A STUDY OF TEST-TAKING SKILLS AND ACHIEVEMENT SCORES UPON SECONDARY STUDENTS

A Dissertation
Presented to
The School of Graduate Studies
Drake University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Education

by
Diane Vida
August 1985
A STUDY OF TEST-TAKING SKILLS AND ACHIEVEMENT SCORES
UPON SECONDARY STUDENTS

by
Diane Vida

Approved by Committee:

Dr. Bruce Vennard, Chair

Dr. Raymond Hock

Dr. Phillip Levine

Dr. W. Russell Abell
Dean of the School of Graduate Studies
A STUDY OF TEST-TAKING SKILLS AND ACHIEVEMENT SCORES
UPON SECONDARY STUDENTS

An abstract of a Dissertation by
Diane Vida
August 1985
Drake University
Advisor: Dr. Bruce Vennard

The problem. This study was conducted to determine if secondary students whose reading abilities were below grade level could improve their reading comprehension and vocabulary test scores on standardized tests, whether the test-taking skills program influenced these students' test scores, and how a test-taking skills program could improve students' future standardized test scores.

Procedure. The Nelson-Denny Reading Tests were used as the measuring instrument. Form E tests were given to all students for the pre-test scores. Students were randomly placed in experimental and control groups. After thirty-six, twenty-minute sessions (a total of twelve hours of instruction time) using published test-taking skills and strategies by World Book, students were post-tested with Form F tests. The statistical findings were calculated using t-Test results on paired and group scores.

Findings. This study provided evidence that there was a significant difference in the mean gain scores of the reading comprehension scores of the experimental group. The null hypothesis was held tenable with the mean gain scores of the vocabulary sections and the reading comprehension of the control group. Data indicated that secondary students whose reading abilities were below grade placement could improve reading comprehension and vocabulary test scores with the use of a published test-taking skills program.

Conclusions and Recommendations. Data suggested that secondary students whose reading abilities were below grade level be given training in test-taking skills because students' scores on standardized tests tend to improve. Teachers should recognize that test scores are not solely the result of content knowledge. Knowing how to select the best possible choice for an answer can improve the student's test results. Other studies in test-wiseness, test-taking skills, long- and short-term training, uses of various published skills programs, and replication of this study with other locales and different subjects should be encouraged.
TABLE OF CONTENTS

List of Tables ........................................................................ v

Chapter

1. Introduction ................................................................. 1
   Statement of the Problem ............................................. 3
   Purpose of the Study ...................................................... 4
   Null Hypotheses ............................................................. 5
   Definition of Terms ....................................................... 5
   Importance of the Study .................................................. 6
   Assumptions .................................................................. 7
   Limitations ..................................................................... 7
   Organization of the Study ............................................... 8

2. Review of Literature ..................................................... 9
   The Studies .................................................................... 9
   Summary ...................................................................... 22

3. Methodology ............................................................ 23
   Introduction .................................................................. 23
   Population .................................................................... 23
   Procedures ................................................................... 23
   Instruments Used .......................................................... 25
   Data Collection ............................................................ 29
   Summary ..................................................................... 30
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Findings and Interpretations</td>
<td>31</td>
</tr>
<tr>
<td>Null Hypotheses</td>
<td>37</td>
</tr>
<tr>
<td>Summary</td>
<td>37</td>
</tr>
<tr>
<td>5. Summary, Conclusions, and Recommendations</td>
<td>38</td>
</tr>
<tr>
<td>Conclusions</td>
<td>39</td>
</tr>
<tr>
<td>Recommendations</td>
<td>39</td>
</tr>
<tr>
<td>Bibliography</td>
<td>41</td>
</tr>
<tr>
<td>Appendix</td>
<td>45</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Nelson-Denny Reading Test, Forms E and F, Descriptive and Reliability Data for Equating Samples</td>
<td>28</td>
</tr>
<tr>
<td>2. Paired $t$-Test Analysis for Experimental Group, Group 1 - Vocabulary Scores</td>
<td>33</td>
</tr>
<tr>
<td>3. Paired $t$-Test Analysis for Control Group, Group 2 - Vocabulary Scores</td>
<td>34</td>
</tr>
<tr>
<td>4. Paired $t$-Test Analysis for Experimental Group, Group 1 - Reading Comprehension Scores</td>
<td>34</td>
</tr>
<tr>
<td>5. Paired $t$-Test Analysis for Control Group, Group 2 - Reading Comprehension Scores</td>
<td>35</td>
</tr>
<tr>
<td>6. Group $t$-Test Analysis for Mean Gain Scores in Vocabulary</td>
<td>35</td>
</tr>
<tr>
<td>7. Group $t$-Test Analysis for Mean Gain Scores in Reading Comprehension</td>
<td>36</td>
</tr>
</tbody>
</table>
CHAPTER ONE
Introduction

A recent emphasis by administrators, teachers, authors, parents, and state leaders was to improve student test scores by updating curriculums, upgrading assignments, increasing graduation requirements, and revamping the teacher certification programs. All of these methods, somehow and some way, have an effect on students and, ultimately, their test performances. In recent years, many students have relied upon a sophisticated and strategic method, called "test-wiseness," that helped them learn to help themselves and their performances on standardized tests.

Too often, students earned lower scores on tests, than their knowledge or aptitude warranted, because they lacked a sophisticated approach to test-taking, that was, "test-wiseness."¹ Test-wiseness, or lack of it, had been a factor in determining these students' test results. These test scores did not indicate a true assessment of a student's capabilities. Teaching all students test-taking skills, that will help them make better choices on standardized

---
tests, was one approach for study of eliminating the "test-wiseness" factor.

Test-wiseness had a relatively short history in educational research.¹ The majority of studies conducted by researchers concentrated on the test-taking abilities of average and college-bound students. According to the studies done by Crehan, McPhail, Jongsma, and Weaver, many of these students had been trained or "coached" for improved test performance on standardized tests.² The overall findings of these studies had concluded that these students' test scores had been improved with stimulation and test-taking skills training.

However, while high-achieving students readily acquired test skills, the average or below-average ability student had problems in learning such skills.³ Few studies on test-wiseness and test-taking skills had been done on the secondary student who had below-average reading abilities. Since reading was the primary method of obtaining the information conveyed in standardized tests, the student who had a reading disability was handicapped by the testing

²Ibid., pp. 33-38.
Students whose reading abilities were below grade level expectations tended to obtain lower scores on standardized tests.\textsuperscript{1} Below-average ability students, in particular, often tended to do poorly on (standardized) tests because they either had given up on the difficult questions or guessed illogically for answers.\textsuperscript{2} These students not only needed to improve vocabulary and reading comprehension levels, they also needed to learn to improve with practice their test-taking skills and strategies.

Improving vocabulary and reading comprehension levels for all students were important skills. But the skills of test-taking and test-wiseness also appeared to be important for all students. Studies done by Jongsma and Warshauer had revealed that both low- and high-achievers who received test-taking instruction and test familiarity practice had obtained higher average scores than their counterparts.\textsuperscript{3}

\textbf{Statement of the Problem}

Researchers have known that test-wiseness cannot be totally eliminated from objective testing. Therefore, an

\footnotesize
\begin{flushleft}
\textsuperscript{1}Rawl, pp. 108-10. \\
\textsuperscript{2}Ibid. \\
\textsuperscript{3}Eugene Jongsma and Elaine Warshauer, \textit{The Effects of Instruction in Test-taking Skills upon Student Performance on Standardized Achievement Tests} (ERIC ED 114 408), p. 36.
\end{flushleft}
attempt was made to control individual differences in test-wiseness by teaching test-taking skills to all students, and especially, those students identified as "low" in test-wiseness.\textsuperscript{1} Test-wiseness could very well be an essential part of an educational curriculum.\textsuperscript{2} Regardless of the grade or subject area being taught, secondary students acquiring test-wiseness and test-taking skills often have improved reading test scores.\textsuperscript{3}

\textbf{Purpose of the Study}

Test scores have been an important part of our academic world. The primary purposes for test scores have been to provide a scale to measure knowledge and capabilities.\textsuperscript{4} Since our nation has been so concerned with test results, it appeared to be important that a correct assessment of students' abilities be made, because many people and institutions make use of these test scores and results.\textsuperscript{5}

The purpose of this study was to determine whether


\textsuperscript{2}Ibid.

\textsuperscript{3}Gail Cohen Weaver, "Teaching Children How to Take Standardized Tests," The Reading Teacher, October 1978, pp. 118-19.

\textsuperscript{4}Pauk and Millman, p. 11.

\textsuperscript{5}Ibid.
given instruction and practice in test-taking skills and test-wiseness, secondary students whose reading abilities were below grade level could improve their vocabulary and reading comprehension scores on standardized tests.

Null Hypotheses

The following hypotheses were tested:

1. There was no significant difference in the vocabulary test scores of those secondary students whose reading abilities were below grade level when trained in a specific test-taking skills program and those below grade level reading students who were in a control group.

2. There was no significant difference in the reading comprehension test scores of those secondary students whose reading abilities were below grade level when trained in a specific test-taking skills program and those below grade level reading students who were in a control group.

Definition of Terms

For the purposes of this study, the term "secondary student" will refer to the student who had completed eight years of primary or elementary levels of education.

The words "grade level" will refer to the actual grade placement of a student in secondary education.

The words "test-wiseness" and "test-taking skills" will refer to a "subject's capacity to utilize the characteristics and formats of the test and/or the
test-making situation to receive a high score."¹

**Importance of the Study**

As long as students are a part of the educational process, they will be subjected to some form of standardized testing in their academic careers. For this reason, it would be to the advantage of the student test-taker to become as learned and skilled about test-wiseness as possible.²

Knowing all the answers to test questions has not been necessary for achieving test results. But knowing **HOW** to select an answer to a question could effectively raise a student's test score. The importance of training all students to make wise and logical choices for answers could be an effective strategy for improving test results.

Another purpose of this study was to promote awareness among educators and administrators of the long-range benefits of these test-taking skills and its effect on students' future test-taking experiences.

---


²Irwin Bergman, "Can Students be Taught to Read and Take Tests More Effectively?" *Reading Improvement*, 17 (Winter 1980), 256-58.
Assumptions

The following assumptions were observed and recognized in this study:

1. This study was an exploratory study.

2. The teaching methods and practices of the two classroom instructors appeared to have the same and/or similar competencies.

3. There was no specific data available in the literature to explain why students score low (below grade placement) on the Science Research Associates (SRA) achievement tests.

4. Reasons for students scoring below grade placement on standardized tests were many. Among them were social class and status, parental education levels and expectations, and economics. The list for speculation was endless.

5. This study was evaluated with test-taking skills.

Limitations

This study was conducted in District 124, Grant Community High School, Fox Lake, Illinois. It had a basic limitation in that the subjects used were secondary students who were assigned to Reading-English 113 because their reading abilities were below grade level placement according to their test scores on the Science Research Associates (SRA) achievement tests. These students were assigned to these reading classes for the second semester of
the school term.

**Organization of the Study**

This study is organized in the following manner: Chapter Two reviews the literature pertaining to studies on test-taking skills, test-wiseness, and its influences on standardized testing results; Chapter Three outlines the procedures and methodology used in this study; Chapter Four presents and analyzes the data and the findings; and Chapter Five presents the summary and conclusions of this study and suggests recommendations for future research.
CHAPTER TWO
Review of Literature

The literature related to test-wiseness, test-taking skills, and standardized test scores was reviewed in this chapter. This review was organized to emphasize the various studies done on test-taking skills and their results and recommendations.

The Studies

No study on test-wiseness had been so frequently investigated, researched, and cited as the pioneer studies done by Millman, Bishop, and Ebel.\(^1\) These authors not only analyzed the process and procedures of testing and test-taking strategies, they also created the term, "test-wiseness," and gave it a meaningful definition for practical purposes and future studies. They defined "test-wiseness" as "a subject's capacity to utilize the characteristics and formats of the test and/or the test-making situation to receive a high score."\(^2\) These authors stated that "test-wiseness" was independent of the examinee's knowledge of the subject matter for which the items were supposedly

\(^1\) Millman et al., pp. 707-26.

\(^2\) Ibid.
measuring.¹

Few studies had been done on test-wiseness at various levels of education, although test-taking programs identified by Bergman have been present at every level of schooling from pre-school through secondary levels.² Bergman cited studies on pre-school (Head Start) programs through secondary school programs which concluded that, in general, test-wiseness had some value in helping facilitate student performances on standardized readiness tests.³ He also encouraged supervisors and administrators at all levels of schooling to re-examine existing practices in test-taking skills and arrange for further assistance for students whenever it should be warranted.⁴

For primary grades, studies conducted by Eakins, Green, and Bushell found that first grade students could be handicapped when taking a standardized test because these students lacked basic familiarity with test-taking skills.⁵ The results of this study indicated that gains

¹Bergman, p. 256.
²Ibid., p. 257.
³Ibid., pp. 257-58.
⁴Ibid.
were enhanced through the practice with the instructional test-taking unit. Students who received multiple instructional experiences produced greater gains than those students who received one or no instructional experience.¹

Taylor and White's three-part study included twenty-four classrooms of 597 second grade test-takers in test-taking skills. These students were trained for one to two weeks before the actual testing.² Although their findings were inconsistent with previous findings, these researchers noted that a difference was due to the length of training or the time factor involved before the actual testing dates.³ Overall, they reported that no statistically significant difference was found between trained and untrained second grade test-takers and their test scores.⁴

Teaching children how to take standardized tests had a positive effect upon their test performances according to Stewart and Green.⁵ They studied test-taking skills and the

¹Eakins et al., p. 67.


³Ibid., p. 206.

⁴Ibid.

⁵Oran Stewart and Dan Green, "Test-taking Skills for Standardized Tests of Reading," Reading Teacher, 36 (March 1983), 634.
application to tests of reading achievement and concluded that "reading from text was one thing, while translating this learning into test performance was another." They referred to other studies and concluded that:

[a] test may not be a valid indicator of a child's reading ability because children vary in their levels and uses of test-taking skills; thus, students vary in their own abilities to demonstrate their learning on a standardized test.¹

The effects of test-taking instruction and reading scores of fifth grade, inner-city students were studied by Jongsma and Warshauer. Their findings indicated that the students of this study who received instruction in test-taking benefited from the instruction because they scored higher than their counterpart control group on the standardized tests.² Although the gains were not significant, Jongsma concluded that test-wiseness existed and could be measured and test-wiseness should be a construction comprised of specific skills so that it should be taught effectively to students of all ages.³

The most dedicated longitudinal studies of test-wiseness were conducted by Crehan, Gross, Koehler, and Slakter. They observed three sets of subjects three times

¹Stewart and Green, p. 635.
²Jongsma and Warshauer, pp. 36-38.
³Ibid.
at two-year intervals. The first set of subjects consisted of seventy-five students in grades five, seven, and nine; the second set consisted of seventy-three subjects in grades seven, nine, and eleven; and the third set of subjects were sixty-four students in grades eight, ten, and twelve.\(^1\) The results of their research found that test-wiseness appeared somewhat stable. Although test-wiseness increased in subjects from grades ten through twelve, large differences in test-wiseness persisted into the high school grades.\(^2\) Crehan concluded with the recommendation that test-taking strategies be taught to all students, especially those identified as "low" in test-wiseness.\(^3\)

A critical analysis of the training of test-taking skills, experimental support for them and the experimental nonsupport for training test-taking skills was presented by Fuego. In her analysis, Fuego cited several studies which have identified and recommended strategies for test preparation, including coaching and practice lessons.\(^4\) According to Fuego, the effects of practice and coaching on various standardized tests had also been investigated and

\(^1\)Crehan et al., pp. 40-43.

\(^2\)Ibid., p. 43.

\(^3\)Ibid.

found to have had significant mean gains.\textsuperscript{1} She summarized her study with the following statements:

Experimental analysis of the factors influencing test performances found that significantly higher scores have been obtained under the following conditions:

1. when the answer sheets were easy to read;
2. when test instructions were clear;
3. when time limits permitted complete skill assessments;
4. when the same form of the test was taken repeatedly;
5. when the skills on one test were preparatory for those being measured on the next test taken;
6. when the tests were administered during the same month; and
7. when test-taking units were used in/with classroom instruction.\textsuperscript{2}

Fueyo concluded her analysis with a recommendation that a "task-specific instructional unit to train necessary skills for test-taking be used to assure that the score on the test is an accurate measure of the skill being assessed."\textsuperscript{3}

Test-wiseness principles have been successfully taught in a classroom setting and have lead to improved achievement test performances.\textsuperscript{4} Wahlstrom and Boersma evaluated the

\textsuperscript{1}\textit{Fueyo}, p. 182.

\textsuperscript{2}\textit{Ibid.}, pp. 182-83.

\textsuperscript{3}\textit{Ibid.}

effects of teaching test-taking strategies on 117 ninth grade students in two schools. Their purpose was to find out if a test-taking strategy taught by classroom instruction would enable a test-wise examinee to obtain a higher score than an equally knowledgeable examinee who lacked test sophistication.¹

Using a modified version of Millman's "Outline of Test-Wiseness Principles," Wahlstrom and Boersma's examinees were given four periods of twenty-five minutes each for a total treatment time of 100 minutes.²

The instruments used were the pre-test, post-test scores of the Verbal Reasoning Tests. These researchers concluded that the post-test mean score of the experimental group was significantly higher than the pre-test mean score of that group.³ They also summarized that the increase in the mean test score from pre-test to post-test for the experimental group appeared to be the result of an increase in test sophistication on behalf of the students.⁴ Differences between pre-test and post-test scores for the control group were nonsignificant.⁵

¹Wahlstrom and Boersma, pp. 413-14.
²Ibid.
³Ibid., p. 415.
⁴Ibid., p. 417.
⁵Ibid.
Test-wiseness was not a general trait, but rather a specific to the particular clue or cue under investigation as noted by the study done by Diamond and Evans. A total of ninety-five sixth graders were used as their subjects to determine if these students could obtain credit without knowledge of the subject matter being tested.

Diamond and Evans gathered information about each subject's intelligence quotient and achievement scores. These students were given fictitious material for study. These researchers concluded that these students were able to verbalize the principles of test-wiseness, and the utilization of the fictitious material was an effective mode for reliability exploring test-wiseness.

In another study, Diamond, Ayrer, Fishman, and Green investigated whether inner-city children possessed test-wiseness skills, and if they did, whether these skills were related to other cognitive abilities, such as those often measured by achievement tests.

Administering several vocabulary, reading

---


2 Ibid.

3 Ibid., p. 150.

comprehension, and language subtests, and the California Achievement Test to seventy-six sixth graders from an inner-city school, Diamond et al. concluded that these students were able to use selected cues to figure out answers to items about which they had no knowledge.\(^1\) They also noted that these students' scores did not seem to be related in any way to their performance on achievement subtests.\(^2\)

A number of researchers considered it important to determine the effectiveness of teaching test-wiseness to minority students. McPhail was among these researchers.

McPhail attempted to show that test-wiseness could be taught to minority student populations by a variety of methods, including programmed materials, lectures, supervision, and reading about test-wiseness cues.\(^3\) McPhail studied fifty-four talented twelfth-grade students from an inner-city school in Philadelphia. Ninety-six percent of these students were Black, while the other 4 percent were Puerto Rican.\(^4\) Using the Iowa Silent Reading Tests, a

\(^1\)Diamond et al., p. 43.

\(^2\)Ibid.


\(^4\)Ibid., p. 169.
test-wiseness curriculum, and a psycholinguistic cues curriculum, this investigation was conducted for forty-seven days. The time allowed for the treatment was approximately ninety-two minutes or two class periods daily.\(^1\) McPhail concluded that the differences between groups were not statistically significant, but there was evidence of treatment effects, especially in the significantly higher reading scores.\(^2\)

Improving test-wiseness through direct instruction in selected test-taking strategies was studied by Woodley.\(^3\) Fifteen different classes with a total of 259 college level students were pre-tested on a test-wiseness scale. Students were grouped into categories. The category of students who received the test-taking instructions was called the program group. Using a short-term approach with a combination of several test-taking strategies, outlines and programs, Woodley post-tested the groups and concluded that there were greater gains in scores for the program group which was given the short-term instruction.\(^4\)

\(^1\)McPhail, "A Psycholinguistic Approach to Training Urban High School Students in Test-taking Strategies, p. 173.

\(^2\)Ibid., pp. 175-76.


\(^4\)Ibid., pp. 15-16.
Shuller concluded that test-wiseness was an important source of variance on standardized tests, independent of content or random error.\textsuperscript{1} Using New York City Public Schools, Shuller made available strategies called Mini Tests for test-taking instructions to elementary schools, grades two through nine. Of the 621 elementary schools involved, 138 schools purchased these Mini Test programs while 483 schools did not. The Mini Tests consisted of ten to twenty instructional packages and were implemented within a four-to eight-week span.\textsuperscript{2} Using the M.A.T. for pre- and post-test results, Shuller concluded that students in the elementary schools which purchased and implemented the Mini Tests programs made significant gains in test scores.\textsuperscript{3}

In a follow-up study, Shuller again concluded that test-wiseness was an important source of variance on standardized tests.\textsuperscript{4} Using the same procedures from his previous study, Shuller selected the same 621 elementary schools, but 194 schools purchased the Mini Tests while the remaining 427 schools did not.\textsuperscript{5} Test scores from those

\begin{flushright}
\textsuperscript{1}Stephen Shuller, A Large-Scale Assessment of an Instructional Program to Improve Test-wiseness in Elementary School Students (ERIC ED 189 143), pp. 4-8.
\end{flushright}

\begin{flushright}
\textsuperscript{2}Ibid.
\end{flushright}

\begin{flushright}
\textsuperscript{3}Ibid.
\end{flushright}

\begin{flushright}
\textsuperscript{4}Ibid., pp. 8-10.
\end{flushright}

\begin{flushright}
\textsuperscript{5}Ibid., pp. 8-10.
\end{flushright}
students whose schools used the Mini Tests were significantly higher than those students whose schools did not purchase the programmed test-taking skills.¹

Jongsma and Warshauer concluded that teaching test-wiseness and test-taking skills can improve students' test scores in inner-city public schools, while producing significantly higher test scores in a suburban public school.² These researchers implemented the same study with two different fifth grade elementary classes. Both classes were pre-tested with the California Test of Basic Skills and post-tested with the Standard Achievement Test. One class consisted of fifty-four black, inner-city school students; the other class consisted of ninety-one white, suburban school students. The "Test-taking Tips" program was implemented. This program was one hour and fifteen minutes in length of training.³ Although the mean scores for the inner-city school students was not significant, the raw scores were 1.5 to 6 points higher, and this appeared to be educationally meaningful.⁴

The post-test scores for the ninety-one white suburban students provided significantly higher scores.⁵ Jongsma and

¹Shuller, pp. 8-10.
²Jongsma and Warshauer, pp. 32-36.
³Ibid.
⁴Ibid.
⁵Ibid.
Warshauer concluded that the highly significant results of the suburban class may have been influenced by the teaching methods of the different teachers.\(^1\) They also encouraged future studies in test-wiseness, especially those studies that would discover information to answer the following:

1. Do all students need training in test-wiseness?
2. What is the best method for teaching test-wiseness?
3. Can test-wiseness produce long-term results?\(^2\)

The Maryland State Department of Education published six booklets to help students acquire test-taking skills. These booklets also assisted teachers in acquiring skills which related to the scope of the curriculum.\(^3\) The researchers' purpose for implementing these booklets with classroom instruction was that the test-taking skills learned and practiced would assure a more accurate assessment of student achievement.\(^4\)

---

\(^1\) Jongsma and Warshauer, pp. 25-30, 32-36.

\(^2\) Ibid.

\(^3\) Maryland State Department of Education, Improving Student Attitudes and Skills for Taking Tests (ERIC ED 128 352), pp. 2-8.

\(^4\) Ibid.
Summary

Chapter Two presented a review of the recent literature on test-taking skills and test-wiseness. Chapter Three will present the methodology, population, and procedures used for this study.
CHAPTER THREE

Methodology

Introduction

The purpose of this study was to find out if teaching test-taking skills to secondary students whose reading abilities were below grade level would increase reading comprehension and vocabulary scores on standardized tests. In this chapter, the population studied, instruments used for testing, instruction time, analysis, treatment of the data, and a summary are presented.

Population

The population studied was secondary students assigned to Reading-English 113, for the second semester of the 1984-1985 school term in District 124. These students were placed in Reading-English classes because of low Science Research Associates (SRA) achievement scores.

A population of thirty-six students was pre-tested. Due to transfers and withdrawals, a population of thirty-three students was post-tested.

Procedures

All secondary students assigned to Reading-English 113 for the second semester of the 1984-1985 school term were
pre-tested with the Nelson-Denny, Form E, Reading Test, which was a vocabulary and reading comprehension standardized test. A random selection of groups was made; the control group, identified as Group 2, received no specific program or unit in test-taking skills; the experimental group, identified as Group 1, received specific program instruction in test-taking skills using teacher-selected materials from the 1982 World Book of Test-Taking Skills: Standardized Tests, Volume 2,¹ and the 1985 World Book Dictionary.²

The experimental group received thirty-six, twenty-minute sessions during their regular class time using instructional materials as described in the previous paragraph. Special permission was obtained from the publisher to reproduce these materials for classroom use (Appendix). The following pages were reproduced and used:

464-473 - Reading Comprehension
504-511 - Word Play: Vocabulary
537-538 - Self-test Answer Key
701-714 - Important Roots to Know³
39-63 - Vocabulary Inventories⁴


Instruction included lectures and discussions on following directions for standardized tests and filling out programmed answer keys. The first two hours of instruction concentrated on lectures, discussions, and actual practice in following test directions and completing answer keys. The next five hours of test-taking instruction concentrated on vocabulary building, word recognition, and word usage skills materials. The last five hours of test-taking instruction concentrated on reading comprehension skills material.

At the end of the thirty-six sessions (twelve hours) of test-taking instruction for Group 1, all students from both Groups 1 and 2 were post-tested with the Nelson-Denny, Form F, vocabulary and reading comprehension tests.

The control group used classroom time for daily reading assignments. No time was given to this group on specific test-taking skills.

Students' post-test scores were used for the data analysis. A t-test was applied to the scores. The results of this statistical study determined if twelve hours of test-taking instruction made a significant difference between means at the .05 level of significance.

Instruments Used

The instruments used in this study were the 1981 Nelson-Denny Reading Test, Forms E and F. The authors of this reading achievement test were Brown, Bennett, and
Hanna. According to these researchers,

The primary purpose of the Nelson-Denny Reading Test, Forms E and F, was to provide a trustworthy ranking of student ability in three areas of academic achievement: reading comprehension, vocabulary development, and reading rate.¹

For this study, the reading comprehension and vocabulary development aspects and scores of this test were used.

The Nelson-Denny Reading Test was composed of two subsets, vocabulary and reading comprehension. The vocabulary section consisted of 100 items, each with five answer choices. This section had a time limit for completion of fifteen minutes.²

The second section of this test was reading comprehension. This section contained eight reading passages. It had a total of thirty-six questions with five answer choices each. The time allotment for this portion of the test was twenty minutes. The total administrative time for the entire test, including answer sheet preparation, was no longer than forty-five minutes.³

According to the authors,


²Ibid.

³Ibid., p. 1.
Forms E and F have been statistically equated and can be used as a pre-test and post-test. Such use provides a measure of student growth and is of value to teachers in determining effectiveness of instruction.\footnote{Brown et al., p. 1.}

Table 1 presents the descriptive and reliability data for Forms E and F for vocabulary and comprehension sections for grades nine through twelve. The standard deviation of these samples generally coincided quite closely with those of the national standardization sample.\footnote{Ibid., pp. 15-17.} The validity studied by Hanna presented context or passage dependence as the kind of validity evidence that was most important.\footnote{Ibid.} The context dependence index findings were as follows:

- Grade 10, Form E \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots .26
- Grade 10, Form F \ldots \ldots \ldots \ldots \ldots \ldots \ldots \ldots .32^4

The median alternate-form reliability coefficient of the vocabulary score was .92. The median across-grades, alternate-form reliability coefficient for the comprehension score was .77. The median for the Total Score was .91.\footnote{Ibid., pp. 15-17.}
Table 1

Nelson-Denny Reading Test, Forms E and F

Descriptive and Reliability Data for Equating Samples

<table>
<thead>
<tr>
<th>Grades</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>409</td>
<td>438</td>
<td>387</td>
<td>343</td>
</tr>
<tr>
<td>r</td>
<td>.89</td>
<td>.89</td>
<td>.91</td>
<td>.93</td>
</tr>
</tbody>
</table>

Form E - Vocabulary

| Mean   | 27.6 | 32.8 | 34.4 | 38.9 |
| SD     | 13.5 | 15.7 | 15.9 | 19.0 |

Form F - Vocabulary

| Mean   | 27.9 | 32.7 | 34.5 | 30.4 |
| SD     | 13.0 | 14.6 | 15.1 | 18.2 |

<table>
<thead>
<tr>
<th>Grades</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>410</td>
<td>438</td>
<td>386</td>
<td>339</td>
</tr>
<tr>
<td>r</td>
<td>.75</td>
<td>.76</td>
<td>.76</td>
<td>.82</td>
</tr>
</tbody>
</table>

Form E - Reading Comprehension

| Mean   | 15.4 | 17.2 | 17.6 | 19.3 |
| SD     | 6.7  | 6.8  | 6.9  | 7.4  |

Form F - Reading Comprehension

| Mean   | 15.5 | 17.5 | 18.3 | 20.0 |
| SD     | 6.8  | 7.0  | 7.3  | 7.6  |

The authors summarized this test as follows:

Reliability findings of these magnitudes by the conservative alternate-form with time interval method might have suggested that the vocabulary, total score, and possibly comprehension scores were sufficiently accurate for prudent professional use with either individual students or group data.

Data Collection

Thirty-six students were assigned to Reading-English 113 for the second semester of the school term. These students were pre-tested with the Nelson-Denny, Form E, in reading comprehension and vocabulary sections. Students were randomly assigned to groups. The experimental group, Group 1, received twelve hours of test-taking skills instruction. The control group, Group 2, did not receive the specific test-taking skills instruction. This group used classroom time using regular reading skills assignments not related to test-taking skills.

Due to transfers and withdrawals, thirty-three students completed this study. Using the thirty-three students' pre-test and post-test scores in vocabulary and reading comprehension, a format was designed for the data entry. The Statistical Package for Social Science, SPSSX, a system of computer programs for analysis, was used to determine the data designs needed to provide evidence in order to test the

---

1Brown et al., p. 15.
hypotheses.

Statistical analysis using rank of scores, means, paired t-tests, group t-tests, and t-values was applied to Hypotheses 1 and 2.

The analysis and presentation of this study were the following:

1. To determine what effect the twelve hours of test-taking instruction had in improving the students' vocabulary and reading comprehension scores on standardized tests; and

2. To determine if test-taking skills were a significant factor in achieving higher scores on standardized tests.

Summary

This chapter presented a discussion of the study including procedures, instrumentation, data collection, and treatment of the data.

Chapter Four will present an analysis of the data received from the post-test scores of the Nelson-Denny, Form F, test.
CHAPTER FOUR

Findings and Interpretations

The data presented in this chapter represents the results of the statistical analysis on the raw scores from the pre-test and post-test scores.

The tables are divided into three groups. Tables 2 and 3 present the findings of a paired t-test based upon the raw scores from the vocabulary pre-tests and the vocabulary post-tests. Tables 4 and 5 present the findings of a paired t-test based upon the raw scores from the reading comprehension pre-tests and the reading comprehension post-tests. Tables 6 and 7 present the findings of a group t-test based upon the actual gains in mean score values for Groups 1 and 2.

The following definitions explain the descriptions of the headings and tables:

Experimental Group: Those secondary students pre-tested and randomly assigned to a test-taking skills class. This group was also labeled "Group 1."

Control Group: Those secondary students pre-tested and randomly assigned to a non-test-taking skills class. This group was also labeled "Group 2."
Group t-Test: A statistical test on independent samples which divides cases into two groups and compares group means on a single variable. This test provides the information needed to distinguish if the two independent samples' means are significant at the .05 level.

Paired t-Test: A statistical test on a paired sample which is the comparison of the pre- and post-test scores. This test provides the information needed to distinguish if sample means are significant at the .05 level.

Gains: The gain score is the difference between the group's mean score on the post-test minus the pre-test group score. These mean scores provide valuable information at the .05 level of significance.

Pre-test Scores: All scores referring to the Nelson-Denny Reading Test, Form E. These scores were used as pre-test measures.

Post-test Scores: All scores referring to the Nelson-Denny Reading Test, Form F. These scores were used as post-test measures.

Totals: The total population of this study is thirty-three secondary students. The experimental group, also referred to as Group 1, includes seventeen of these students. The control group, also referred to as Group 2, includes sixteen students.

The data in Table 2 presents a paired t-test and the mean scores for the seventeen students in Group 1 on the
vocabulary pre-test and the vocabulary post-test. Group 1 shows a mean difference of 4.7059.

Table 2
Paired t-Test Analysis for Experimental Group, Group 1 - Vocabulary Scores (n=17)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Dev.</th>
<th>Mean Diff.</th>
<th>Standard Dev.</th>
<th>T-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Test</td>
<td>34.7647</td>
<td>9.47</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Post-Test</td>
<td>39.4706</td>
<td>9.799</td>
<td>4.7059</td>
<td>11.973</td>
<td>1.62</td>
</tr>
</tbody>
</table>

*Critical t-value at the alpha level .05 = 1.746.

The data in Table 3 presents a paired t-test and the mean scores for the sixteen students in Group 2 on the vocabulary pre-test and the vocabulary post-test. Group 2 shows a mean difference of 1.125.
Table 3

Paired t-Test Analysis for Control Group, Group 2 - Vocabulary Scores (n=16)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Dev.</th>
<th>Mean Diff.</th>
<th>Standard Dev.</th>
<th>T-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Test</td>
<td>27.6875</td>
<td>9.877</td>
<td>1.125</td>
<td>9.50</td>
<td>0.47</td>
</tr>
<tr>
<td>Post-Test</td>
<td>28.8125</td>
<td>7.503</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Critical t-value at the alpha level .05 = 1.753.

The data in Table 4 presents a paired t-test and the mean scores for the seventeen students in Group 1 on the reading comprehension pre-test and the reading comprehension post-test. Group 1 shows a mean difference of 6.5294.

Table 4

Paired t-Test Analysis for Experimental Group, Group 1 - Reading Comprehension Scores (n=17)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Dev.</th>
<th>Mean Diff.</th>
<th>Standard Dev.</th>
<th>T-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Test</td>
<td>13.2941</td>
<td>6.282</td>
<td>6.5294</td>
<td>6.811</td>
<td>3.95</td>
</tr>
<tr>
<td>Post-Test</td>
<td>19.8235</td>
<td>7.029</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Critical t-value at the alpha level .05 = 1.746.
The data in Table 5 presents a paired t-test and the mean scores for the sixteen students in Group 2 on the reading comprehension pre-test and the reading comprehension post-test. Group 2 shows a mean difference of 1.3750.

Table 5

Paired t-Test Analysis for Control Group,
Group 2 - Reading Comprehension Scores
(n=16)

<table>
<thead>
<tr>
<th></th>
<th>Mean</th>
<th>Standard Dev.</th>
<th>Mean Diff.</th>
<th>Standard Dev.</th>
<th>T-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-Test</td>
<td>11.875</td>
<td>5.818</td>
<td>1.375</td>
<td>5.608</td>
<td>0.98</td>
</tr>
<tr>
<td>Post-Test</td>
<td>13.250</td>
<td>4.219</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Critical t-value at the alpha level .05 = 1.753.

The data in Table 6 refers to the group t-test and the overall mean score gains in vocabulary for Groups 1 and 2.

Table 6

Group t-Test Analysis for Mean Gain Scores in Vocabulary

<table>
<thead>
<tr>
<th>Number of Cases</th>
<th>Mean</th>
<th>Standard Dev.</th>
<th>T-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>17</td>
<td>4.7059</td>
<td>11.973</td>
</tr>
<tr>
<td>Group 2</td>
<td>16</td>
<td>1.1250</td>
<td>9.50</td>
</tr>
</tbody>
</table>

*Critical t-value at the alpha level .05 = 1.697.
The data in Table 7 refers to the group t-test and the overall mean score gains in reading comprehension for Groups 1 and 2.

Table 7

Group t-Test Analysis for Mean Gain Scores in Reading Comprehension

<table>
<thead>
<tr>
<th></th>
<th>Number of Cases</th>
<th>Mean</th>
<th>Standard Dev.</th>
<th>T-value*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>17</td>
<td>6.5294</td>
<td>6.811</td>
<td>2.36</td>
</tr>
<tr>
<td>Group 2</td>
<td>16</td>
<td>1.3750</td>
<td>5.608</td>
<td></td>
</tr>
</tbody>
</table>

*Critical t-value at the alpha level .05 = 1.697.

The results produced in this study presented one hypothesis to be significantly different at the .05 level. The overall mean average score of 6.594 and the t-value of 2.38 in Group 1's reading comprehension results indicated this finding. Thus, Null Hypothesis 2 was rejected when applied to the mean difference between the two groups and within Group 1.

According to this study's findings, Null Hypothesis 1 was held tenable.
Null Hypotheses

1. There was no significant difference in vocabulary test scores of those secondary students whose reading abilities were below grade level when trained in a specific test-taking skills program and those below grade level reading students who were in a control group.

2. There was no significant difference in the reading comprehension test scores of those secondary students whose reading abilities were below grade level when trained in a specific test-taking skills program and those below grade level reading students who were in a control group.

Summary

Chapter Four presented the analyzed data. The summary, conclusions, and recommendations of this study are reported in Chapter Five.
CHAPTER FIVE
Summary, Conclusions, and Recommendations

This study was designed to obtain information on the influence of test-taking skills and test-wiseness on standardized test scores.

Secondary students were pre-tested and randomly assigned to two groups: experimental group, Group 1, and control group, Group 2. A total of thirty-three secondary students participated and completed this study.

This study was designed to obtain data to answer these questions:

1. Can test-taking skills and test-wiseness influence and improve a student's score on a standardized vocabulary test?

2. Can test-taking skills and test-wiseness influence and improve a student's score on a standardized reading comprehension test?

Descriptive and inferential statistics were used for the analysis of data. Paired t-test and group t-test procedures were used to test Null Hypotheses 1 and 2 for differences between the groups' mean values at the .05 level of significance.
Conclusions

Two null hypotheses were tested.

1. There was a significant difference in vocabulary test scores of those secondary students whose reading abilities were below grade level when trained in a specific test-taking skills program and those below grade level reading students who were in a control group.

The results of the paired t-test and group t-test procedures reported in Tables 2, 3, and 6 presented no significant differences at the .05 level. Thus, Null hypothesis 1 was held tenable for this study.

2. There was no significant difference in reading comprehension test scores of those secondary students whose reading abilities were below grade level when trained in a specific test-taking skills program and those below grade level reading students who were in a control group.

The results of the paired t-test and group t-test procedures reported in Tables 4 and 7 presented significant differences at the .05 level. Thus, Null Hypothesis 2 was rejected for this study.

Consideration must be given to the possibility that the differences between the two teachers, two classrooms, organizational procedures and settings may have introduced factors other than those being tested in this study.

Recommendations

The results of this study led this writer to two types of recommendations: first, those recommendations for school use; and second, those recommendations for future studies.
Recommendations for school use were the following:

1. All students should be given continual instruction and training in test-taking skills and test-wiseness.

2. When implementing such a program or course of study, teachers should be trained to use the same methods and materials for instruction.

3. Teachers composing teacher-made tests should apply the principles of test-wiseness.

Recommendations for future studies were the following:

1. Continued research should be investigated in locating and/or creating other methods, materials, and procedures for promoting the learning of test-taking skills and test-wiseness.

2. Length of time, short-term or long-term training, should be a major consideration for other studies.

3. Research should be continued, a la McPhail and others, on the application of teaching test-wiseness skills to members of minority groups.

4. Replication of this study in other locales with different subjects, to determine if both learning and test-wiseness can be improved, should be encouraged.
BIBLIOGRAPHY
BIBLIOGRAPHY

Books


Periodicals


ERIC Documents

Ford, Valeria. *Everything You Wanted to Know About Test-wiseness.* ERIC ED 093 912.


Shuller, Stephen. *A Large-Scale Assessment of an Instructional Program to Improve Test-wiseness in Elementary School Students.* ERIC ED 189 143.


Dissertations

APPENDIX
June 20, 1985

Ms. Diane Vida
c/o Nickelson
613 Winegardner Road
Des Moines, IA 50317

Dear Ms. Vida:

To confirm our telephone conversation of June 18, you wanted some validation for the 'Vocabulary Inventories' and material from The World Book of Test Taking which we gave you permission to use for your dissertation.

Enclosed is a copy of the "Preface" and "Introduction" to The Living World Vocabulary by Professor Edgar Dale and Dr. Joseph O'Rourke. Information in this publication was used as the basis for the inventories, and these few pages should provide you with sufficient background.

For the development of The World Book of Test Taking, we consulted with two advisers who had expertise in this area. Freelancers were involved in the research and writing of the text, as well as our own staff. A copy of the staff page is enclosed for reference.

We hope this information meets your needs. If you have any further questions, please let me know.

Sincerely,

Janet T. Peterson
Permissions Editor

JTP:pjm
Enclosures
Preface

To prepare superior written materials for any purpose, two conditions are required. First, there must be effective communicators—writers and editors who can create clear, compelling copy. Second, there must be a source of available information on the familiarity of the words and concepts to be used. There must be an almost perfect match between the reader and the material to be read.

Talented writers do exist, but until the appearance of The Living Word Vocabulary list of tested words and terms, we have had little data on the familiarity of most of the words and concepts used. Instead, we have had to depend upon personal judgment and nonsemantic frequency studies of words to provide this information—an inadequate and somewhat unreliable approach. Professor Edgar Dale and Dr. Joseph O'Rourke, coauthors of The Living Word Vocabulary, believe that the data presented in this work will mark the beginning of a revolutionary approach to the preparation and presentation of materials that fit not only the reading abilities, but the experience and background of the learner as well.

This word list includes more than 44,000 items—each identified by a simple word meaning. Unlike any other word list, it also presents a percentage score on those words or terms familiar to students in grade levels 4, 5, 6, 7, 8, 9, 10, 11, 12, and 16. Each word-familiarity score was obtained by administering a three-choice test to students from schools and colleges throughout the United States. This testing began in 1954 and is still in progress. A significant number of items in this volume were updated within the current year.

Test forms were arranged for computerizing, and as the data for the words became available, writers and editors for World Book—Childcraft International, Inc., were provided with these word scores. This information is used by editors of The World Book Encyclopedia as one of several checks on the reading level of World Book manuscripts. Since World Book Encyclopedia articles are also field tested prior to publication, a continuing check on the validity of the word scores is assured.

The Living Word Vocabulary can be an invaluable resource to writers, editors, teachers, and librarians in determining the grade level at which the specific meaning of a word can be readily understood. During a recent reading workshop Dr. Edgar Dale, coauthor of the Dale-Chall Readability Formula said, "We overestimate the familiarity of words selected . . . This situation is most likely to occur when teachers or writers work without objective data on the familiarity of words selected at specific grade levels." The Living Word Vocabulary makes it easy to find the right word . . . at the right time . . . for the right person.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score</th>
<th>Word</th>
<th>Word Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>78%</td>
<td>abrupton</td>
<td>a sudden breaking off</td>
</tr>
<tr>
<td>08</td>
<td>71%</td>
<td>abscess</td>
<td>wound with pus</td>
</tr>
<tr>
<td>12</td>
<td>31%</td>
<td>abscond</td>
<td>to cut apart</td>
</tr>
<tr>
<td>16</td>
<td>72%</td>
<td>abscissa</td>
<td>horizontal coordinate</td>
</tr>
<tr>
<td>16</td>
<td>84%</td>
<td>abscond</td>
<td>run away and hide</td>
</tr>
<tr>
<td>04</td>
<td>67%</td>
<td>absence</td>
<td>being away</td>
</tr>
<tr>
<td>08</td>
<td>91%</td>
<td>absence</td>
<td>not having something</td>
</tr>
<tr>
<td>04</td>
<td>84%</td>
<td>absent</td>
<td>not here</td>
</tr>
</tbody>
</table>

Actual sample showing type size and content
The number before each word in the list (see sample words) tells at what grade level that percentage of students understand the word in terms of the word meaning, if The Living Word Vocabulary indicates that a necessary word is too difficult for the grade level of the audience to whom the written material is being directed, substitute a suitable synonym or use the word and define it in context.

A word of a certain grade level may be used without definition in an article for use by readers one grade below that level. Thus, a 10th grade word may be used for 9th graders and a 6th grade word may be used for 5th graders.

Because many words have more than one meaning the user of The Living Word Vocabulary is cautioned to be certain that the exact meaning is determined before checking the assigned grade level. The Living Word Vocabulary is in no way a substitute for the traditional dictionary, and the word meaning list is not, in most instances, a proper synonym for readability purposes.

World Book-Childcraft International, Inc., is pleased to publish this important and significant work by Edgar Dale and Joseph O'Rourke. The Living Word Vocabulary has been an invaluable editorial tool for our World Book staff, and it is in daily use as new World Book material is prepared and edited. By publishing the Dale/O'Rourke analysis we continue World Book's commitment to effective readability and better written communication.

W. H. Nault
Editorial Director
The World Book Encyclopedia
Introduction

This 25-year study of vocabulary provides objective, familiarity scores on 44,000 word meanings. The Living Word Vocabulary may be used in many ways. With this list writers now can find the familiarity scores of the words they are using. If the score indicates that a word is unfamiliar, the writer can explain it either directly or indirectly, or remove it. In the past we have had extensive word lists by E. L. Thorndike, The American Heritage Word Frequency Book, and others solely on the basis of frequency in literature, textbooks and other writing. This is the first time we have had an extensive list of word meanings and their familiarity scores.

Persons preparing directions for filling out official applications and blanks of various kinds can now check the familiarity of words for intended readers. Public documents prepared for universal consumption can be written at the sixth to eighth grade level instead of at the tenth to twelfth, as is now common. Writers of textbooks preparing material for stated grade levels can turn to The Living Word Vocabulary for guidance. For example, writers of articles for The World Book Encyclopedia use this list to guide them in producing readable copy at varied age and grade levels.

Test makers can be sure that the words in their tests are not too hard (or too easy) for prospective users. Vocabulary tests can be made to fit the persons being tested. In short, materials of instruction and information can be tailor-made.

In presenting this list of 44,000 words and their familiarity scores we can understand what Samuel Johnson meant in his "Preface to the Dictionary." He notes the possible presence "of a few wild blunders, and rash absurdities, from which no work of such multiplicity was ever free, [and] may be for a time furnish folly with laughter." He points out that "no dictionary of a living tongue ever can be perfect, since while it is hastening to publication, some words are budding, and some falling away; that a whole life cannot be spent upon syntax and etymology, and that even a whole life would not be sufficient." (Johnson's Dictionary—A Modern Selection. Pantheon Books, 1963, p. 28.)

This list provides an inventory of the written words known by children and young people in grades 4, 6, 8, 10, 12, 13, and 16. It discloses what they know and don't know. It means, in part, the definition of language given by Jerome S. Bruner, Jacqueline J. Goodnow and George A. Austin in A Study of Thinking (New York: John Wiley & Sons, 1956, p. 311): "For language is nothing less than inventory of all the ideas, interests, and occupations that take up the attention of the community. The study of semantics, in this extended sense, cannot be distinguished from the general study of culture."

There were three stages in the development of The Living Word Vocabulary. Our first concern was to develop a list of common terms used widely in all writing: "A Comparison of Two Word Lists," (Educational Research Bulletin, December 9, 1911, Vol. X, No. 18, pp. 484-87). These 769 words appeared (1) in the international Kindergarten Word List of 2596 words (International Kindergarten Union, Child Study Committee, A Study of the Vocabulary of Children before Entering the First Grade. Washington, D.C.: Association for Childhood Education, 1928), and (2) in the first 1,000 words of E. L. Thorndike, The Teacher's Word Book (New York: Teachers College, Columbia University, 1921).

Clarence R. Stone revised the Dale 769-word list by removing 173 words and adding 173 words. Here are some of the words he removed: bank, blind, blood, body, bone, born, born, coal, cool, dead, dozen, finger, mis, tell, ship, silver, sit.

He also added such children's words as bow-wow, chick, cock-a-doodle-doo, pet, moo, puppy, iced, doll, racetr, km, rabbit.

* Haughton Mifflin Company, Boston, MA 02167.
The 769-word list was used by Irving Lorge in "Predicting Readability," "The American College Record," Vol. 45, No. 6, pp. 404-419, March 1944.

E. W. Dolch also revised the 769-word list, extending it to 1,000 words. (Problems in Reading, Garrard Press, 1948, pp. 108-129.) This list was classified into categories dealing with: The Child’s Person, Home Environment, School, Nature, War and Fighting, General Things, Qualities and Relationships, Miscellaneous Parts of Speech, Verbs. This classification makes the meanings much clearer and does not depend primarily upon a frequency score.

Twenty years later, George Spache added 361 words to the revised Stone word list and removed 87 words, making a total of 1,041 words. (Good Reading for Poor Readers, revised 1974, Garrard Publishing Co., pp. 195-207.) Some of the words added by Spache were from an unpublished list developed by Edgar Dale and Emily Schuh titled, "A List of 1400 Words Known by 75% or More of First Grade Children in the Enrichment Program of the Columbus (Ohio) Public Schools," March 1970.

Next we developed a list of 8,000 common words and used student judgment tests to check their familiarity. There were data to indicate a high correlation between student judgment and the familiarity of the word. We asked students in grades 4, 6, and 8 to check whether they knew the words which we presented to them in list form. The procedure worked well with easy words and with hard words, but not as well with the words in between. Further, it does not adequately take into account the large number of simple words which have multiple meanings, although we did present some of the words in short phrases. Students were reacting to words whose spelling sometimes confused them. Thus title was stated as known because some thought it meant stomach. Euphem was confused with ute.

A second approach was the use of pupil judgment involving having children in the fourth grade read Black Beauty and underline the unknown words. We got some valuable data to use in our 3,000-word list, but the findings lacked needed accuracy. It did give a useful indication of familiarity.

A third approach involved placing The Brownie Converse Dictionary in the hands of college freshmen at The Ohio State University and Stephens College and asking them to read stated portions of the dictionary and check unknown words. This list proved valuable in discovering “hard” words but did not give us data on easy- and middle-level words. Here are samples of words checked as unknown by many of these college freshmen: aerosol, antipathy, aqua, crescent, effusion, agouti, aggregates, agnostic, alleles, aliquot, alluvial, aliquot, amphora, amphibious, anachronism.

It became clear that we would have to turn to a more rigorous form of testing, namely a multiple-choice test. One of the writers had explored the nature and possibilities of such a test in his doctoral dissertation titled, “Factual Basis for Curriculum Revision in Arithmetic with Special Reference to Children’s Understanding of Business Terms.” (Edgar Dale, University of Chicago, 1929.) First of all, it involved an analysis of the nature and frequency of business terms in investment literature written for the layman. We found 2,276 terms and tested the 200 most common ones using a four-choice multiple choice test.

What is a word and what determines its meaning?

In this study we have assumed that a word is a unit of meaning presented in a dictionary with its semantic variations. Some words in the dictionary are given only one clear-cut meaning, other words may have ten to twenty meanings. In The World Book Dictionary, for example, there are 23 meanings of good. The Shorter Oxford English Dictionary presents more than one hundred variant meanings of the word go.

The definitions used for the words presented here came chiefly from The World Book Dictionary. Many dictionaries, however, were consulted to find the number of different meanings of certain words and to note especially simple, clear-cut definitions. Some word meanings are our own, sometimes involving a shortened version of the longer dictionary entry.

When is a word known?

Knowledge of a word can be placed on a continuum, starting with “I never saw or heard the word before.” For example, the reader of this introduction does not know the words biotic, pointilist, or jambalaya because they do not exist as meaningful words. Second, we may say, “I know there is such a word but I don’t know what it means.” Such words might be jigger-mugger, adstrum, adstrum, P.E. ratio, secondstory. A third stage is “the contextual placing of the word. You know you make the written ring and that hussies has to do with elections, but what do these words mean specifically? You bask in the sun. Can you bask in the shade? What are kirk and kent? What is a pind paper? Was President Andrew Johnson impeached? (He was.) These words are in our “twilight zone.”

Fourth, and finally, we reach the stage where we have pinned the word down. We know it. We would recognize it again if we saw it, and we are likely to remember it. This is the level at which we have tried to test our words.

What about multiple meanings of words?

When the list of entries was established, the next problem was the selection of important multiple meanings. The high frequency words have many meanings. The selections of the different meanings to be tested were usually made by the late Lynton Bercer, a student of words, who analyzed various dictionaries and selected the multiple meanings of key words. He spent much of his time—for ten years—working on this task.

For example, the word pery has more than 40 separate definitions in The World Book Dictionary, and with com...
pound words and idioms occupies a full page. We tested the following meanings:

- 4.76% point = a compass direction
- 6.81% point = a place
- 6.81% point = an important idea
- 4.82% point = a mark or dot
- 8.72% point = a certain stage
- 4.71% point = to show with finger
- 4.76% point = to aim
- 0.75% point = a unit of scoring
- 4.77% pointed = sharp-ended
- 12.47% needle-point = a kind of lace

Few major studies of the frequency of the occurrence of multiple meanings of a word have been made. Professors Lorge and Thorndike, with the aid of a staff of 270 persons, counted the frequency of the occurrence of different meanings of a sample of about four and a half million words. (Irving Lorge, "The English Semantic Count," Teachers College Record, 39:65-77, October 1937. Irving Lorge and Edward L. Thorndike, A Semantic Count of English Words. New York: The Institute of Educational Research, Teachers College, Columbia University, 1938.)

What about the inflections of words?
The standard for testing inflections was that: If the inflection is likely to provide a variant score, we should test it. We usually did not test regularly inflected words—the addition of -ed, ing, fy, etc. But we considered retesting if the visual pattern of the word was irregularly changed in the inflecting, for example, substantize and substantive. Lengthening the word did tend to make it "harder" and in the lower grades inflection may make the word less familiar since the affixes might not be known.

Persons and Places?
When we discuss the vocabulary of an individual we are usually thinking about common nouns, not proper nouns. Yet there is much significant learning tied to the names of persons, trademarked products (cereals, automobiles), the states in the Union, the Presidents of the United States, the countries of the world, characters in literature. In the past twenty-five years we have been busy learning new words such as Zaire, Botswana, Namibia, Malagasy, Malawi—all in Africa and increasingly in the news.

Most scores of persons and places were low. We discovered that names of persons and places will only be known when they are specifically taught in school or have appeared repeatedly in the news over time. We tested only a few trademarked items.

A limited number of proper nouns appear in this list. However, we have additional test scores on about five thousand such words and may publish them later.

How were the words tested?
We gave up hope of developing an inexpensive, quick method for determining word knowledge and turned to a multiple-choice test. We chose a three-choice multiple-choice test for several reasons. Amos Tversky has pointed out in the Journal of Mathematical Psychology, 1964 that:

Whenever the amount of time spent on the test is proportional to its total number of alternatives, the use of three alternatives at each choice point will maximize the amount of information obtained per time unit . . . (p. 390).

Clearly, in the writing and testing of some 44,000 different items this factor of saving time is critical. Tversky also says:

There exists some empirical evidence, based on the study of auto-instructional items, which indicates that three-alternative test items are indeed optimal . . . (p. 390)

Taken together with Garner's conclusions concerning human capacities to process multi-dimensional information, these results suggest that the use of three levels per dimension may be the most efficient way to code and process information (p. 391).

How did we know what words to test at what levels?
We had access to Gesell's Spelling Difficulties in 3,876 Words, which gave us a familiarity score on each of those words. (Bureau of Publications, Teachers College, Columbia University, New York, 1937.) We also had access to the unpublished study by Paul Dietrich and Osmond Palmer. They did multiple-choice testing of over 4,700 words in grades 11 and 13. These were chosen from the Thorndike word list. Out of these 4,700 words in the 11th grade list about one-third were known (67 per cent or better). However, over one-half of the 13th grade words were known (67 per cent or better). In addition, we made extensive use of the graded vocabulary list included by B. R. Buckingham and E. W. Dolch in A Combined Word List (Ginn and Co., 1936).

Our aim was to secure a score on each word of at least 67 per cent and not more than 84 per cent. If a word got a score of 66 per cent in the 4th grade, we tested it again in the 5th grade. If a word in the 5th grade got a score of 85 per cent or above, we tested it in the 4th grade. Thus most words in this list have scores of 67 per cent or above. However, we did not retest all words in grade 11 which got less than 67 per cent—usually from lack of time.

We tested in grades 4, 6, 8, 10, etc. and did not test in the intervening grades—5, 7, 9, 11—because this would have almost doubled the cost of the project. However, the scores in the missing grade levels can be inferred from the fact that between the two tested grades there is usually an average difference of 20 per cent on word scores. Hence scores between single grade levels are likely to differ by about 10 per cent. Testing words in the 5th and 7th grades would have been useful because the child is then learning to read well and his vocabulary is growing rapidly. Many of the harder words at the 12th, 13th, and 16th grade level have scores
Introduction

below 67 per cent. Some are being tested as a part of our continuing word testing program.

We offer a note of caution about words with scores below 50 per cent. More chance would give a score of 33 per cent, which might mean no knowledge of the word at all. Thus, a word with a score of 50 per cent or less is generally a hard word.

Should we test the word in some kind of context?

Unfortunately, context can vary from as 

help to complete help 

in revealing the meaning of a word. So when someone says, "Test the word in context," we must ask: "What context?"

Victor H. Kelley compared the scores received when he correlated the scores of students who used the word in a sentence and those from matching or multiple choice tests, and concluded: "The multiple choice and matching tests appear to be the best technique with no real difference existing between the two." (Victor H. Kelley, "An Experimental Study of Certain Techniques for Testing Word Meanings," Journal of Educational Research, Vol. 27, December 1933, pp. 277-282.)


Construction of Multiple-Choice Items

In constructing the three-choice tests, the critical problem lies in selecting the two incorrect choices, the distractors. We set up these rules for selecting them but art and skill are important in choosing these two distractors.

1. The order of the correct choice was based on a table of random numbers, hence is one entirely of chance.

2. The incorrect choices are assumed to be on the same level of familiarity as the word being tested. The correct definition must not be harder than the word being tested. For example, we would not test 

garbage 

by the noun refuse or uncertain by the word dubious. When we began testing we did not know for certain how hard the distracting choices were but as the testing continued we used our own data to determine the difficulty level of words used as distractors. Distractors should not include terms ordinarily not known beyond the grade at which the word is tested. This is, of course, a counsel of common sense, but persons who make tests are mature adults with superior vocabularies and they may use distractors which the child doesn’t know.

3. The test usually occupied only one line.

4. We aimed to make all choices about the same length.

5. In general, if the word being tested is a negative one, the choices should be negative. If affirmative, the choices should be affirmative.

6. There is a tendency among test writers to use distractors which require spelling discriminations. The student is then tested not on the meaning of the word but on his spelling ability. So we would not test 

dual (twofold) with the distractor “to fight with swords” or test 

dar (coffin stand) as the distractor for “an alcoholic drink.” True, the ability to spell well may be required to tell the difference between the two words, but meaning, not spelling, was the aim of this testing. Inability to spell correctly may cause reading errors.

7. Care must be taken so that the word tested cannot be guessed by easy elimination. For example, if we were testing 

counter which, we would not use (a) answer to password (b) rear flag on a ship (c) inside advertisement. Below is a current list of actual test samples and scores ranging from the fourth grade to the college senior level.

<table>
<thead>
<tr>
<th>Grade</th>
<th>Score</th>
<th>Word</th>
<th>Sample Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>16</td>
<td>58%</td>
<td>styphon</td>
<td>(a) study of insects (b) figure of speech (c) a root word</td>
</tr>
<tr>
<td>16</td>
<td>68%</td>
<td>evoke</td>
<td>(a) call upon for help (b) take back again (c) call forth</td>
</tr>
<tr>
<td>13</td>
<td>67%</td>
<td>magma</td>
<td>(a) moon rocks (b) molten earth (c) sea bed</td>
</tr>
<tr>
<td>13</td>
<td>55%</td>
<td>bismark</td>
<td>(a) two-horned (b) with two lens openings (c) having two houses</td>
</tr>
<tr>
<td>12</td>
<td>79%</td>
<td>cardinal</td>
<td>(a) 1, 10, 20 (b) 1st, 2nd, 3rd, etc (c) 20th, 30th, 40th</td>
</tr>
<tr>
<td>12</td>
<td>72%</td>
<td>meacho</td>
<td>(a) cruel (b) speed (c) many</td>
</tr>
<tr>
<td>12</td>
<td>78%</td>
<td>edemalin</td>
<td>(a) gland secretion (b) harmful disease (c) stiff fabric</td>
</tr>
<tr>
<td>16</td>
<td>79%</td>
<td>gale</td>
<td>(a) bitter-tasting liquid (b) a celebration (c) a star cluster</td>
</tr>
<tr>
<td>10</td>
<td>67%</td>
<td>medley</td>
<td>(a) interference (b) illness (c) a mixture</td>
</tr>
<tr>
<td>8</td>
<td>79%</td>
<td>nitrogen</td>
<td>(a) 4/5 of the air (b) medicine (c) explosive</td>
</tr>
<tr>
<td>8</td>
<td>79%</td>
<td>architectural</td>
<td>(a) bowed (b) dartlike (c) about buildings</td>
</tr>
<tr>
<td>6</td>
<td>85%</td>
<td>gas guzzler</td>
<td>(a) requires much fuel (b) a stomach problem (c) drinks too much</td>
</tr>
<tr>
<td>6</td>
<td>74%</td>
<td>method</td>
<td>(a) good excuse (b) loud noise (c) plan for doing</td>
</tr>
<tr>
<td>4</td>
<td>71%</td>
<td>meteor</td>
<td>(a) rocket ship (b) falling star (c) crater on moon</td>
</tr>
<tr>
<td>4</td>
<td>74%</td>
<td>capable</td>
<td>(a) a kind of act (b) busy (c) real</td>
</tr>
</tbody>
</table>
Making the Tests

When we tested one form of a word it was helpful to have the earlier test scores on other forms, e.g., vary, variant, variable, variation, invariant, and the like. Hence the more we tested, the more data we gathered on the scores of related words.

We studied the scores of tested words to see whether they agreed with previous scores on inflected words or agreed with our personal knowledge of familiarity of words. The computer gave each word a score noting the differences between the score in the upper 10% of the group tested, and the lowest 25% per cent. We could quickly spot and revise weak tests on which the less able students did as well or better than the able students.

Furthermore, the average correct score on our tests ran around 75% per cent, hence we were working with words which are, on the whole, fairly well known. This was not true, however, at the 13th and 16th grade levels. We also had computer data on the per cent choosing each distractor which helped us determine their power or weakness as distractors. Words were retested when we concluded they tested too high or too low. But sometimes it was hard to discover why certain distractors overpulled or underpulled.

The Nature of the Sampling

In a study of the vocabulary knowledge of students in the United States the question of sampling immediately arises: What students and how many were sampled from the total population?

Here are questions asked about the sample of persons:

1. Why was a minimum of 200 students used for testing?
2. How were the students chosen for the study? Was it a random sample? Are these students representative?

Most of our tests were randomly sampled in cities and public school districts as found in the Education Directory of the U.S. Office of Education, Public School Systems.

School Systems

We had a few refusals to cooperate, but no evidence that the schools selected the more able classes. The few schools which did not give our tests usually indicated that they already were engaged in extensive testing programs. To avoid sectional bias we sent 50 tests or less of a single form to a single school district. We tested in all sizes of school systems—both rural and urban.

From 1954 to 1960 we used a three-choice test to determine the familiarity of 17,350 words in grades 4, 6, 8, 10, and 12. The sampling was nationwide and included 41 states and 281 schools. The familiarity scores and the names of all cooperating schools were presented in Children’s Knowledge of Words (Edgar Dale and Gerhard Eichholz, The Ohio State University, 1960, out of print).

After 1960 we tested about 320,000 students throughout the United States, approximately 1,500 tests. We present here familiarity scores on more than 44,000 word meanings in grades 4, 5, 6, 8, 10, 12, 13, and 16. Not less than 200 subjects were tested on each word. Keep in mind, however, that many words were tested more than 20 times. For example, the word adrenal gland was tested 200 times at the 8th (42%), 200 times at the 10th (65%), before it attained a passing score of 78% at the 12th grade level (the score in this book). The word copier (crafty) was tested at the 8th grade (46%), the 8th grade (57%), and at the 10th grade (72%)—its final score to date.

Our computerized data furnished a Kuder-Richardson reliability coefficient for each test. These correlations were rarely below 90. As noted earlier, we found that a sample of 200 students (per word) would provide stable averages. Arthur O. Gates of Teachers College found in his testing vocabulary that: “In some cases, it was necessary to obtain two-hundred or more responses from each of the six grades; in others, fewer grades and fewer cases were needed to stabilize the placement for the several percentages.” (Ibid., pp. 2-3.)

Validity

The ultimate test of the word list is its value in preparing reading materials. Does the use of the list enable writers to anticipate the difficulties students are likely to have in the reading material? Do the writers and editors use The Living Word Vocabulary when they have easy access to it? Here is one answer:

For many years the editorial staff of The World Book Encyclopedia has used the familiarity scores in their writing and editing. This is what James Shuster, Chief Copy Editor of The World Book Encyclopedia, says:

The value of your readability testing has certainly been proven again by the results on the articles recently returned to us. The correlation between the words cited as difficult by students and/or teachers, and the grade levels on the Dale word list, is indeed remarkable.

Has the use of this list by a major publisher resulted in clear-cut benefits? Here is the judgment of Kenneth F. Kuster, Editor, Encyclopedia Buying Guide, A Consumer Guide to General Encyclopedias in Print, who points out:

... For many years World Book has paid stricter attention to vocabulary control and precise levels of readability than any other general encyclopedia on the market...

... New and technical terms are italicized and defined in context. This approach to vocabulary is based on a 44,000-word graded list developed by Dr. Edgar Dale, a leading authority on readability and special consultant to World Book. In point of fact, Dr. Dale and his staff review all new and significantly revised material in each edition of World Book prior to publication to make sure that the vocabulary is geared to the encyclopedia’s intended readership. This procedure ensures a high degree

of comprehension of material as well as assists in vocabulary development . . . (p. 333).

To elaborate briefly, *World Book* is unquestionably one of the best designed and meticulously edited American encyclopedias ever made. As the digest above indicates, the encyclopedia passes every major critical test with flying colors. It is constructed to serve a very broad readership, ranging from students in the upper elementary grades to adults seeking basic information or reliable overview material on both academic and practical subjects (p. 337).

The set's authority is unimpeachable, its contents thoroughly accessible, and its up-to-dateness and breadth of annual revision impressive. The text is clear, direct, unusually interesting, and, when appropriate, the articles are written in pyramid style (i.e., from simple to complex).

In addition, the vocabulary is strictly controlled and, again when appropriate, articles are written to grade level . . . (p. 337).

We shall not complete our testing with this list of more than 44,000 words. We shall continue to test new words and check the scores of previously-tested words. As Samuel Johnson, we trust that there are "few wild blunders" and few "rustle absurdities."

May 1981

Edgar Dale, Ph. D.
Joseph O'Rourke, Ph. D.