)	(47%)	(19%)	(1%)