DOCUMENT RESUME

ED 139 281 FL 008 587

AUTHOR	Stansfield, Charles
TITLE	Dictation as a Measure of Spanish Language Proficiency.
PUB DATE	Apr 77
NOTE	16p.; Paper presented at the annual convention of the Teachers of English to Speakers of Other Languages (Miami, Florida, April 27, 1977)
EDRS PRICE	MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS	Achievement Tests; College Students; *Language Instruction; *Language Proficiency; *Language Research; *Language Tests; Phonology; Second Language Learning; *Spanish; Teaching Methods; Testing
IDENTIFIERS	*Dictation

ABSTRACT

Recent research by Valette, Oller, and others has shown the utility of dictation as a measure of general language competence when correlated with achievement and proficiency batteries for French and English as a second language. No such studies have been conducted with Spanish. The investigator hypothesized that since Spanish is a phonetic language permitting easy transcription without comprehension, the dictation would not serve as a good substitute measure of language competence. In order to test this hypothesis 127 students enrolled in first-year Spanish at the University of Colorado were read a 106-word dictation together with a 100-item final examination. The results of each test were then correlated by computer and a Pearson product-moment coefficient of .50 was obtained. The scores on both tests for all students are displayed on a scatter diagram, and the reliability of both tests was ascertained using the Kuder-Richardson 21 formula. The results indicate general confirmation of the research hypothesis that the dictation is significantly less useful as a proficiency measure for Spanish than for French and English. (Author)

"Dictation as a Measure of Spanish Language Proficiency"*

The dictation is one of the oldest techniques known for testing progress in foreign language learning. It has long been associated with the traditional or grammar translation method and for this reason was rejected by Gouin and others who propagated the "natural" method during the second half of the nineteenth century. Later it became popular again under the direct method, especially as phonetic dictation or phonetic transcription of spoken language, a task which was especially pleasing to direct methodologists because of their scientific interest in phonology. The reading method, which was popular during the 1930's and 1940's, employed the dictation only sparingly since this method's emphasis on listening comprehension and spelling was slight.

With the advent of the audio-lingual method at the beginning of the 1960's, dictation again received considerable criticism partly due to its association with the writing skill, which was to be postponed, and partly because of it's association with the grammar translation method which became the whipping boy of new-key methodologists. Therefore, in spite of substantial support for research in foreign language learning during the decade following N.D.E.A., no research or interest in the dictation was demonstrated, but for a single exception.

In 1964, Valette reported on a study she conducted at the University of South Florida. During a first semester French course, she divided six beginning French classes into two treatment groups. Group A received regular dictations during each class meeting throughout the semester. Group B received only sporadic dictations for a total of only three or four during the semester. At the close

*The following paper was presented at the Seminar on Tests and Testing held at the 1977 TESOL Convention on April 27, 1977 in Miami, Florida under the sponsorship of the International Association of Applied Linguistics.

of the course, all sections received the same departmental examination which included a 50 word dictation. The maximum possible raw score on the examination was 155, while maximum raw score on the subpart was 20. After calculating a Pearson product-moment coefficient for both groups between part score on the dictation and total score on the test, Valette found correlations of .78 for Group A and .89 for Group B. She concluded that for French at least the dictation is a reasonably good measure of overall student proficiency, especially when practice in taking dictation has not been offered.

Ironically, Vallette's finding went relatively unnoticed for the remainder of the 1960's. Interest in the dictation returned again in the 1970's due to the extensive research into integrated measures of language proficiency spearheaded by John Oller. After reporting on the high correlations obtained between short cloze tests and multi-section proficiency tests, Oller and others turned their attention to the dictation. In 1971, Oller reported a correlation of .86 between a section on dictation and total score on the UCLA English as a Second Language Placement Examination. Four years later in response to criticism of his figures by Breitenstein, Oller and Streiff published a corrected correlation coefficient of .94.

In more recent article on testing E.S.L. university students in Iran, Irvine, Atar, and Oller reported similar findings, although of lesser magnitude, after correlating scores on a cloze test and a dictation with scores on the Test of English as a Foreign Language published by Educational Testing Service.

Thus far research on the dictation has focused on two languages, English and French. The results of this research have greatly simplified the task of proficiency and placement testing in these languages and rejuvenated confidence in the use of dictation by E.S.L. and French teachers. Oller and others have posited that the success of the dictation is due to the fact that it itself is an integrated measure of language competency, testing many factors such as sound discrimination, word recognition, rapid decoding of speech, information storage in short term memory, recoding and spelling. In several interesting articles, he refers to the cognitive model of the active listener who constructs in his mind what the speaker is saying or will say, and then compares this expected model with what he actually hears. The ability to accurately and rapidly construct this model which Oller calls a grammar of expectancies, and then accurately compare it with the perceived stream of speech, is viewed as the active application of the listener's underlying linguistic competence in the language.

Yet it remains to be seen whether the activity is as useful in languages such as Spanish and German which show much simpler phonological and orthographic systems thereby allowing the learner to merely transcribe what he hears. In order to ascertain whether the dictation correlates highly with comprehensive language skills test scores in Spanish, and can therefore serve as a good "quick and dirty" Spanish proficiency measure, the researcher conducted the following experiment.

Methodology. All one hundred twenty seven (127) students enrolled in the second semester of a first year Spanish course at the University of Colorado were selected as subjects. Undergraduate students at the University of Colorado are generally above average in intelligence and show mean scores of about 550 on the verbal and 570 on the quantitative sections of the College Entrance Examination Board's Scholastic Aptitude Test. The subjects were given a 100 item final written examination three hours in length. In addition, approximately one hour after the test began, students were administered a 106 word dictation constructed by the investigator and a graduate student. When both tests were graded they were turned over to the researcher for statistical analysis.

Since the purpose of this study is to determine the suitability of the dictation as a substitute measure of achievement or proficiency, the two sets of data were correlated using the Pearson product-moment correlation coefficient (r_{xy}). This statistical technique is appropriate when we wish to correlate scores on two interval scales. The following decision rules were made regarding the interpretation of the data. Since a high correlation coefficient is necessary to demonstrate the empirical validity of an instrument, and since the N was large enough to give substantial power to the statistical technique, it was decided that the null hypothesis, $p_{xy} = 0$, would not be rejected unless the probability of the obtained r was less than one percent ($p < .01$). Furthermore, since correlations based on a single sample are subj ct to sampling error, it was decided to construct a 95% confidence interval around the obtained correlation coefficient using Fisher's Z-transformation of r as described by Glass and Stanley. Such a confidence interval offers considerable certainty of capturing the true correlation between the two instruments while providing a truer picture of the generalizeability of the findings.

The data were then analyzed as described above by a CDC 6400 computer employing the standard statistical programs included in the Statistical Package for the Social Sciences.

Calculation of the reliability of a teacher scored test is often tedious, particularly when the test contains a large number of items and is given to a large number of students. In order for the normal point-biserial coefficient to be calculated, it would be necessary to tally correct and incorrect responses to every item on the two tests. This would involve the collection of some 27,000 pieces of data.

Fortunately, there is a much simpler procedure available to researchers known as Kuder-Richardson Formula 21. KR 21 requires knowledge of only the test mean, the standard deviation, and the number of items on the test, and these statistics are readily available through computer analysis of scores. KR 21, however, is only an estimate of true reliability. In analyzing 58 tests, Lord found that KR 21 consistently underestimated the true reliability, though usually by .05 or less. Because of this, KR 21 is often used as a lower-bound or minimal estimate of reliability (Stanley and Hopkins, 1972, p. 127). In this study, KR 21 reliability coefficients were calculated by the investigator on a hand calculator.

Instrumentation. The final examination was an achievement test based on the content of the textbook Espanol a lo vivo by Hansen and Wilkens. It was graded by each student's regular instructor based on a previously agreed upon system of scoring. Since some sections required the student to write several words or a sentence, some errors counted only one-half point. The examination consisted of sentence rewrite exercises using various syntactical transformations, and fill-in-the-blank exercises for testing morphology. It was a totally discrete point test and was given during final examination week in May, 1976. The test was designed to be cumulative in nature and covered the content of the entire text. It was not merely limited to the second half of the book. Because of this, the test can be considered a good indicator of the beginning student's grammatical competence.

The dictation was likewise based on the vocabulary and structures encountered by students taking first year Spanish with the Hansen and

Wilkins text. Designed to be somewhat challenging in order to obtain a reliable spread of scores, it was administered in the normal manner as described by Valette and others, and lasted about twelve minutes. Students first listened to the entire paragraph for meaning. Then each breath group of five to eight words was read twice by a graduate student from Mexico. Finally, the entire selection was reread at normal speed. Following the dictation, students continued work on the examination. Since one test was given in the middle of the other, history and maturation can be disregarded as possible threats to the internal validity of this study.

The dictations were graded by a graduate student under the dir :tion of the investigator based on a system which counted one point off for each word which was incorrectly written in any way or which should not have appeared. No partial credit was given since previous research on the cloze test has shown no change in the respective ranks of subjects when more elaborate scoring systems are used. Oller (1975) has also employed this same procedure in his research on the dictation.

Analysis of Results: Table 1. depicts some descriptive statistics for both tests. The correlation between the two indices was .495 which is significant at the .001 level. This again confirms the validity of the dictation as a proficiency measure.

By employing Fisher's Z-transformation of r, we can produce a confidence interval on p, the true correlation between the two indices for the population, by a process which captures p within its limits 95% of the time. The resulting interval is .36 - .72. This means that we can be reasonably certain that the true correlation produced by taking an infinite number of samples is greater than .36 and less than .72.

TABLE I
DESCRIPTIVE STATISTICS

Final Exam Score			Errors on Dictation	
Mean	68.9		Mean	17.9
Std Dev	16.6		Std Dev	7.2
Std Err	1.4		Std Err	.6
High	94.5	$r_{xy} = -.495$	High	40.0
Low	12.5	$p < .001$	Low	6.0
Range	82.0		Range	34.0
r_{KR21}	.93		r_{KR21}	.72

Note: The score on the final exam was determined by counting the number of right answers. The score on the dictation was determined by counting the number of errors. The result is a negative correlation coefficient. A positive correlation of equal magnitude would be derived by scoring according to the number of right answers or wrong answers on both tests.

The reliability of both tests is good, particularly when one considers that the coefficients reported here, .72 for the dictation and .93 for the final exam, are minimal or lower bound estimates. The reliability of the final exam indicates that the test functioned as an effective discriminator between different levels of knowledge among first year students.

Because the relationship between two variables will be weakened by any lack of reliability in the measurement of either or both variables, statisticians have developed a technique for deriving the true correlation. This procedure, called the correction for attenuation, estimates what the correlation between two variables would be if both tests were perfectly reliable. The formula for deriving an estimate of the "true" relationship is

$$r_{t_x t_y} = \frac{r_{xy}}{\sqrt{r_x r_y}}$$

where: $r_{t_x t_y}$ = the correlation between true scores on variables x and y.

r_{xy} = the obtained correlation between variables x and y, and

$r_x r_y$ = the reliability coefficients of variables x and y, respectively.

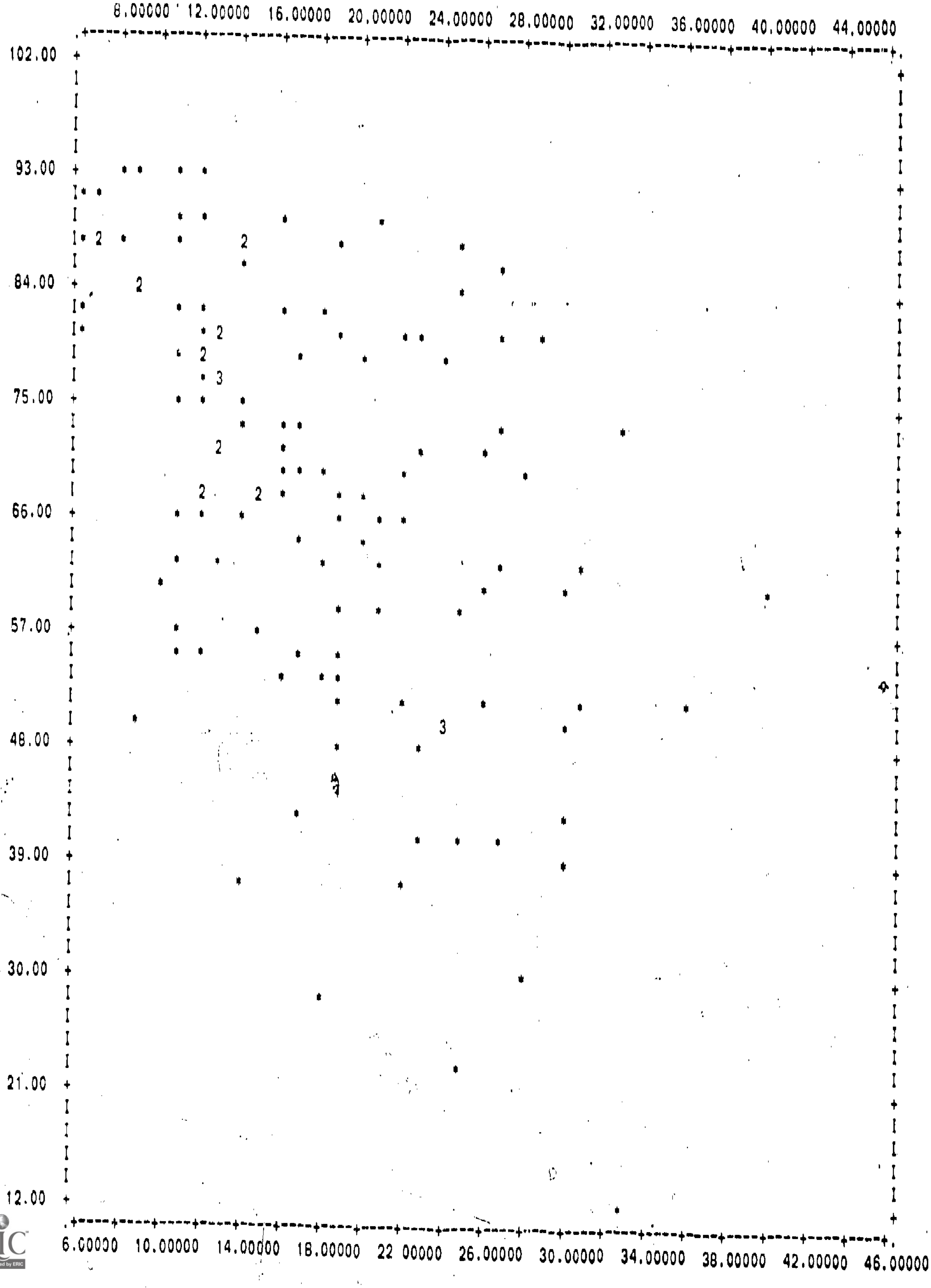
By substituting the obtained coefficients into the above formula, we get:

$$r_{t_x t_y} = \frac{.495}{\sqrt{(.93)\ (.72)}} = .61$$

This procedure further corroborates the confidence interval (.36 - .72) which was developed earlier. Again it appears that even if both tests had been harder, resulting in a greater dispersion of scores and differention among students, the resulting correlation between test scores would still be moderate, rather than high.

Since correlations may be linear, curvilinear, or random, it is best to depict them on a scattergram (or

SCATTERGRAM OF (DOWN) FINAL , FINAL EXAM SCORE (ACROSS) DICTADO ERRORS ON DICTATION



scatter diagram). A visual understanding of the strength of a relationship can be gained by studying a two-way scattergram of tallies. Each asterisk represents the intersection of two scores for a single subject. If two or more subjects fall into the same position on the scattergram, the actual number of subjects is printed.

The relationship between two indices is linear if an imagined straight line through the center of the tallies, called a regression line, more closely fits the pattern of the scattergram than does any curved line. The scattergram reproduced here indicates a definite linear relationship between the two indices, so that as the score on the final decreases, errors on the dictation increase. Nevertheless, the relationship depicted is far from perfect as one can readily perceive many scores which do not fit the regression line closely. In such cases, the score on one test will not serve as a predictor of the score on another since the difference between the predicted score and the actual obtained score is considerable. It is on these differences, otherwise known as errors in prediction, that the correlation coefficient is based.

Conclusions and Discussion. This study compared scores on a dictation with scores on a 100 item Spanish achievement test considered to be a valid indicator of overall grammatical competence. It found the dictation to be only a moderately good indicator of overall competence when used with learners of Spanish. It does not purport to contradict the findings of other researchers who have found higher correlations for learners of English and French. Indeed, as the investigator hypothesized before collecting the data, Spanish can be transcribed by the language learner with considerable facility, and this reduces the dependency on the internalized grammar of expectancies posited by Oller. While the Spanish learner will employ his internalized grammar of the language in taking a dictation, he is not left to depend on it alone. If he does not recognize a word he may still transcribe it correctly due to the good fit of the language. He is not forced to construct a distorted version of what the "dictator" has said (though he sometimes does) through the active use of his internalized grammar. He can instead rely on simple spelling conventions to fill in the gaps when his linguistic competence fails him. Oller (1976, p. 77) has stated, "Low intercorrelations must be interpreted as indicating low test validity, i.e., that one of the tests being correlated does not tap underlying linguistic competence or that it does so to an insufficient extent." It is my belief, supported by the findings described here, that the dictation does not sufficiently tap the learners underlying competence so that the learner must depend on that competence exclusively in order to perform correctly in Spanish. On the other hand, the validity of the close test as a proficiency measure would be generalizeable to Spanish because in constructing an appropriate response the learner is depending exclusively on clues provided him by his internalized grammar. If such is the case, we can expect a large disparity in correlations on close tests and dictations in Spanish. It is probable that future research applying integrated measures to languages with good fit will demonstrate this.

Charles Stansfield
University of Colorado
Boulder, Colorado

References

Breitenstein, P.H. (1972). "Reader's Letters." English Language Teaching. 26:2, 202-3.

Glass, Gene V. and Julian C. Stanley (1970). Statistical Methods in Education and Psychology. Englewood Cliffs: Prentice Hall, pp. 308-309.

Hansen, Terrence L. and Ernest J. Wilkins (1974). Espanol a lo vivo. New York: John Wiley.

Irvine, Patricia Parvin Atai, and John W. Oller, Jr. (1975). "Cloze, Dictation, and the Test of English as a Foreign Language." Language Learning. 24:2, 245-252.

Neisser, Ulric (1967). Cognitive Psychology. New York: Appleton-Century-Crofts.

Nie, Norman H., C. Hadlai Hull, Jean G. Jenkins, Karin Steinbrenner and Dale H. Bent (1975). Statistical Package for the Social Sciences. New York: McGraw Hill.

Oller, John W., Jr. (1971). "Dictation as a Device for Testing Foreign Language Proficiency." English Language Teaching. 25:3, 254-9.

Oller, John. (1972). "Scoring Methods and Difficulty Levels for Cloze Tests of Proficiency in English as a Second Language." Modern Language Journal. 56, 151-158.

Oller, John W., Jr. (1973). "Discrete-Point Tests Versus Tests of Integrative Skills." in Oller and Richards. Focus on the Learner: Pragmatic Perspectives for the Language Teacher. Rowley, Mass.: Newbury House.

Oller, John W., Jr. (1975). "Dictation: A Test of Grammar Based Expectancies." in Jones and Spolsky. Testing Language Proficiency. Arlington, VA: Center for

Applied Linguistics.

Stanley, Julian C. and Kenneth D. Hopkins (1972). Educational and Psychological Measurement and Evaluation. Englewood Cliffs, N.J.: Prentice Hall.

Valette, Rebecca M. (1964). "The Use of the Dictée in the French Language Classroom." Modern Language Journal. 48:7, 431-4.

Valette, Rebecca M. (1967). Modern Language Testing: A Handbook. New York: Harcourt, Brace and World, pp. 140-141.