DOCUMENT RESUME

ED 377 464 CS 011 930

AUTHOR Smith, Dean R.
TITLE A Reexamination of Semantic Difficulty: Validating the "Living Word Vocabulary."
PUB DATE Apr 93
NOTE 8p.; Paper presented at the Annual Meeting of the American Educational Research Association (74th, Atlanta, GA, April 12-16, 1993).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Elementary Secondary Education; Readability; Reading Research; Regression (Statistics); Semantics; Test Use; *Test Validity; *Vocabulary Development
IDENTIFIERS *Living Word Vocabulary

ABSTRACT

Four studies examined the validity of the "Living Word Vocabulary" (LWV), a corpus of approximately 44,000 alphabetized words with multiple meanings tested at different grade levels. Regressions were performed between the grade level p-values (percentages of students responding correctly to a vocabulary test item) reported by LWV and word frequency, grade-level p-values obtained from three nationally standardized tests, logit difficulties obtained from the Peabody Picture Vocabulary Test (PPVT), and observed difficulties obtained from two sets of words. Results indicated that the LWV seemed to be a valid measure of semantic difficulty. A regression analysis between the LWV and the PPVT produced an equation that allows users of the LWV to place all of the word difficulties upon a common scale, thus allowing cross-grade-level comparisons of the same word. This equation now allows those interested in readability access to the huge corpus of data in the LWV. (Contains 10 references and 5 tables of data.) (RS)

A Reexamination of Semantic Difficulty: Validating the *Living Word Vocabulary*

Presentation at the 1993 American Education Research Association Conference in Atlanta, Georgia

Dean R. Smith
Department of English
Shippensburg University
Shippensburg, PA 17257

Abstract: The *Living Word Vocabulary* is a corpus of approximately 44,000 alphabetized words with multiple meanings tested at different grade levels. Four studies were performed to validate the *Living Word Vocabulary*. Regressions were performed between the grade level p-values reported by this corpus and word frequency, grade level p-values obtained from three nationally standardized tests, logit difficulties obtained from the *Peabody Picture Vocabulary Test* (*PPVT*), and observed difficulties obtained from two sets of words. Raw correlations ranged from .768 to .844 without being corrected for measurement error or range restriction. Perhaps a more important result comes from regressing the *Living Word Vocabulary* (*LWV*) and the logit difficulties from the *PPVT*. The *LWV* word difficulties are reported by grade level and p-value or the percentage of students responding correctly to a vocabulary test item. However, the manner in which word difficulty is reported prevents cross grade comparisons from being made because each word difficulty is locked into a single grade level interpretation. For example, the word bed as used to indicate a part of a pickup truck is known by 70% of all eighth graders tested, but there is no way of interpreting how many sixth graders or fourth graders know this use of the word bed. The regression analysis between the *LWV* and the *PPVT* produces an equation that allows users of the *Living Word Vocabulary* to place all of the word difficulties upon a common scale thus allowing cross grade level comparisons of the same word.

A Reexamination of Semantic Difficulty: Validating the *Living Word Vocabulary*

Traditionally, readability formulas have used average word length or syllable counts to measure vocabulary difficulty. In hopes of finding a better measure of vocabulary difficulty, Stenner, Smith, and Burdick (1983) analyzed over fifty semantic variables which may have contributed to the difficulty of the vocabulary items on the *Peabody Picture Vocabulary Test-Revised* (*PPVT*) *Forms L and M* (Dunn and Dunn, 1981). Variables included in the study were part of speech, number of letters, the number of syllables, the modal grade at which the word appeared in school materials, content classification of the word, the frequency of the word from two different word counts, and numerous algebraic transformations of these measures. Correlations were then run between the logit difficulties of the test items and each targeted variable. The best operationalization of vocabulary difficulty was found to be the log of the word frequencies obtained from the American Heritage corpus (Carroll, Davies, and Richman, 1971). The use of word frequency as a measure of semantic difficulty seems logical since the more often a word appears in text, the more likely readers will encounter it and construct its meaning.

However, word frequency, as well as word length and syllable counts, does not take into account the variety of definitions associated with a particular word. The word *bed* has the same number of letters and syllables and word frequency count regardless of whether it is being used in a passage to denote something in which we sleep or something in which we plant flowers or the back of a pickup truck. Therefore, each of these estimates will assign the word *bed* the same measured difficulty. However, common sense reveals that most second grade students will recognize the use of *bed* to denote a place of rest, but not necessarily recognize the word as a place to put flowers or 2x4's.

In response to this limitation, Dale and O'Rourke (1981) developed the *Living Word Vocabulary* (*LWV*), a corpus of approximately 44,000 alphabetized words with multiple meanings tested at different grade levels. Word difficulty is reported by grade level and p value or the percentage of students responding correctly to a vocabulary test item. A high p value indicates that a large number of students recognized the word, and therefore, is easier than a word with a lower p value. In order to test the validity of the *LWV* corpus, four different studies were performed.

Study 1: *Living Word Vocabulary* and Word Frequency Counts

An initial test of the content validity of the *LWV* involved a simple correlation between LWV difficulties and the log of the word frequency counts obtrained from the *Word Frequency Book* (Carroll, Davies, and Richman, 1971). The words studied were taken from the *Peabody Picture Vocabulary Test Forms L and M* (Dunn and Dunn, 1981) The resulting correlation between the *LWV* difficulties and the respective word frequencies was an r = .768 (see Table 1).

Table 1

Correlation between the Living Word Vocabulary Difficulties and Word Frequency Counts from the Word Frequency Book

	Mean	SD	Kur	Skew	Range
LWV logit difficulties	.06	2.41	.92	.98	13.41
Log of word frequency counts	1.76	.83	-.58	-.08	4.02

In order to establish the power of the *LWV* corpus in measuring vocabulary difficulty, other validations were needed that go beyond a comparison with word frequency. Three other validity tests of the *LWV* corpus were performed, one of which resulted in an equation that allows researchers to convert the *LWV* data into a more usable format. Because the difficulty of words in the *LWV* corpus is reported as grade level p-values, cross grade comparisons cannot be made. For example, the word bed as used to indicate a part of a pickup truck is known by 70% of all eighth graders tested, but how many sixth graders or fourth graders know this use of the word bed?

Secondly, the *LWV* data cannot be used to measure passage difficulty because of the way it is reported. P-values should not be averaged since they are not on an interval scale. This is unfortunate because the *LWV* corpus could serve as the semantic measure of a powerful readability formula. In fact, Fry (1990) attempted to tap the *LWV* for measuring the readablity of short passages since the impact of vocabulary difficulty on short passages would require a more sensitive instrument than is currently available. However, Fry disregarded much of the power of the corpus by focusing only on the grade level at which a word was tested. He ignored the p value, and opted to average only the grade levels of the words within the passage because he could not average the *LWV* difficulties as they were reported.

One way of making better use of the *LWV* word difficulties would be to place the data reported on a common interval scale. The logit scale which is the basis of the Rasch model has this capability (Wright and Stone, 1979). A logit difficulty is merely a log transformation of a p-value. The advantage of working with logits is twofold. One, the transformation removes the curvalinearity found in percentile scale. Two, it places all of the items on a common scale so that it is possible to compare ite is administered to a group of fourth graders to items adminstered to a group of eighth graders.

Study 2: *Living Word Vocabulary* and the *Peabody Picture Vocabulary Test*

In order to place the difficulties from the *LWV* on a common scale, a regression analysis was performed using the *Peabody Picture Vocabulary Test* (*PPVT*) *Forms L and M* (Dunn and Dunn, 1981). By comparing the logit difficulties of the 350 *PPVT* items to the *LWV* difficulties, two important things can be accomplished. One, a high correlation would provide concurrent validation of the *LWV*. Second, the grade level/p-value scores from the *LWV* can be translated into logit difficulties using the regression equation which results from running the statistical correlation. This formula can be used to convert all of the *LWV* grade-level p-values to logits thus allowing comparisons to be made across different grade levels.

The regression analysis between the *LWV* and the *PPVT* (see Table 2) produced a correlation of .842 (n = 348). This correlation is significantly higher than a similar analysis between the PPVT test item difficulties and the log of word frequency counts obtained from the *Word Frequency Book* (Carroll, Davies, and Richman, 1971) which produced an r = .772 (n = 331).

Table 2

Means and Standard Deviations for PPVT Logit Difficulties and the LWV Logits

	Mean	SD	Kur	Skew	Range
PPVT logit difficulties	.15	2.84	-1.03	-.08	11.89
Logit transformations of grade level p values	.15	2.39	.04	.85	11.02

More importantly however, the regression equation produced allows the grade level p-values from the *LWV* to be computed into logits. The formula follows with G equal to the grade level and P equal to the p-value:

$$G(.49) - [lg(P/(1-P)](3.59) - 1.03$$

By using this equation, we can convert the p-value of a given word for grades other than the targeted grade. For example, bed as used to reference the part of a pickup truck is known by 70% (or a p value of .70) of all 8th graders. If we wanted to know how many 7th graders know the word, merely plug the variables into the equation and solve for P. The results of this analysis for 7th grade as well as other grades is found in Table 3:

Table 3

P-value Estimates for Multiple Grades for the Word Bed as Defined in the LWV as a Part of a Truck

Grade	P-value
10	.81
9	.76
8	.70
7	.63
6	.55
5	.47

This formula can be used to either place all of the words in the *LWV* on a logit scale, or can be used to compute p-value difficulties for grade levels other than the originally tested grade, a procedure also advocated by Gershon (1991). This should make the *LWV* corpus accessible for the development of more sensitive readability formulas such as Fry's (1990). This conversion formula was also used in two other studies designed to test the validity of the *Living Word Vocabulary* difficulties.

Study 3: *I iving Word Vocabulary* and Standardized Vocabulary Tests

Concurrent validity may also be established by correlating the *Living Word Vocabulary* difficulties with item difficulties from nationally standardized tests. Difficulties were obtained for the vocabulary items found on the *Stanford Achievement Test Form J* (Psychological Corporation, 1985), the *California Achievement Test Form E* (CTB/McGraw-Hill, 1987) and the *Gates-MacGinitie Forms 1 and 2* (Riverside, 1978). Separate regression analyses were performed, one for each test, in which the *LWV* difficulty was correlated with the standardized test difficulty for all vocabulary items. The item difficulties for each of the standardized tests were reported in grade level p-values. The formula used to convert *LWV* grade level p-values was used to convert the standardized test p-values to logits. The results for each regression analysis are reported in Table 4.

Table 4

Means, Standard Deviations, and Correlations between the *Living Word Vocabulary* Logit Difficulties and Item Difficulties from Three Major Standardardized Tests

	Mean	SD	n	r
LWV logits	.37	1.38	270	.844
California Achievement Test	-.39	1.63		
LWV logits	-.83	2.95	322	.840
Stanford Achievement Test	-.44	2.19		
LWV logits	-.61	1.93	563	.828
Gates-MacGinitie Test	-.14	1.83		

Study 4: *Living Word Vocabulary* and Student Performance on Sampled Vocabulary Words

The fourth study involved testing the predictive validity of the *LWV*. Fifty vocabulary words with fourth grade difficulties and fifty more vocabulary words with sixth grade difficulties were randomly sampled from the *Living Word Vocabulary*. Test items were then developed for each of these words. Each test item consisted of the target word and five single-word foils. Each of the foils was checked against the target word to make certain that they were easier than the word being tested (see Figure A).

Figure A: Sample Test Item for Grades 3 and 4 of Validation Study 4

1. small
   a. happy
   b. little
   c. nice
   d. pretty
   e. sun

These words were then administered to a small sample of students in heterogeneously mixed classrooms in a rural elementary school. The fourth grade words were tested with 3rd and 4th grade students and the sixth grade words were tested with fifth and sixth grade students. It was important to test how well the *LWV* grade-level difficulties could predict the performance of students at the same grade used to develop the corpus. However, it was also important to see how well the *LWV* could predict the performance of students from grades different than those used to develop the corpus. Hence, two data runs were made. One correlation compares the 4th and 6th grade *LWV* word difficulties with the observed data collected from 4th and 6th grade students. This is called the on-grade analysis. The second analysis compares how well the observed data collected from 3rd and 5th grade students compare with the 4th and 6th grade word difficulties. This is called the off-grade analysis.

After the tests were administered, the observed difficulties were then calculated as logits and regressed against the difficulties reported by the *Living Word Vocabulary*. Before performing the regression analysis, the grade level p-values reported in the *LWV* were converted to logits as well using the formula obtained from Study 2. For the on-grade analysis, r = .790; for the off-grade analysis, r = .776 (Table 5).

Table 5

Correlation between LWV Logit Difficulties and Observed Difficulties for 100 Vocabulary Words

	Mean	SD	Kur	Skew	Range
Observed logit difficulties for grades 4/6	-1.90	1.97	-1.12	-.35	6.45
Observed logit difficulties for grades 3/5	-1.30	1.72	-.71	-.65	5.85
LWV word logits	-1.12	1.10	-1.06	-.39	4.27

It might be noted that the correlations are relatively high given the small sample sizes (3rd grade = 24; 4th grade = 43; 5th grade = 23; and 6th grade = 18). However, the size of the student samples remains a concern. A caveat must be noted in that the students used in the study were not randomly sampled. Instead, entire classes were used where teacher volunteers could be found to make time for the test. It should be further noted that because entire classes were used and because the classes were heterogeneous, ability variance within classes ranged from above grade level readers to students who were being mainstreamed and who had been classified with learning disabilities. Perhaps it could be argued that the sample population in this study more realistically reflects the norm found in the public schools. Finally, it must be noted that the correlations obtained in Study 4 are most likely deflated due to range restriction. Had data been collected that more adequately reflects the variance found through the entire range of grades (K-12), the correlations would be dramatically higher (Smith, Stenner, Horabin, and Smith, 1989; Thorndike, 1949).

Conclusion:

The regression analyses performed in these studies produced correlations ranging from .768 to .844. Based on these findings, the *Living Word Vocabulary* seems to be a valid measure of semantic difficulty. It provides a better measure of word difficulty than does word frequency. It also is more functional and accurate since different difficulties are available for different uses of the same word. Perhaps the most important finding of these studies is related to the development of an equation that can be used to convert the grade level p-values reported by the *LWV* corpus to logits, the measurement units based upon the Rasch model. The logit conversion allows for cross grade level comparisons to be made for the same word, whereas the *LWV* data, in its original format, was locked into a single grade interpretation of difficulty. This now allows those interested in readability access to the huge corpus of data collected by Dale and O'Rourke (1981).

References:

CTB/McGraw-Hill (1987). *California Achievement Test Form E*. Monterey, CA: McGraw-Hill.

Dale, E. and O'Rourke, J. (1981). *The living word vocabulary*. Chicago: World Book.

Dunn, L. M. and Dunn, L. M. (1981). *Peabody Picture Vocabulary Test Revised, Forms L and M*. Circle Pines, MN: American Guidance Service.

Fry, E. (1990). A readability formula for short passages. *Journal of Reading*, 33, 594 597.

Gershon, R. C. (1991). LWV to JOCRS conversion formula (Research memorandum 1991-1). Chicago: Johnson O'Connor Research Foundation.

Psychological Corporation (1985). *Stanford Achievement Test Form J*. Orlando, FL: Harcourt, Brace, and Jovanovich.

Riverside (1978). *Gates-MacGinitie Forms 1 and 2*. Boston, MA: Houghton-Mifflin.

Smith, D., Stenner, A, Horabin, I., and Smith, M. (1989). The Lexile framework: Final report for NIH grant HD-19448. ERIC Document Reproduction No. ED 307 577.

Thorndike, R. L. (1949). *Personnel selection*. New York: Wiley.

Wright, B. D. and Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.