

## THE NELSON-DENNY READING TEST: FAIR MEASURE FOR NON-NATIVE SPEAKERS?

Mary Heise

### Abstract

The purpose of this study was to determine whether reading test intended for native speakers of English can be used to determine the reading proficiencies of non-native speakers of English. The test was administered to groups of both native English-speaking and Japanese freshmen at Creighton University in Omaha, Nebraska, U.S.A. This article includes a description and critical analysis of the test and a statistical analysis of the test results for each of the two groups of students. The conclusion is that the test is inappropriate for use with non-native speakers of English, and that the test should not be used as a predictor of academic success or as an admissions criterion for non-native English-speaking applicants to North American universities.

The purposes of this study are to determine whether or not the norm-referenced *Nelson-Denny Reading Test* (NDRT) can be used as a criterion in determining a foreign student's readiness to compete with American students in the skill area

---

Mary Heise has an M.S. degree in Applied Linguistics from Indiana University. She has taught ESL at Indiana University and at Creighton University since 1978. She coordinated the English Language Program at Creighton University from 1980-82. The largest proportion of Creighton University's international students are from Japan.

of reading at an American university, and to compare the relative performance of this select group of non-native English-speaking freshman students with a typical group of American freshman students.

The sample groups in the study are quite small. Nevertheless, a statistical analysis of the test results has been prepared. This analysis includes means, ranges, medians, standard deviations, item difficulty and item discrimination indexes, reliability quotients and correlations with the *Test of English as a Foreign Language* (TOEFL).

The NDRT is intended to serve predictive, screening, and broadly diagnostic purposes. The test provides measures of vocabulary, comprehension and rate for students in grades 9-15+.

The vocabulary subtest of 100 multiple-choice items has a time limit of ten minutes. Most of the items are presented in isolation, although a few have minimal context. The words are presented as an incomplete definition which the student is asked to complete by choosing the correct option. Fifty of the items are prefixed; 50 are non-prefixed.

The comprehension and rate subtest consists of eight passages with 36 multiple-choice questions. The time limit is 20 minutes. The questions require the students to find details, draw conclusions, make generalizations, recognize relationships, styles and moods, as well as to determine the writer's purpose and the main idea. The readings and the questions deal equally with subject areas within the humanities, natural and social sciences. Reading rate is tested by asking the students to mark the number of words they read during the first minute of the reading comprehension subject.

Four forms of the NDRT are available. Forms A and B were published in 1960; forms C and D in 1973. The authors state that the four forms reflect slightly different emphases of subject matter and levels of difficulty; therefore, they should provide greater opportunity for users in selecting the

## The Nelson-Denny Reading Test

most appropriate form for a given situation. However, no guidelines are provided for selecting the best form (Buros, 1978: 1207).

The norms for the NDRT, Forms C and D, were established following testing in 1972. The reliability of comparing the scores of a sample group in 1980 to these norms is questionable. The norms include percentile ranks and grade equivalents. The test manual contains a thorough description of norms for grades 9 through 12. However, the data supplied for grades 13 through 16 was based on a significantly smaller test population. The make-up of this last sample group is unusual for a test which reports norms for grades 6 through 15+, as the sample group included college students of all levels as well as graduates and faculty.

### Method

The NDRT, Forms C and D, was administered to two sample groups. The first group consisted of 24 Japanese students who had begun their freshman year at Creighton University in August 1980. Form C was completed by 14 students; Form D by 10. The test was administered to this group on July 14, 1980 on a volunteer basis. There was a range in the English language abilities of these students as evidenced by their TOEFL scores. The TOEFL scores of these students as of May 1, 1980 ranged from 430 to 557.

The second test group consisted of 31 American college freshmen. Form C was completed by 16 students; Form D by 15. The two testing dates for this group were September 10, 1980 and September 30, 1980. These students had also begun their freshman year at Creighton University in August 1980. Three of these students had been placed in a remedial reading course during their first semester of college as a result of their American College Testing Program (ACT) scores and/or their

high school records. According to their ACT scores, the academic abilities of these students as reported by this particular measure (ACT) ranged from low to extremely high.

## Results and Discussion

### Mean, Range, Median and Standard Deviation

All scores were computed on the scale provided for grade 13 in the examiner's manual. A close observation of the data for the vocabulary subtest reveals that both the percentile ranks and the grade equivalent scores are substantially lower for the Japanese than for the American students (see Appendix B). The medians in the case of each form are equally if not more indicative than the means. The range of the Japanese students' percentile ranks is narrow compared to that of the American students.

An observation of the data for the comprehension subtest (Appendix B) shows an even sharper discrepancy between the two groups. In general the comprehension scores are much lower for the Japanese students. In Form D the ranges of the comprehension scores for the two groups are quite close. However, in Form C there is a startling difference at the higher end of the range of the percentile ranks (28 for the Japanese; 96 for the Americans).

The total scores (total = vocabulary + comprehension scores, Appendix B) confirm the results discussed for the two subtests, i.e., the generally lower scores of the Japanese students, which are much narrower in the percentile ranks than in the grade equivalents.

Statistics for reading rate were not included in the reported data because a significant number of students neglected to report their rates. Moreover, the norm data in the manual was not complete enough to allow for sufficient interpreta-

## The Nelson-Denny Reading Test

tion of the reported scores as a group. However, a glance at the reported scores reveals that there is a wide range in the scores of both groups, with the scores of the Japanese being generally lower.

### Reliability

The reliability of the test according to the examiner's manual varies according to the subtests. The vocabulary subtest is reported to be the most reliable, and the reliability of the comprehension and rate subtests are low. The intercorrelations of the subtests are high (Buros, 1978: 1209). According to one authority on reading and reading tests, subtests are often too short to have adequate reliability (Farr, 1969: 64). This could be true in this case.

The reliability quotients for the two test groups in question were calculated according to the Kuder-Richardson formula. The results are in Table 1 (Appendix A).

In most cases (the exception being the vocabulary subtest for the Japanese students), the reliability of Form D is better for both groups. For the American group the reliability of Form D is quite good.

The reliability of the measure of reading rate was not obtained. The reliability and face validity appear to be low for three reasons. First, rate should be tested over a period of at least three minutes, with five minutes being preferable (Farr, 1970: 223). The second reason is that under normal circumstances most readers experience a slack in their speed. Readers are in effect being told to read fast during a test of reading rate, and therefore their speed will often increase by as much as 40 percent (Farr, 1969: 50). The third reason is that there is no one reading rate for an individual. The various reading rates of an individual vary according to the nature of the material as well as according to the purpose of the reader (Farr, 1970: 222).

Caution must be used when using the NDRT for diagnostic purposes because of the questionable reliability of many of the subtests. It is especially presumtuous to make any diagnosis with reference to a student's reading rate based on the one-minute sample in the NDRT.

### **Item Difficulty and Discrimination**

According to more than one source, the level of material presented in the NDRT is quite difficult (Pumfrey, 1976: 126). For example, if one accepts the widely-accepted criterion which states that a test will discriminate among a test population in which half of the items are answered correctly, one would have to go up to the grade-16 level before this is true of the population described in the examiner's manual (Buros, 1978: 1210).

Presented in Table 2 (Appendix A) are the item discrimination data for both groups. Provided for each subtest are the percentages of test items with a satisfactory item discrimination index of  $+0.30$  or above. The information was obtained by using the upper and lower 25 percent tests of each group (Forms C and D combined) according to their raw total scores.

There is little beneficial information which can be gained from this data other than the fact that Form D discriminates better than Form C for both groups.

The item difficulty information is presented in Table 3 (Appendix A) as percentages of excessively easy items (over 90 percent correct) and excessively difficult items (under 30 percent) in each subtest.

The high percentage of difficult items on the vocabulary subtests for both groups are in large part due to the time limit. The total number of items out of the 100 present in the vocabulary subtest completed by all members of the group ranged from 34 for the American students on Form D to 48 for the Japanese students on Form D.

## The Nelson-Denny Reading Test

In summary, both tests were significantly more difficult for the Japanese students. The time pressure did not appear to be the only factor, as a significant number of errors were made on the questions which were answered by all the students. Each subtest is discussed below together with some possible reasons for its apparent difficulty.

One major factor contributing to the difficulty of the vocabulary subtest is its time limit. Students have only ten minutes to answer 100 items. A strict time limit may increase random guessing which in turn affects the test reliability. No data relating to the effect of the time limit is supplied in the manual. Strict time limits reduce the test's validity as a measure of vocabulary. When both speed and vocabulary are tested together, what is actually being measured is some unknown combination of these two skills (Farr, 1969: 35).

The test manual refers to a study by Holmes and Singer which has been widely criticized in part because of its inappropriate interpretation of the reading process. The manual refers to three factors which influence vocabulary according to this study: vocabulary in context, vocabulary in isolation, and prefixes. The manual advises that since one factor (prefixes) could be related to the NDRT items, such information should be used for diagnosis (Buros, 1978: 1210). The information provided in the manual, namely which items are with or without prefixes, is of no value in this study. No pattern developed among the answers of either group.

For the most part the vocabulary items are presented without context. In a few cases there is minimal context. As the test is one of reading, and a reader relies heavily on semantic and syntactic clues to determine word meanings, the validity of asking students taking a reading test to define words largely out of context is questionable.

Assessing a student's performance in reading comprehension on the basis of the Nelson-Denny reading comprehension subtest is a complex matter. Some factors affecting scores

in many reading comprehension tests are time limits, interest, the subject matter, the level of the material, organizational patterns, structural language patterns, the reader's purpose, the reader's command of basic decoding skills, the level(s) of questioning used, whether the examinees are allowed to refer back to the selection, and scheduled or unscheduled interruptions during the test (Farr, 1969: 52ff.). Those factors which apply are discussed below.

The time limit is quite strict for this subtest as well. Twenty minutes are allowed for the reading of eight passages of varying length and difficulty and for the answering of 36 questions. Most students in both groups did not complete this particular subtest. The students were allowed to refer back to the selections when answering the questions.

The reading levels of the passages were checked with the Fry Readability Graph (West, 1978: 31). The results should be viewed as approximations, because of the reliability of not only the Fry Graph but of any readability formula and because most of the selections are too short to obtain the suggested number of three samples for the use of the Fry Graph. For Form C the readability estimates range from grade levels 10.0 to 17.0+ with an average of 12.5. For Form D the passages range in difficulty from 7.0 to 17.0+ with an average of 10.8. These figures represent a wide range of levels. A discrepancy in the difficulty of the two forms may or may not be present, depending on the accuracy of the graph and the possible altering of the results due to the short length of the selections.

The manual provides information regarding the authors' alternative to the traditional levels of questioning. They include only the literal and interpretive levels in the above category, while evaluations are considered with details and writer's purpose in a category devised according to the levels of abstraction presented by the questions. For the purposes of this study, concentration was given chiefly to the literal,



## The Nelson-Denny Reading Test

evaluative and interpretive levels of questioning. The method employed to determine "easy" and "difficult" questions is that method described in the section of the study concerned with "Item Difficulty and Discrimination." On Form C, all of the easy questions (6/6) for the American group fall into the literal category. Most (5/6) of the difficult questions are at the interpretive or evaluative level. On Form D for the same group the one easy question is literal, while the one difficult question is interpretive. On Form C for the Japanese group there are no easy questions, while approximately two-thirds (13/18) of the difficult questions are at the interpretive or evaluative levels. Note that a significant number (5/18) are, however, at the literal level. On Form D, there is both one easy and one difficult question for the Japanese at the literal level, while most (15/16) of the difficult questions are at the interpretive or evaluative level. On both forms and for both groups time was evidently a distinct factor in the above results. Many of the difficult questions are often clustered in the last half of the 36 questions. In spite of this factor, the results described above indicate that, in general, even literal questions are not easy for the Japanese, while they constitute about 50 percent of the easy questions for the Americans. Most of the difficult questions for both groups are at the interpretive or evaluative levels, but there are many more difficult questions for the Japanese than for the Americans.

The comprehension questions are divided into subject areas (humanities, natural sciences, social sciences) in the manual. On Form C, four out of six of the easy questions were in the natural sciences for the American group. There was not a sufficient number of easy questions for this group on Form D to make a similar assessment. On both forms for the American group, a significant number of questions was answered correctly by a majority of the students (88 percent); however, no pattern emerges which points to any one or two subject areas as strong points for this group. On Form C for

the Japanese students there were no easy questions. The difficult questions are distributed in the following manner: out of a possible total of 18, 2 are in the natural sciences, 6 are in the humanities, and 10 are in the social sciences. The expected pattern above does not, however, hold true for Form D. The one easy question for the Japanese is in the humanities. The distribution of the difficult questions is as follows: out of a total 16, 7 are in the natural sciences, 3 are in the humanities, and 6 are in the social sciences. Therefore, for this select group of Japanese, subject area does not appear to have played a role in item difficulty.

There are factors related to the general effectiveness and validity of the standard type of reading comprehension tests (tests which include a reading passage and questions). One such factor is passage independence, that is, when questions can be answered without reading the text. In one study it was found that inferential questions are more passage independent than literal factual questions (Duffelmeyer, 1980: 131-33).

Prior knowledge and guessing are influencing factors. According to Farr, a comprehension score of 55 percent (this figure as meant merely as an example, not as a point of reference) without reading the selection would not be unusual due to these combined factors (Farr, 1969: 46).

Passage dependence, that is, matching the possible answers to phrases in the reading passages, is another factor. In a recent study 8 items from the Nelson-Denny test were used. The "stem" or opening phrase of each question was removed. The appropriate reading passages and only the four possible answers to each of the 8 questions were given to 27 graduate students. The students were instructed to select the correct answers by selecting those answers which included words that also appear in the passage or by selecting the one with words mentioned most frequently in the passage. The mean score for this group was 6.7 out of 8. Upon analyzing the test

## The Nelson-Denny Reading Test

items, it was found that even a slight change of word form increased the difficulty of word matching and counting. For example, if "pursuit of perfection" in the passage was changed to "pursued perfection" in the answer choice, some students were confused. The point of this study was that in tests of reading comprehension there are two alternative possibilities: words from the passage should be used in every answer choice, or the answer choices should be entirely reworded (Pyrczak, 1980: 162).

In view of the factors mentioned above, it is questionable whether such reading comprehension tests provide an accurate measure of achievement or proficiency.

The recording of the students' reading rates after the first minute of the reading comprehension test is a critical interruption. The interruption is not worthwhile for a number of reasons, some of which were discussed previously in this article. Another major difficulty with such a measure is that speed is not important unless at least a minimal level of comprehension is maintained (Farr, 1969: 45f.). One problem with the NDRT is that rate and comprehension are measured independently of each other.

### Correlations

Correlation figures for the NDRT and the *Preliminary Scholastic Aptitude Test* (PSAT) and the *Scholastic Aptitude Test* (SAT) were quite high according to the examiner's manual (Brown, 1976: 30). Most of the students who participated in this study took the academic tests offered by the American College Testing Program (ACT). There are no tests offered by this program which are solely devoted to reading comprehension or vocabulary knowledge. The tests which do not include reading comprehension have, in addition, general background questions. Therefore, correlations with the ACT are not included in this study.

Twenty-one of the Japanese students who participated in this study also took the TOEFL examination on May 1, 1980. The TOEFL includes subtests in reading, structure and listening comprehension. As the skills measured by the TOEFL are diverse, a correlation was not made between the Nelson-Denny reading subtests and the TOEFL total. However, the correlations presented in Table 4 (Appendix A) were made. There is little correlation between the two tests.

The evidence in the NDRT's examiner's manual presents weak evidence of predictive validity. The scores are related to scores in several college courses as well as overall grade point averages (Buros; 1978: 1209). No attempt was made to correlate the NDRT with grade point averages of the students in this study, as none of the students had completed a full semester of at least 12 credit hours at the time this study was completed.

### Conclusion

In spite of the small size of the two test groups, the following conclusion can be drawn. As any standardized test of reading comprehension and vocabulary, the NDRT has severe limitations, especially for an audience for which it was not originally intended. These limitations, some of which are the time limit, difficulty, guess factor, etc., seriously affect the validity of the test results. Therefore, the test should not be used as a factor in college admission or as a predictive device for a group of non-native English speakers.

The study does indicate statistically that a group of non-native speakers does experience, for the most part, a tremendous amount of difficulty in competing with native English-speaking American college students in situations comparable to a standardized test of reading comprehension and vocabulary knowledge. Therefore, their deficiencies in reading will adversely affect their performance in tests, especially when a severe time limit is imposed.

# The Nelson-Denny Reading Test

## Appendix A

### TABLE 1 – RELIABILITY

	Form C			Form D		
	Voc.	Comp.	Total	Voc.	Comp.	Total
Americans	.77	.75	.79	.93	.93	.96
Japanese	.58	.73	.59	.52	.84	.84

### TABLE 2 – ITEM DISCRIMINATION

	Form C		Form D	
	Voc. (100 items)	Comp. (36 items)	Voc. (100 items)	Comp. (36 items)
Americans	26%	28%	57%	72%
Japanese	34%	11%	42%	63%

### TABLE 3 – ITEM DIFFICULTY

	Form C		Form D	
	Voc. (100 items)	Comp. (36 items)	Voc. (100 items)	Comp. (36 items)
<b>Easy Items</b>				
Americans	9%	36%	3%	3%
Japanese	0%	0%	0%	3%
<b>Difficult Items</b>				
Americans	49%	17%	41%	3%
Japanese	85%	50%	69%	44%

### TABLE 4 – CORRELATIONS WITH TOEFL

	TOEFL Reading Subtest	TOEFL Total
Nelson-Denny Reading Subtest	.21	—
Nelson-Denny Total	.39	.45

# JALT JOURNAL, VOLUME 6 (1984)

## Appendix B

Following is the raw score data for the vocabulary and reading comprehension subtests and the total scores for both test groups.

FORM C		Voc.	Comp.	Total (Voc. + Comp.)
Mean	American	54	45	50
	Japanese	15	12	11
Range	American	28-93	8-96	17-90
	Japanese	1-46	0-28	1-27
Median	American	43	41	41
	Japanese	12	8	10
Standard Deviation	American	19.7	23.3	22.3
	Japanese	11.6	9.0	7.5

FORM D		Voc.	Comp.	Total (Voc. + Comp.)
Mean	American	68	57	62
	Japanese	12	14	11
Range	American	19-99	1-99	1-99
	Japanese	2-35	2-83	1-57
Median	American	77	72	79
	Japanese	7	4	6
Standard Deviation	American	27.2	35.9	35.5
	Japanese	10.0	23.4	15.8

FORMS C and D		Voc.	Comp.	Total (Voc. + Comp.)
Mean	American	61	45	55
	Japanese	13	13	11
Range	American	19-99	1-99	1-99
	Japanese	1-46	0-83	1-57
Median	American	57	43	52
	Japanese	10	8	7
Standard Deviation	American	24.5	26.7	27.6
	Japanese	10.7	10.7	11.7

## The Nelson-Denny Reading Test

### References

- Brown, J. I., Denny, E. C., & Nelson, M. J. *The Nelson-Denny reading test: Examiner's manual*. Boston: Houghton Mifflin, 1976.
- Buros, I. K. (Ed.). *The eighth mental measurements yearbook*. Highland Park, New Jersey: The Gryphon Press, 1978.
- Duffelmeyer, F.A. The passage independence of factual and inferential questions. *Journal of Reading*, 1980, 24 (2), 131-133.
- Farr, R. *Reading: What can be measured?* ERIC/CRIER Reading Review Series. Newark, DE: International Reading Association, 1969.
- Farr, R. *Measurement and evaluation of reading*. New York: Harcourt, Brace and World, 1970.
- Pumfrey, P. D. *Reading: Tests and assessment techniques*. Edinburgh, Great Britain: Hodder and Stoughton, 1976.
- Pyrczak, F. Another twist on passage dependence. *Journal of Reading*, 1980, 24 (2), 162.
- West, G. B. *Teaching reading skills in content areas: A practical guide to the construction of student exercises* (2nd ed.). Oviedo, Florida: Sandpiper Press, 1978.