A psychometric study of a domain-specific sentence verification technique test.

Urip Purwono

University of Massachusetts Amherst
A PSYCHOMETRIC STUDY OF
A DOMAIN-SPECIFIC SENTENCE VERIFICATION TECHNIQUE TEST

A Thesis Presented

by

URIP PURWONO

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

MASTER OF SCIENCE

February 1998

Department of Psychology
A PSYCHOMETRIC STUDY OF A DOMAIN-SPECIFIC SENTENCE VERIFICATION TECHNIQUE TEST

A Thesis Presented

by

URIP PURWONO

Approved as to style and content by:

James M. Royer, Chair

Ronald K. Hambleton, Member

Arnold D. Well, Member

Melinda Novak, Department Head
Department of Psychology
ACKNOWLEDGEMENTS

This thesis could never have been completed without the assistance, encouragement, and support of many individuals. It is, therefore, a great pleasure to me to convey my appreciation to all of the distinguished individuals who have made this work possible to complete.

Special recognition has to be extended to my chairman and advisor, Dr. James M. Royer. Without his help in arranging the collaboration with the Springfield Technological Community College, the luxury of collecting all the valuable four semesters of data utilized in this study would have been next to impossible for me to get. Above all of that, his understanding, support, patience, and intellectual guidance since the initial stage of this thesis has always encouraged me to keep moving and complete all the work.

Dr. Ronald K. Hambleton consistently encouraged his students to always set high standards and take pride in their resultant works. I am grateful for his motivation.

Dr. Arnold D. Well also deserves special thanks for his valuable feedback, encouragement, and, in particular, his patience.

I am also indebted to the staff at the Springfield Technological Community College for providing me the opportunity to collect the data. In particular, I wish to express my special gratitude to Ms. Georgina van Straat for her generous assistance in finding all the supporting data necessary to complete this project.

Finally, I would like to acknowledge the support of my peers. The intellectual-emotional support from both my office-mate at the Department of Psychology, Loel Tronsky, and my fellow graduate students at the Research and Evaluation Methods
Program was invaluable. It has always been great to have them around, particularly in times when clouds seemed to hide the sun.
# TABLE OF CONTENTS

ACKNOWLEDGEMENTS ........................................................................................................ iii
LIST OF TABLES .................................................................................................................. vii
LIST OF FIGURES ............................................................................................................... viii

CHAPTER

1. INTRODUCTION ............................................................................................................ 1
   1.1 Two Research Traditions ....................................................................................... 2
   1.2 Limitation of Traditional Reading Comprehension Assessment ....................... 4
   1.3 Purpose of the Study .............................................................................................. 8
   1.4 Plan of the Chapters ............................................................................................. 10
   1.5 The Definition of Comprehension ........................................................................ 10
   1.6 The Sentence Verification Technique (SVT) Procedures ................................... 13
   1.7 Psychometric Investigations Related to the SVT ............................................... 16
      1.7.1 Reliability of the SVT .................................................................................. 17
      1.7.2 Validity of the SVT .................................................................................... 17
   1.8 Test Development and the SVT ........................................................................... 24
   1.9 Research Questions .............................................................................................. 25

2. METHOD ......................................................................................................................... 29
   2.1 Participants ............................................................................................................ 29
   2.2 Materials ............................................................................................................... 29
   2.3 Administration Procedures .................................................................................... 30
   2.4 Variables and Analysis .......................................................................................... 30

3. RESULTS ......................................................................................................................... 33
   3.1 General Overview ................................................................................................. 33
   3.2 Classical-test-theory-based Analysis .................................................................... 35
      3.2.1 Reliability of the SVT .................................................................................. 35
      3.2.2 Item Analysis ............................................................................................... 38
      3.2.3 Difference in Item Statistics ......................................................................... 47
3.3 Validation Study .......................................................... 48

  3.3.1 Using the Bio-SVT to Predict College Performance ........ 49
  3.3.2 Factor Analysis Study ............................................. 53

4. SUMMARY, CONCLUSION, DISCUSSION, AND IMPLICATION
   FOR FURTHER RESEARCH ............................................. 57

APPENDIX .............................................................................. (in pocket)

REFERENCES ........................................................................... 65
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Number of Participants</td>
<td>34</td>
</tr>
<tr>
<td>3.2</td>
<td>Descriptive Statistic of the SVT</td>
<td>35</td>
</tr>
<tr>
<td>3.3</td>
<td>Reliability of each Passage in the Test</td>
<td>36</td>
</tr>
<tr>
<td>3.4</td>
<td>Reliability of each Item Type in the Test</td>
<td>37</td>
</tr>
<tr>
<td>3.5</td>
<td>Item Statistics</td>
<td>39</td>
</tr>
<tr>
<td>3.6</td>
<td>Items with Negative or Small Corrected-R</td>
<td>42</td>
</tr>
<tr>
<td>3.7</td>
<td>Cross Tabulation of Items and Item Types</td>
<td>44</td>
</tr>
<tr>
<td>3.8</td>
<td>Cross Tabulation of Item and the Passage from which the Items were Derived</td>
<td>44</td>
</tr>
<tr>
<td>3.9</td>
<td>Descriptive Statistic of the Selected Data</td>
<td>50</td>
</tr>
<tr>
<td>3.10</td>
<td>Anova Table for Regression Analysis between SVT scores and Final Grades in Introductory Biology Class</td>
<td>51</td>
</tr>
<tr>
<td>3.11</td>
<td>Correlation between Passages, Item types, and Total scores with the Final Grade in Introductory Biology Class</td>
<td>52</td>
</tr>
<tr>
<td>3.12</td>
<td>Regression analysis of Item types and Final Grade in Introductory Biology Class</td>
<td>52</td>
</tr>
<tr>
<td>3.13</td>
<td>Regression analysis of Passages and Final Grade in Introductory Biology Class</td>
<td>53</td>
</tr>
<tr>
<td>3.14</td>
<td>Rotated Factor Matrix</td>
<td>54</td>
</tr>
</tbody>
</table>
### LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Plot of p-values and corrected item-total correlation</td>
<td>43</td>
</tr>
<tr>
<td>3.2</td>
<td>Factors scree plot</td>
<td>54</td>
</tr>
<tr>
<td>3.3</td>
<td>Factor structures of the Bio-SVT</td>
<td>55</td>
</tr>
</tbody>
</table>
CHAPTER 1
INTRODUCTION

The measurement of reading comprehension has been a focus of research for almost a century. Devising a measure of reading ability was initiated by Rice, once a student of Wilhelm Wundt, when he attempted to use reading ability as an indicator of educational achievement. Since this early effort, interest in developing reading comprehension tests has continued (Johnston, 1984a).

The interest of many investigators in the measurement of reading comprehension can be seen as a manifestation of their recognition of the importance of the issue. Indeed, our modern culture relies heavily on written materials as a way of transferring, transforming, and translating information about our world. Consequently, it is understandable that reading (and also writing) has become an important attribute of individual success. Techniques for measuring reading, therefore, become critical because they provide us with information about mastery of reading skills, and with information that is useful for designing reading instruction. Such measures also enable us to evaluate the merit of an instructional program. Royer (in submission) elaborated reasons for measuring reading ability where he indicated that there are at least three possible purposes for assessing one’s reading comprehension ability: (1) to predict comprehension performance at some future point in time, (2) to diagnose reading comprehension difficulties, and (3) to assess the extent an individual has gained in reading comprehension competence. Reading comprehension performance has also been considered as a good indicator of academic performance. It is for this reason that for
many years comprehension tests have been included in several aptitude tests aimed at predicting future academic achievement such as the Scholastic Aptitude Test (SAT), Graduate Management Admission Test (GMAT), and Graduate Record Examination (GRE).

1.1 Two Research Traditions

An important event in the history of the assessment of reading comprehension was the adoption of Rice’s idea of comparative assessment of spelling performance by the Department of the Superintendent of the National Education Association in 1914 (Johnston, 1984a). Rice’s idea stimulated comparative testing and, in turn, also shaped the direction of reading comprehension tests. Most reading comprehension tests developed during this time were designed merely for comparison purposes and, therefore, did not usually consider psychological theories of the comprehension process.

Another factor that shaped the direction of the assessment of reading comprehension was the fact that many testing companies that developed standardized reading comprehension tests did so for practical reasons. The main practical purpose for reading tests was to predict future performance. This led to a situation where evidence supporting the predictive validity of the tests became more important than the underlying theoretical foundation of reading comprehension. Validity evidence in the form of the extent to which tests of interest could predict some future performances (e.g. academic or job performance, depending what the test was initially designed for) was assumed to be influenced by "reading comprehension ability". This emphasis on predictive validity again resulted in tests that were not influenced by theories of the psychological processes underlying reading comprehension.
The minimal interaction between researchers who are interested in the measurement issues of reading comprehension assessment and those who are interested in reading comprehension theory was also driven by the fact that, at the other end of the continuum, researchers who devoted their time to understand the phenomena of reading comprehension were primarily concerned with discovering the psychological processes that underlie reading and text comprehension and not on the idea of how to assess this human capability. Triggered by the work of Neisser (1967) that had stimulated the re-emergence of cognitive psychology, the "zeitgeist" during the time was primarily to uncover the process inside the "black box" of human cognition.

Fortunately, for the last two decades, both theoreticians in the field of cognitive psychology, who mainly work in uncovering the cognitive process of text comprehension, and test developers, whose main work is to develop a reliable and valid measures of reading comprehension ability, realize the importance of developing cognitive-theory-based reading comprehension assessment procedures. Emphasis on theory-based procedures was mainly driven by the fact that (1) the traditional format of reading comprehension test carries several problems, particularly if it is to be used for purposes other than comparing students belonging to different population or predicting the student’s general performance in the future, and (2) the types of observations and the patterns in data that reflect the ways that student think, perform and learn cannot be accommodated by traditional test models and methods (Mislevy, 1993). For these reasons, there has been an increase in the number of assessment procedures based on cognitive theory.
From the cognitive psychology perspective, the latent variable that was measured by cognitive theory-based assessment is a cognitive skill. In their review on techniques and procedures for assessing cognitive skills, Royer, Cicero, and Carlo (1993) described as many as 47 cognitive skill assessment techniques found in the literature which measure knowledge acquisition, knowledge structure and organization, depth of problem representation, mental models, meta cognitive skills, automaticity/encapsulation of performance, and efficiency of procedures.

The trend toward a cognitively-based test in the domain of reading comprehension has not only been a trend in the cognitive and educational psychology area. Researchers working in the psychometric area were also starting to consider the importance of the theory of human cognition in understanding mental processes that underlie student performance. Mislevy (1993), for example, acknowledged that the view of human abilities implicit in standard test theory is incompatible with the view rapidly emerging from cognitive and educational psychology. What he meant by this is that it is not quite true to view the learners increasing knowledge as merely an accumulation of facts and skills. He noted that learners increase their competence by reconfiguring their knowledge structures, by automating procedures, and chunking memory to reduce memory loads, and by developing strategies and models "... that tell them when and how facts and skills are relevant" (pp. 20). This view, undoubtedly, has also been shared by researchers in the area of cognitive and educational psychology.

1.2 Limitation of Traditional Reading Comprehension Assessment

Traditional standardized reading comprehension tests have been successful in indicating the examinee's general comprehension abilities which involves many abilities
including general cognitive ability and the tests are generally in predicting the examinees future general performances. They also function well for the purpose of comparison between groups of students and between a given student with a given norm group in terms of their reading text ability.

From the reading diagnostic perspective, however, the procedures have little value in explaining the source of failure in the test since the stimulus presented to the examinees were not constructed according to a specific cognitive theory of reading comprehension. Traditional tests, therefore, have a limited utility for diagnostic purposes.

Another limitation of traditional reading comprehension tests is that they have problems when used to evaluate the merit of an instructional program. As mentioned elsewhere in this chapter, traditional reading comprehension tests also measure general cognitive ability. Theoretically, this general cognitive ability is not penetrable by instructional program. Since performance of an examinee in traditional reading comprehension test is confounded with his/her general ability, it follows that the traditional standardized reading comprehension test would not be a good measure of educational gain in term of text comprehension ability. In some cases, the tests may not have detected improvement in reading comprehension ability resulting from carefully planned educational interventions (Royer & Cunningham, 1981).

Traditionally, reading comprehension has been measured by presenting students with passages accompanied by several multiple choice questions related to the passages. The questions usually consist of (1) direct questions, in which the answer can be found in the passage, (2) questions concerning the main idea of the passage, and (3) inferential
questions, where students have to combine the information contained in the passage with other prior information to answer this type of question correctly. The problem with this format is that success in answering the questions does not merely depend on the examinee's ability to comprehend the text.

Drum, Calfee, and Cook (1981) demonstrated this in a study that examined 52 text variables and related them to the student performance on traditional standardized reading comprehension test. They used several standardized reading achievement tests commonly used nationwide, namely the California Achievement Test (CAT) Form A, Levels 1, 2, and 3, Comprehensive Test of Basic Skills (CTBS), Form Q, Levels 1, 2, and 4, and Sequential Test of Educational Progress (STEP), Form A, Level 4. In one part of their study they investigated the contribution of the 16 orthogonal variables to the student's performance. These variables were grouped into four structural variables which are “Passage Components”, “Stem Components”, “Correct Choice Components”, and “Incorrect Choice Components”. Three out of these four variables were related to the properties of the test question and only one variable was related to the characteristic of the passages. Analysis utilizing multiple regression analysis were completed in order to assess the variance attributable to each of the predictor variables and to obtain the prediction equation.

Using the percent of correct responses on each of the total 210 items as criterion, the investigators found that only 21% of the variance in the number of items correctly answered was related to the characteristics of the passages that the item was based on, whereas 69% of the variance is associated with the properties of the test questions themselves. All in all, it was concluded that properties of the text passage do not
contribute substantially to student performance. This result suggested that performance on individual test items was more dependent on the examinees' ability to comprehend the question than it was on their ability to comprehend the passage. Drum, Calfee, and Cook (1981) further noted that this result fits with previous findings on the presence of passage independence of items in reading achievement tests (Pp.503). This result, for example, is consistent with Tuinman's (1974) finding that the average probabilities of correctly responding to reading comprehension test items with no passage present ranged between .32 and .50, which is well above the chance score of .25. Tuinman's study was based on administering 5 widely used standardized tests of reading comprehension to a total of 9,451 students equally divided over grades 4, 5, and 6 across a school system in the state of Indiana.

The result of studies described above raise the question of what standardized reading comprehension tests measure. One of the answers, as indicated by Johnston (1984a), is that they provide a fairly good indicator of IQ, just as do standardized vocabulary tests. Johnston (1984a) suggested asking a central question and preventing the reader from referring to the text while answering the questions to improve reading comprehension measurement. This suggestion was based on an assumption that what we mean by comprehension is the forming of a coherent cognitive model of the text meaning.

Another common format of reading comprehension test is the "cloze" technique. This technique involves presenting the examinee with passages that have incomplete sentences and asking them to fill the blank in each of the sentences with a word presented in the choice options. This procedure, as demonstrated by Kibby (1980),
Shanahan & Kamil (1982), and Shanahan, Kamil & Tobin (1982) also has substantial drawbacks in that cloze tests often measure sentence comprehension rather than passage comprehension.

These recognized shortcomings of traditional reading comprehension tests for diagnostic, evaluation, and placement purposes, coupled with the subsequent development in the cognitive-psychological theory of reading comprehension, have ultimately motivated many researchers to develop cognitive-based reading assessment. The Sentence Verification Technique (SVT) is one of such measures. The procedures were developed by James M. Royer in response, initially, to the need for the availability of a reliable reading comprehension measure to assess the impact of a well controlled instructional interventions (Royer and Cunningham, 1981).

1.3 Purpose of the Study

The present study was aimed at investigating the psychometric properties of one variant of the SVT based on material draw from Biology textbook at the introductory level in the higher education.

Since its first development in 1979, the SVT has been used for several different purposes outside the laboratory setting. The results have always been promising, suggesting the possibility of utilizing the SVT in various other settings. A study by Royer, Abranovic, and Sinatra (1987) in particular, showed that domain- relevant SVT tests were significant predictors of course performance, but not the irrelevant SVT. This study, involving a total of 148 college students enrolled in psychology and business statistics course, was designed to examine if SVT performance could predict learning performance and to identify whether the prediction was derived from the general or
specific route. SVT tests based on excerpts from a textbook used in psychology course and a textbook used in business statistics course were administered to all of the participants. The result showed that the psychology-based SVT test predicted course performance in psychology and business statistics-based SVT predicted performance in business statistics course, but not the other way around. That is, the psychology-based SVT turned out did not predict performance in business statistics class and business statistics-based SVT did not predict performance in psychology course.

The result of Royer, et al. (1987) above indicated that relevant SVT tests were significant predictors of course performance, but the irrelevant SVT tests were not, and that this prediction was derived from reading expertise in a specific content area (e.g., psychology) rather than from general reading expertise. Royer, Marchant, Sinatra, & Lovejoy (1990) replicated this result using psychology based-SVT and biology-based SVT as predictor measures. From test development perspective, results of these two studies suggests the possibility of using domain-relevant SVT as predictors of future course performance.

The present study focused on examining SVT test performance as a predictor of college student academic performance in a specific domain, namely Biology. The goal of the overall investigation was to examine the possibility of using SVT procedures to predict future academic performance in a specific subject such as Biology.

The data was collected at the Springfield Technological Community College (STCC) located in the state of Massachusetts. For many years, the performance of students enrolled in the introductory Biology class at the STCC has formed a bi-modal distribution. Some students do fine while others perform very poorly. Experience with
the course resulted STCC in personnel pre-testing students who are going to take an introductory Biology class so that the school can make a preliminary determination concerning who might not benefit from the regular instructional program and, therefore, need a special educational intervention. The school has been using the Nelson-Denny reading comprehension test for this purpose, and now seeks an alternative procedure. The present study, therefore, is also going to be beneficial to the school system in that it provides them with an alternative of an easy to build, easy to administer, and easy to expand selection procedure.

1.4 Plan of the Chapters

In describing the overall research, the chapters are organized as follows. The remainder of this introductory chapter contains a brief description of the background of the development of the SVT followed by a presentation of its theoretical foundation. The procedure of developing and administering the SVT will then be described. The chapter concludes with the research questions preceded by a literature review of relevant research.

In the second chapter, the overall design of the study is presented. The methods of gathering and analyzing data are described, followed by a chapter describing the results of the study. In the last chapter, conclusions are discussed and elaborated, leading to a final recommendation of both the use of the SVT and future research on the technique.

1.5 The Definition of Comprehension

One of the problems with traditional reading comprehension tests is that they usually are not developed with a clear definition of reading comprehension in mind. Test
developers do not usually define the meaning of comprehension as their basic operational
definition and, therefore, these measures are difficult to interpret on individual basis.
Operational definition of comprehension based on a valid theoretical foundation of
human cognition needs to be described before beginning an attempt to develop the test.

Carroll (1972) distinguished two different psychological stages involved in text
comprehension:

1. The apprehension of the linguistic message contained in text
2. Relating the information to a broader context.

Carroll’s notion above suggests that while adequate comprehension of a message
could be achieved at the time of reception, the “total meaning” might become available
only after a period of time allowing the message receiver to relate the message to other
information. Royer and Cunningham (1981) defined comprehension as apperception at
the time the message was received, and called the “total meaning” stages thinking.
Consequently, the two authors asserted that text comprehension results in the extraction
of the surface meaning of the text, while thinking results in the elaboration and
integration of the message.

In their further elaboration, Royer and Cunningham (1981) emphasized that the
meaning of a message is constructed rather than assembled. This notion is consistent
with current understanding that comprehension is a process of constructing coherent
mental representation of the text (Kintch and van Dijk’s, 1978; Kintch, 1988; van den
Broek, 1990; Vonk and Noordman, 1990; Myers, 1990; Garrod and Sanford, 1990), and
that this process occurs to a great extent at the point of message reception (Balota,
d’Arcais, and Rayner, 1990). This leads to an argument that one possible strategy to
assess text comprehension is to present the examinee with a passage followed by questions that investigate the memory representation in order to examine if the meaning of the text has been preserved.

The key concept in this argument is that if the product of comprehension was the creation of memory representation, then comprehension could be measured through the use of a technique that was sensitive to the extent to which an examinee has been successful at establishing a meaning preserving memory representation of a message that had been read (Royer, J. M., Hastings, C. N., & Hook, C., 1979). This idea is consistent with Carol's (1972) suggestion that verification techniques were a promising way of measuring text comprehension.

The use of verification format was motivated by the work of Sachs (1967; 1974), Pezdek and Royer (1974), and Anderson (1972). In investigating memory for sentences and discourse, Sachs (1974) presented participants with sentences and, at a point of time, asked them to verify whether a sentence was identical to the earlier presented sentences. Three types of sentences were used in the verification stages: (a) Semantic changes, (b) passive/active sentences, and (c) formal changes. This research concluded that the ability to detect meaning change stayed significantly longer than the ability to detect changes in the surface structure of the sentences. Pezdek and Royer (1974) used verification techniques in their study to investigate whether memory traces resulting from exposure to concrete sentences had a different form than those resulting from abstract sentences. They used three type of sentences to be verified by the participant: (a) identical sentences, (b) meaning change sentences, and (c) word change sentences. The
distinctive feature of the Pezdek and Royer (1974) assessment procedure was, instead of sentences, participants were presented with paragraphs to be read.

The work of Anderson (1972) also motivated the development of the SVT, particularly in the use of paraphrased sentences. In his paper concerning the measurement of comprehension, Anderson (1972) affirmed that students' acquisition of the meaning of instruction could be measured by presenting students with paraphrased questions and asking them to judge whether the meaning of the presented questions was equivalent to the meaning of the original instruction. The idea behind his suggestion was that since the surface features of the paraphrased sentences differ from those of the original events, students' success in verifying the sentences could not be based merely on memory of the exact test.

1.6 The Sentence Verification Technique (SVT) Procedures

As mentioned before, Royer, Hasting, and Hook (1979) proposed the use of Sentence Verification Technique (SVT) as a measure of text comprehension. The SVT test consists of a set of three to six passages, each of which is followed by a set of test sentences. The test sentences are arranged randomly with the only constraint that sentences derived from the first half of the original passages have to appear first in the test. This restriction is imposed to minimize the possibility of the examinee responding to the questions based on the content of short-term memory.

The first step in constructing an SVT test involves developing three different sentence types from each sentence appearing in the passage. The three sentence types are: (a) original sentence, (b) paraphrase, and (c) meaning change, each of which has rule for construction. Test development results in an item pool consisting of the three
different sentence types. The next step is to assemble test questions that are equal in number to the sentences in the passage by selecting one sentence type for each sentence in the passage and arranging it so that sentences derived from the first half of the original passages appear first in the test. The result should be that each of the three test sentence types equally appear in the test.

In addition to developing the three sentence types, the same number of distractor sentences is also prepared. In the final stage of test assembly, distractor sentences are added and randomly spread throughout the test. The final form of the 12 sentence length passage test, therefore, will consist of 16 test sentences: 4 originals, 4 meaning change, 4 paraphrases, and 4 distractors.

The general rule in developing each of the sentence types is as follows.

ORIGINAL

Original sentences are simply a copy of the original sentences that appear in the text passage.

PARAPHRASE

Paraphrase sentences are constructed by changing as many words as possible in an original sentence without altering the meaning of the original sentences. In addition, the meaning of the paraphrased sentence should fit with the meaning of the passage as a whole.

MEANING CHANGE

Meaning change sentences are constructed by changing one or two words of the original sentence so that the meaning of the sentence is altered. Meaning change sentences are created with the constraint that the meaning of the sentences should be
inconsistent with the meaning of the passage as a whole but should not be bizarre in any way. Royer (in submission) reported that meaning change sentences are the most difficult test sentences to write because the nature of a meaning change can vary from being highly obvious to very subtle.

DISTRACTER

A distractor is a sentence that differs in meaning from any original sentence in the passage. The meaning of a distractor sentence should not be related to any of the original sentences. The syntactic structures of the sentences, however, should be similar to the syntactic structure of the sentences that appear on the passage. In addition, distractor sentences should also be consistent with the thematic content of the passage. Royer (in submission) suggested that, in practice, a copy of sentences that appear in the text surrounding the target would satisfy the requirements of a good distractor test sentence.

Another form of SVT, called the Meaning Identification Technique (MIT), does not contain original or distractor sentences. Instead, the MIT only consists of two item types: paraphrase and paraphrase meaning change test sentences. The reason for this is that both distractor and original item types were suspected of not doing a good job in differentiating people, therefore lowering the overall reliability of the test. The MIT, however, do not possess diagnostic capability derived from the use of the four sentence types in the SVT.

An SVT test is administered by giving the students a test booklet and an answer sheet whenever available. The test administrator explains the materials by reading the direction provided in the first page of each test booklet. The proctor also explains that
the student’s task is to read a passage and to answer the test questions by marking it as “true” if the sentence has the same meaning as the sentence that appears in the passage, and marking it as “false” if the test sentence meaning is different from the one that appears in the paragraph without going back to the text passage.

Students are then typically given practice paragraphs, which are followed by test questions, to make sure that they understand the procedures. These practice paragraphs are also included in the direction section of the test. As in any other test, the proctor will not start the real test until everybody understands the procedures.

One of the most important aspects of the SVT administration is that the students should know that they cannot go back to the text material while they are answering the question. To ensure this, the text passage and the test question are printed on different pages so that once the students start to work on the test sentence they will not be able to return back to the passage without being observed to do so.

1.7 Psychometric Investigations Related to the SVT

Since the appearance of the first SVT article in the Journal of Reading Behavior (Royer, Hasting, and Cook, 1979), many investigations of the SVT have been reported. The studies range from examinations of the reliability and the validity of the SVT, to investigations of the use of SVT sub-scores to provide a diagnostic description of reading comprehension problems. In the following section, some of this research is reviewed, particularly the research related to the general psychometric properties of the SVT.
1.7.1 Reliability of the SVT

The reliability of SVT tests reported in the literature were mostly obtained from single administration procedures. More specifically, most of them were estimated utilizing Cronbach’s coefficient alpha. As an example of a study that examine the reliability of SVT tests and a new form called the Meaning Identification Technique (MIT), Marchant, Royer, and Green (1988) administered both the SVT and the MIT to one-hundred-three students from an undergraduate educational psychology class. They found that the coefficient alpha of the three paragraph length SVT was .53. Using the Spearman-Brown formula, the estimated reliability of the six paragraph length was .63. This was the lowest reliability of the SVT that has been reported in the literature. Higher reliability of the SVT has been reported by Greene, Royer, & Anzalone (1990). The three investigators developed SVT tests for use in Grenada. Based on a total of 786 examinees, the reliability of the SVT was reported to be .76 for a booklet with three passages, .78 for four passages, and .83 for five passages.

The reliability of SVT tests was also reported by Royer & Hambleton (1983). They administered 24 different booklets of six passage SVT to over 1,000 students enrolled in grades three to seven. Each of the booklets was administered to a minimum of 25 students and a maximum of 90 students. The reliability of the SVT was reported to range from .84 to .98 with an average reliability of .92.

1.7.2 Validity of the SVT

Royer (in submission) has reviewed evidence relevant to the construct validity of SVT tests. The paragraphs to follow examine evidence indicating that SVT tests (1) are sensitive to differences in reading skill, (2) are sensitive to text difficulty, (3) measure
passage comprehension as opposed to sentence comprehension, (4) are sensitive to intervention in the form of instruction, (5) are sensitive to working memory capacity. Each of these types of evidence will be briefly discussed in the following section.

The general approach in studying the validity of SVT tests is to investigate if the test is sensitive to reading skill and to text difficulty. The study is usually conducted by (1) giving an SVT test is that based on the same passage to students from different grade levels or (2) giving SVT tests that vary in difficulty levels to examinees with same reading ability. The logic is that because students from higher grades must have better reading skill, and also since student’s reading comprehension must vary as function of the difficulty level of the text, the student’s performance on the SVT tests should vary according to both grades levels and text difficulty.

In an investigation of whether SVT tests are sensitive to variation in reading skills and text difficulties, Royer, Hasting, & Hook (1979) administered SVT tests to twenty fourth graders and twenty-two sixth graders from a medium sized school district in Central Illinois. Three different levels of difficulty of SVT tests, the below level, on level, and above level, were administered to each group. The fourth graders, for example, had to do the second, fourth, and sixth grade level difficulty of the SVT, while the sixth graders had to do the fourth, sixth, and eighth grade level difficulty of the SVT. Using analysis of variance techniques that considered both proportion correct and $d'$ indices as dependent variables, the three authors found that there was a significant effect of difficulty level. These results indicate that the SVT was sensitive to differences in text difficulty, thereby supporting the construct validity of the SVT as a measure of reading comprehension.
Evidence for this differential function of the SVT scores have been replicated many times. Green, Royer, and Anzalone (1990), for example, also reported that SVT performance varies as a function of student ability. This conclusion is in accordance with the previous study conducted by Royer, Kulhavy, Lee, & Peterson (1986) in which they found that six grade students score significantly better on the SVT test than fourth grade students, and that performance on the SVT tests systematically declines as a function of increasing readability of the passages.

Evidence suggesting SVT tests are sensitive to reading ability was also gathered through the use of external criterion. With the same logic mentioned in the previous paragraphs, the general approach is to examine relationship between examinees’ performance on the SVT and their performance on an external criterion indicating their reading ability. The external criterion may either be a standardized reading test or teacher judgment of reading ability.

Using this approach, Rasool and Royer (1986) administered SVT tests to 44 third graders who attended two elementary schools in a medium-sized western Massachusetts city. The students were divided into teacher-defined good, medium, and poor reading groups which, in this case, was used as the external criterion measure. The results indicated that the SVT was sensitive to teacher-defined differences in reading competence. Moreover, using interview data capturing each student’s exposure to, understanding of, and attitude toward printed text, the two investigators also found that the SVT was significantly related to a student’s understanding of printed text. These two results confirm the validity of SVT tests as a measure of reading comprehension.
Royer, Lynch, Hambleton, and Bulgareli (1984) used a different strategy in investigating the validity of SVT tests as a measure of reading comprehension. Based on an assumption that readers with greater subject matter expertise should comprehend technical text in their area better than readers with less subject matter expertise, the four investigators conducted their validation study using subject matter expertise as criterion measure. The article reported four different experiments, two of which tested the hypothesis derived from the subject matter assumption mentioned before.

In Experiment-1, an SVT test was administered to graduate students in psychology, advanced undergraduate psychology majors, and undergraduates who had never had a psychology course. The test was a twelve-passage SVT, six of which were based on psychology material and the other six based on non-psychology material modified from the New York Times Sunday Book Review. In experiment 2 an SVT test was administered both at the beginning and the end of the semester to 82 undergraduates enrolled in a psychology methods course. The test was a 4-passage SVT consisting of two passages of psychology-based material and two passages of non-psychology-based material taken from experiment 1. Results of the two experiments confirmed the prediction that (1) student performance on the technical SVT test is related to subject matter expertise, and (2) the differences in student performance on technical SVT was not due to differences in their general ability. This result was also replicated by in a more current study of Royer, Carlo, Dufrense, and Mestre (1994). Moreover, another study also found that performance on SVT tests are related to performance on a standardized reading test. Royer, Sinatra, & Schumer (1990) reported that third and fourth grade students’ performance on the SVT was associated with their performance on the
California Test of Basic Skills. In addition, Royer (in submission) noted that SVT tests also correlated with various other standardized reading comprehension tests such as the Iowa Test of Basic Skills \((r=.73)\), California Achievement Test \((r=.52)\), and Stanford Achievement Test \((r=.50)\). All of these results support the validity of SVT tests as a measure of reading comprehension.

The study of SVT validity was also conducted by examining whether the tests were sensitive to text difficulty. One strategy is to administer SVT tests based on passages drawn from text used in different grades. One example of this is the study conducted by Royer et. al. (1979, Experiment 1) reviewed earlier in this chapter. The result of this study indicated that student performance on SVT tests declined systematically as a function of text difficulty. This result has replicated many times by other researchers (Green et al., 1990; Royer & Carlo, 1991; Royer et al., 1986).

Another way to investigate the sensitivity of SVT tests to text difficulty was by administering SVT tests based on passages that have been written using formal readability analysis such as the Dale-Chall readability formula and Kintch & Vipond structural text analysis. Using this approach, Royer et al. (1984, Experiment 1 & 3) has reported that performance on the SVT tests was significantly associated with text difficulty defined by the Kintch & Vipond method of analysis. This result is consistent with result obtained from a different study conducted by Lynch (1984) confirming that SVT tests are sensitive to text difficulty.

Another line of evidence necessary to support the construct validity of SVT tests is that the SVT tests measure passage comprehension as opposed to sentence comprehension. These studies are based on an assumption that a coherent text includes
causal relationships between sentences in the text. These causal relations play an important part in text comprehension as noted by Trabasso and Sperry (1985) that causal structures underlie the construction of the memory representation of a story. Comprehension of a text, therefore, involve something more than recognition of words and parsing of sentences into propositional units (Myers, 1990).

The argument that coherent text is more than merely a concatenation of sentences warranted the necessity of establishing evidence that a reading comprehension test indeed measures passage or text comprehension and not sentence comprehension. In the case of SVT tests, the Royer et al. (1984, Experiment 4) study provides this evidence. In an experiment using an SVT test consisting of 12 coherent and scrambled sentences passages administered to an experimental and control group, respectively, they found that examinee performance on SVT tests based on coherent passages was significantly better than performance on tests based on the same 12 scrambled sentences. This result supports the validity of SVT tests as measure of passage comprehension rather than sentence comprehension.

Validation study of SVT tests have also been conducted to examine the sensitivity of SVT tests to relevant instruction. The reasoning behind the studies involves a distinction between a measure of general ability and comprehension ability, and between measures of prior knowledge and knowledge about the presented reading material. One of the critiques of standardized reading comprehension tests is that the tests are suspected of measuring general cognitive abilities rather than reading ability. If this was also true for the SVT tests, then, the tests should not be sensitive to reading
instruction due to the nature of general cognitive abilities which are impenetrable by relevant instruction.

The second critique of standardized reading comprehension tests was based on empirical findings that students performed well above chance on answering the question correctly even without reading the passages. Citing the work of Johnston (1984), Royer (in submission) attributed this performance both to the students’ general cognitive abilities as well as the extent of their prior knowledge. If this was also the case in the SVT tests, then, the tests should also not be sensitive to relevant instruction.

A study showing that SVT tests was sensitive to relevant instruction was reported by Royer et al. (1984, Experiment 2). As described earlier in this chapter, the study involved administering SVT tests to college undergraduates both at the beginning and end of the semester. The students were enrolled in a psychology course that required them to read and interpret psychology journal articles. The tests given to the students were based on material drawn from psychology journal articles and material drawn from the New York Times Sunday Book Review. The results of the study showed significant pre- to post-instruction gains on the SVT test that were based on material drawn from psychology journal articles. However, there were no gains on the SVT test that were based on material drawn from the New York Times Sunday Book Review. This result supports the hypothesis that SVT tests are sensitive to relevant instruction. This result has also been replicated in a study by Royer, Sinatra, and Schumer (1990).

SVT test performance has also been related to an individual’s working memory capacity. Working memory is a central concept in reading theory. It is conceived as part
of memory that constitutes conscious awareness and thought as having limited capacity and limited duration. After a short period of time in the working memory, information will either be stored in long term memory or it will decay. In the case of reading, therefore, an individual’s ability to retain linguistic units such as ideas or proposition in working memory is related to a reader’s capabilities for comprehending text materials. In general, readers with inadequate working memory capacity have comprehension difficulties because they cannot retain a sufficient number of interrelated linguistics units in their working memory to construct a meaningful segment.

Evidence that performance on SVT tests is related to working memory capacity has been gathered by Lynch (1986, 1987). Lynch, as reported by Royer (in submission), found a significant correlation between SVT performance and working memory capacity.

1.8 Test Development and the SVT

From the test development perspective, one attractive feature of the SVT is its practicality. SVT tests are relatively easy to develop and can be based on virtually any text material. Example of materials SVT tests have been based on include: New York Times Sunday Book Review articles, abstracts of psychology journal articles, a text book used in an environmental biology course, a text book from a business statistics course, and a text book used in an introductory psychology course (Royer, Lynch, & Bulgareli, 1984; Royer, Abranovich, & Sinatra, 1987; Royer, Marchant, Sinatra, & Lovejoy, 1990). Within the school setting, the development of the SVT based on text passages selected by fourth- and sixth-grade teachers has also been reported as satisfactory (Royer, Hasting, and Hook, 1979).
Virtually any local personnel with no prior expertise in psychometrics can be easily trained to develop their own SVT tests. A typical training-workshop for developing the SVT would take about three working days of several personnel who are familiar with the subject matter of the assessment. The scoring procedures can be done by hand, and the test performance can be interpreted without norms established in large norming studies. This features allows for an inexpensive test development. Furthermore, involvement of the local personnel would contribute to the validity of the test because of their familiarity with the cultural and linguistic context of the target population.

This simple, easy to develop, and practical nature of SVT has resulted in the use of SVT tests as a measurement tool in several developing countries. For example, the procedure has been used in the second-year evaluation of the impact of a computer-assisted instructional system on student learning in Grenada (Greene, Royer, and Anzalone, 1990) to determine whether reading materials in use in the schools in Belize were matched appropriately to the reading skills of the students (Royer, Greene, 1990), and the listening version of the SVT has been reported as being adopted to evaluate whether the programs broadcast on an educational radio network in Nepal were understood by its target audience (Anzalone & Mathima, 1989).

1.9 Research Questions

The present study has two general purposes. The first purpose is to investigate the psychometric properties of SVT tests, and the second purpose is to investigate the use of SVT tests in predicting future learning.

It has been reported in the literature that different item formats have a different effect on examinee performance. SVT tests utilize the same item format in the entire test.
All the items are in the “true-false” format where the examinee task is to make a judgment on each of the question sentences. In this regard, the item format is not a problem. The test sentences themselves, however, are of four different types. As mentioned earlier, there are original, paraphrase, meaning change and distractor sentences in the test. These different item types may impact differently on the examinee’s performance. The present study was, therefore, aimed at investigating the effect of different item types as it relates to examinee performance. The investigation included examining (1) if the different item types yield a different proportion of correct answer, (2) if the different item types equally discriminate examinee performance, and (3) if the different item types contribute differently to the prediction of the criterion measure.

The present study also investigated other psychometric properties of the SVT test based on biology text. First of all, the overall reliability of the Bio-SVT was examined. The investigation was based on a single administration method. In addition, the item difficulty and item discrimination index of the Bio-SVT in regard to the five different passages in the test was also investigated.

As mentioned earlier in the chapter, one reason for developing the Bio-SVT was to use the test to identify students who were at risk in a standard biology course. In particular, one application of the test would be to use it to identify students at Springfield Technological Community College who should take a lower level biology course rather than the standard course. A cut-off score for making this decision, therefore, would be needed in the future. In this regard, the present study was also aimed at investigating the dependability of various cut-off scores in the test.
Another focus of the present study was to investigate the contribution of the Bio-SVT in predicting student's performance in introductory Biology class. The analysis was performed using final grades as criterion measure.

The idea of using reading comprehension performance as a predictor of future learning is not new. As mentioned at the beginning of this chapter, various tests developed to be used for college and university admission have a reading comprehension section. According to Royer, Abranovic, and Sinatra (1987) this is because traditional reading comprehension tests are sensitive to cognitive abilities that are relevant to comprehending a wide variety of material such as general knowledge, inferential reasoning ability, working memory capacity etc. and, therefore, can be thought of as tests that also measure general cognitive abilities. It follows that this type of test should be able to predict general learning performance. The three authors call this a general route of how reading comprehension performance can predict general learning performance.

Royer, et al. (1987) suggested there was another possible route for predicting learning performance that could be called the specific route. As an illustration of the specific route, we would expect a clinical psychologist to comprehend clinical psychology reading materials easier than a historian or an accountant or any other professional, even though they did not differ in their general reading capabilities. Likewise, the clinician would also be likely to learn the content of the reading materials with greater ease and efficiency than the other persons.

The expectation that reading performance would be related to subject matter expertise is consistent with theories of human cognition which indicate that an
individual’s ability to comprehend text material is, to some extent, dependent on working memory capacity and the amount of load in that memory.

In summary, then, the research questions of the present study can be stated as the following:

1. What is the overall reliability of the Bio-SVT test? Does each of the Bio-SVT item types contribute equally to the total reliability of the test?

2. What are the item statistics of the Bio-SVT? Does each of the item types in the SVT have equal difficulty? Does each of the item types in the SVT equally discriminate examinees who are taking the test?

3. Does the Bio-SVT predict examinee’s performance in an introductory biology class? What is the total contribution of the SVT test in predicting examinee’s performance in an introductory biology class?

4. What is the contribution of each of the item types in the SVT in predicting examinee performance in an introductory biology class?

5. Does the SVT test measure a single latent variable? If not, how many latent variables does the SVT test measure? What is the contribution of each of the item types in the SVT in measuring the latent variable of interest?

The analyses required to answer the above questions include a factor analysis, latent variable analysis, regression analysis, and correlation techniques. The complete design of the study is discussed further in the next following chapter.
CHAPTER 2

METHOD

This chapter contains the description of participants involved in the study, the material used, variables, and the analyses performed to answer the research questions.

2.1 Participants

The study involved a total of 461 students of the Springfield Technology Community College (STCC) who were enrolled in a biology class for the first time in the Fall semester, 1994, Spring and Fall semester, 1995, and Spring semester, 1996.

In the Fall semester, 1994, the SVT test was administered at the end of the semester, in the Spring and Fall semester, 1995, the SVT test was administered both at the end and at the beginning of the semester, and in the Spring semester 1996 the SVT was administered at the beginning of the semester. In addition, each student’s final grade in the biology class, age, and gender were also collected. This information were provided by the STCC record office. Analyses in the present study, however, were based only on data collected at the beginning of the semester.

2.2 Materials

The SVT test administered contains five passages that were developed for the purpose of the present study by biology course instructors at STCC who had been trained in SVT test development. The test, which is called Bio-SVT, consists of a set of five passages, each twelve sentences long, taken from an introductory biology textbook. Each of the passages is followed by a set of 16 sentences consisting of four of each of
the four types of test sentences which are the original, paraphrase, meaning change, and distractor.

2.3 Administration Procedures

The SVT test was administered at the beginning of the class by the class instructor. First the instructions were read by the instructor while the students read it from the printed material distributed earlier. Students were then instructed to read the two example passages and answer the test questions.

The student’s task on the test was to decide whether each of the test sentences presented to them has the same meaning as the sentences on the passage without coming back to the original passage. The number of correct answers determined each student's score. Since each passage consisted of 12 sentences, there were 16 test sentences (4 original, 4 paraphrase, 4 meaning change, and 4 distractor sentences) to be answered. It follows that the possible score ranged from 0 to 80.

The testing took place in December 1994, January 1995, October 1995, December 1995, and February 1996 at the Springfield Technology Community College, Springfield, Massachusetts, administered by the class instructor who was briefed on how to administer the SVT.

The procedures were the same for all the students who participated in the study. Instructions were given in English. Examples were given to ensure that every student understood the task precisely before the actual test was started.

2.4 Variables and Analysis

Variables that were collected throughout the study consisted of:
1. Overall Bio-SVT item responses as well as Bio-SVT item responses broken down into both passages and item types.

2. Final grade in the Biology class.

The analyses included:

1. Analysis of the reliability of the test. Both “split-half” and Cronbach’s coefficient alpha were computed as the reliability estimates.

2. Analysis of the p values. This analysis was based on the same data analyzed before. The proportion of correct answers was computed for each item. To examine if there is a different in item difficulty across the four item types, an analysis of variance procedure was performed by considering item types as the independent variable.

3. Analysis of the discrimination indices of each of the item. This analysis was also based on the same data configuration mentioned above. To examine if the four item types discriminate differently, analysis of variance procedures was performed by considering item types as the independent variable and the item-total correlation as the dependent variable.

4. Exploratory factor analysis to investigate latent variable measured by the test. To conduct this analysis, all of the student’s responses were coded “1” for the correct answer, and “0” for the wrong answer. All the data were then combined, and various different latent variable models were tested using this data. The analysis itself was performed with the help of the SPSS statistical program.

5. Regression analysis to investigate the contribution of the overall test as well as of each of the item types in the prediction of the criterion measure. These analyses
were performed based only on the data that had valid observations on both the SVT variable and the final grades variable.
CHAPTER 3

RESULTS

The results of the investigation will be described in the next three sections. To give a perspective on the characteristics of the participants involved in the study as well as a general description of the rest of the data, the first section presents a general overview of the study and the data that were collected.

The remaining two sections present results that pertain to the two major purposes of the study. The first of these address questions derived from classical test theory. In particular, the reliability of the Bio-SVT was investigated based on both a “split half” method and by applying the Cronbach’s coefficient alpha formula. Also included in the second section are analyses of the items and item types in the test.

Finally, the results of a predictive validity study of the Bio-SVT test is presented in the fourth section of this chapter. This section primarily addresses the question of how well the Bio-SVT test predicted the student’s performance in an introductory biology class. As an initial step of collecting empirical construct validity evidence, a preliminary factor analysis of the data was also conducted. Although somewhat inconclusive, the result of this analysis is also briefly discussed in this final section of the chapter.

3.1 General Overview

As mentioned in the earlier chapter, data was collected throughout the beginning of the Spring 1995, Fall 1995, and Spring 1996 semesters. Table 3.1 below shows the number of participants involved in the study.
Table 3.1  
Number of Participants

<table>
<thead>
<tr>
<th>Semester</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spring 1995</td>
<td>161</td>
</tr>
<tr>
<td>Fall 1995</td>
<td>182</td>
</tr>
<tr>
<td>Spring 1996</td>
<td>170</td>
</tr>
</tbody>
</table>

After some cleaning up procedures to exclude participants that did not make an attempt to do the last two passages of the Bio-SVT test, 461 participants were included in the data sample. Only about half of the participants identified their gender and 34.7% of those were female.

The average age of the participants was 27 years with 19 being the youngest and 54 being the oldest. Of the original 461 participants, the final grade in the Biology class was only available for 195 cases. This was due to the fact that a significant portion of students who were initially registered for the class either dropped the class or withdrew at some point during the semester. Another reason for loss of data was some participants either did not give their social security number or did not identify it properly. Since the social security number was the only number used in this study to match and merge the data, data from unidentified students had to be discarded from the analysis.
3.2 Classical-test-theory-based Analysis

3.2.1 Reliability of the SVT

The initial analysis performed in the present study was aimed at estimating the reliability of the Bio-SVT test. Table 3.2 below shows descriptive statistics of the SVT based on the 461 participants mentioned above.

Table 3.2
Descriptive Statistic of the SVT

<table>
<thead>
<tr>
<th>Scale</th>
<th>N Items</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Overall SVT</td>
<td>80</td>
<td>56.39</td>
<td>6.76</td>
</tr>
<tr>
<td>Item types:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>20</td>
<td>16.49</td>
<td>2.21</td>
</tr>
<tr>
<td>Distractor</td>
<td>20</td>
<td>15.37</td>
<td>3.15</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>20</td>
<td>14.60</td>
<td>2.62</td>
</tr>
<tr>
<td>Meaning Change</td>
<td>20</td>
<td>9.92</td>
<td>3.19</td>
</tr>
<tr>
<td>Passages:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage-1</td>
<td>16</td>
<td>10.61</td>
<td>1.99</td>
</tr>
<tr>
<td>Passage-2</td>
<td>16</td>
<td>11.93</td>
<td>2.03</td>
</tr>
<tr>
<td>Passage-3</td>
<td>16</td>
<td>11.82</td>
<td>1.79</td>
</tr>
<tr>
<td>Passage-4</td>
<td>16</td>
<td>11.82</td>
<td>1.90</td>
</tr>
<tr>
<td>Passage-5</td>
<td>16</td>
<td>10.21</td>
<td>2.44</td>
</tr>
</tbody>
</table>

Using the data set shown in Table 3.2, Cronbach coefficient alpha was 0.71.

Another estimate of the reliability was obtained using the “split half” method which involved splitting the test into two 40 items test. To insure equality between the two forms, item types in the test were divided equally so that each of the forms consisted of 10 original, 10 distractor, 10 paraphrase, and 10 meaning change items distributed
equally across the five passages of the test. This procedure yielded a reliability estimate of .55, and, after a correction using the Spearman-Brown formula, the final estimate was .71.

Results from the two procedures above suggests that the reliability of the Bio-SVT test was in a range of 0.70. Even though this reliability is within an acceptable range, the estimate is small for a five passages SVT test. Royer (in submission) cited that the average reliability for a five passages SVT test would typically be near 0.90. A closer look at each of the items helped to explain this somewhat unusual result.

Another question in the study concerned the magnitude of the reliability of each passages in the Bio-SVT. Using the same set of data, a Cronbach's coefficient alpha was computed for each passage. The result was then corrected using the Spearman-Brown formula to obtain an estimate of an 80 item test. Table 3.3 below shows the result of this analysis.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Scale</th>
<th>N Items</th>
<th>Mean</th>
<th>SD</th>
<th>Reliability (r)</th>
<th>Corrected -r</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Passage-1</td>
<td>16</td>
<td>10.61</td>
<td>1.99</td>
<td>.32</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>Passage-2</td>
<td>16</td>
<td>11.93</td>
<td>2.03</td>
<td>.38</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>Passage-3</td>
<td>16</td>
<td>11.82</td>
<td>1.79</td>
<td>.26</td>
<td>.63</td>
</tr>
<tr>
<td></td>
<td>Passage-4</td>
<td>16</td>
<td>11.82</td>
<td>1.90</td>
<td>.32</td>
<td>.70</td>
</tr>
<tr>
<td></td>
<td>Passage-5</td>
<td>16</td>
<td>10.21</td>
<td>2.44</td>
<td>.47</td>
<td>.81</td>
</tr>
</tbody>
</table>
The results in Table 3.3 suggest that the corrected reliability of each passages in the SVT was in the range of .6 to .8. Passage three, which was about cells as a building block of living organism, had the lowest reliability, while passage five discussing the two types of cells, had the highest reliability.

Another interesting finding concerning the five different passages was that passage 5 was associated with the largest standard deviation. As described further in the next section, however, there were no differences in item statistics among the five passages. According to the analysis, the five passages had about the same difficulty and had also about the same average item-total correlation.

The only difference between passage five and the rest of the passages in the test was that passage five was presented as the last passage in the test. It was possible, therefore, that the location of passage five in the test was contributing an increase in performance variability. An independent investigation, however, is needed to confirm this hypothesis.

In terms of items presented to the students to be verified after they read the passage, the Bio-SVT consisted of four different item types: the original, distractor, paraphrase and meaning change. Reliability of each of the item types was also investigated in the study using the same set of data with the same method as the investigation of passage based reliability. Table 3.4 shows the result of the analysis.
Table 3.4
Reliability of each Item Type in the Test

<table>
<thead>
<tr>
<th>Scale</th>
<th>N Items</th>
<th>Mean</th>
<th>SD</th>
<th>Reliability (r)</th>
<th>Corrected -r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>20</td>
<td>16.49</td>
<td>2.21</td>
<td>.47</td>
<td>.78</td>
</tr>
<tr>
<td>Distractor</td>
<td>20</td>
<td>15.37</td>
<td>3.15</td>
<td>.71</td>
<td>.91</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>20</td>
<td>14.60</td>
<td>2.62</td>
<td>.53</td>
<td>.81</td>
</tr>
<tr>
<td>Meaning Change</td>
<td>20</td>
<td>9.92</td>
<td>3.19</td>
<td>.59</td>
<td>.85</td>
</tr>
</tbody>
</table>

Table 3.4 above suggests that the distractor items were the most reliable of the item types.

3.2.2 Item Analysis

Item analysis is an important step in every test development process. The two most important item statistics are item difficulty and item discrimination indices. The present study was also aimed at obtaining the two item statistics for the test. Item difficulty was estimated by the "proportion correct", which is the proportion of examinees correctly answer the item of interest. Item discrimination were obtained from the correlations between a given item and the rest of the items in the test. Table 3.4 shows these item statistics.
Table 3.5
Item Statistics

<table>
<thead>
<tr>
<th>Item Number</th>
<th>p</th>
<th>Corrected-r</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.91</td>
<td>.08</td>
</tr>
<tr>
<td>2</td>
<td>.90</td>
<td>.18</td>
</tr>
<tr>
<td>3</td>
<td>.91</td>
<td>.16</td>
</tr>
<tr>
<td>4</td>
<td>.31</td>
<td>.09</td>
</tr>
<tr>
<td>5</td>
<td>.75</td>
<td>.11</td>
</tr>
<tr>
<td>6</td>
<td>.26</td>
<td>.29</td>
</tr>
<tr>
<td>7</td>
<td>.71</td>
<td>.12</td>
</tr>
<tr>
<td>8</td>
<td>.36</td>
<td>-.07</td>
</tr>
<tr>
<td>9</td>
<td>.47</td>
<td>.14</td>
</tr>
<tr>
<td>10</td>
<td>.86</td>
<td>.13</td>
</tr>
<tr>
<td>11</td>
<td>.68</td>
<td>.15</td>
</tr>
<tr>
<td>12</td>
<td>.92</td>
<td>.27</td>
</tr>
<tr>
<td>13</td>
<td>.82</td>
<td>.00</td>
</tr>
<tr>
<td>14</td>
<td>.73</td>
<td>.11</td>
</tr>
<tr>
<td>15</td>
<td>.63</td>
<td>.23</td>
</tr>
<tr>
<td>16</td>
<td>.40</td>
<td>.17</td>
</tr>
<tr>
<td>17</td>
<td>.30</td>
<td>.04</td>
</tr>
<tr>
<td>18</td>
<td>.88</td>
<td>.08</td>
</tr>
<tr>
<td>19</td>
<td>.75</td>
<td>.22</td>
</tr>
<tr>
<td>20</td>
<td>.95</td>
<td>.12</td>
</tr>
<tr>
<td>21</td>
<td>.72</td>
<td>.23</td>
</tr>
<tr>
<td>22</td>
<td>.72</td>
<td>-.06</td>
</tr>
<tr>
<td>23</td>
<td>.87</td>
<td>.13</td>
</tr>
<tr>
<td>24</td>
<td>.81</td>
<td>.20</td>
</tr>
<tr>
<td>25</td>
<td>.75</td>
<td>.15</td>
</tr>
<tr>
<td>26</td>
<td>.78</td>
<td>.26</td>
</tr>
<tr>
<td>27</td>
<td>.54</td>
<td>.22</td>
</tr>
<tr>
<td>28</td>
<td>.89</td>
<td>.26</td>
</tr>
<tr>
<td>29</td>
<td>.56</td>
<td>.27</td>
</tr>
<tr>
<td>30</td>
<td>.73</td>
<td>.15</td>
</tr>
<tr>
<td>31</td>
<td>.81</td>
<td>.27</td>
</tr>
<tr>
<td>32</td>
<td>.87</td>
<td>.16</td>
</tr>
<tr>
<td>33</td>
<td>.83</td>
<td>.14</td>
</tr>
<tr>
<td>34</td>
<td>.45</td>
<td>.20</td>
</tr>
<tr>
<td>35</td>
<td>.82</td>
<td>.19</td>
</tr>
<tr>
<td>36</td>
<td>.96</td>
<td>.08</td>
</tr>
<tr>
<td>37</td>
<td>.87</td>
<td>-.16</td>
</tr>
<tr>
<td>38</td>
<td>.81</td>
<td>.06</td>
</tr>
<tr>
<td>39</td>
<td>.84</td>
<td>.09</td>
</tr>
<tr>
<td>40</td>
<td>.88</td>
<td>.26</td>
</tr>
</tbody>
</table>

Continued next page
Table 3.5 (Continued)

<table>
<thead>
<tr>
<th>Item Number</th>
<th>p</th>
<th>Corrected-r</th>
</tr>
</thead>
<tbody>
<tr>
<td>41</td>
<td>.61</td>
<td>.24</td>
</tr>
<tr>
<td>42</td>
<td>.90</td>
<td>.27</td>
</tr>
<tr>
<td>43</td>
<td>.47</td>
<td>.19</td>
</tr>
<tr>
<td>44</td>
<td>.25</td>
<td>-.17</td>
</tr>
<tr>
<td>45</td>
<td>.78</td>
<td>.22</td>
</tr>
<tr>
<td>46</td>
<td>.54</td>
<td>.04</td>
</tr>
<tr>
<td>47</td>
<td>.91</td>
<td>.23</td>
</tr>
<tr>
<td>48</td>
<td>.89</td>
<td>.28</td>
</tr>
<tr>
<td>49</td>
<td>.92</td>
<td>.27</td>
</tr>
<tr>
<td>50</td>
<td>.91</td>
<td>.06</td>
</tr>
<tr>
<td>51</td>
<td>.91</td>
<td>.25</td>
</tr>
<tr>
<td>52</td>
<td>.93</td>
<td>.16</td>
</tr>
<tr>
<td>53</td>
<td>.80</td>
<td>.09</td>
</tr>
<tr>
<td>54</td>
<td>.85</td>
<td>.18</td>
</tr>
<tr>
<td>55</td>
<td>.79</td>
<td>.33</td>
</tr>
<tr>
<td>56</td>
<td>.49</td>
<td>.25</td>
</tr>
<tr>
<td>57</td>
<td>.74</td>
<td>.09</td>
</tr>
<tr>
<td>58</td>
<td>.57</td>
<td>.07</td>
</tr>
<tr>
<td>59</td>
<td>.92</td>
<td>.10</td>
</tr>
<tr>
<td>60</td>
<td>.34</td>
<td>.01</td>
</tr>
<tr>
<td>61</td>
<td>.81</td>
<td>.02</td>
</tr>
<tr>
<td>62</td>
<td>.51</td>
<td>.02</td>
</tr>
<tr>
<td>63</td>
<td>.49</td>
<td>.23</td>
</tr>
<tr>
<td>64</td>
<td>.84</td>
<td>.25</td>
</tr>
<tr>
<td>65</td>
<td>.61</td>
<td>.15</td>
</tr>
<tr>
<td>66</td>
<td>.34</td>
<td>.09</td>
</tr>
<tr>
<td>67</td>
<td>.58</td>
<td>.30</td>
</tr>
<tr>
<td>68</td>
<td>.56</td>
<td>.29</td>
</tr>
<tr>
<td>69</td>
<td>.52</td>
<td>.29</td>
</tr>
<tr>
<td>70</td>
<td>.50</td>
<td>-.03</td>
</tr>
<tr>
<td>71</td>
<td>.90</td>
<td>-.03</td>
</tr>
<tr>
<td>72</td>
<td>.83</td>
<td>.08</td>
</tr>
<tr>
<td>73</td>
<td>.40</td>
<td>.02</td>
</tr>
<tr>
<td>74</td>
<td>.80</td>
<td>.31</td>
</tr>
<tr>
<td>75</td>
<td>.78</td>
<td>.13</td>
</tr>
<tr>
<td>76</td>
<td>.72</td>
<td>.09</td>
</tr>
<tr>
<td>77</td>
<td>.77</td>
<td>.03</td>
</tr>
<tr>
<td>78</td>
<td>.58</td>
<td>.31</td>
</tr>
<tr>
<td>79</td>
<td>.67</td>
<td>.25</td>
</tr>
<tr>
<td>80</td>
<td>.64</td>
<td>.34</td>
</tr>
</tbody>
</table>
Table 3.5 shows that among the 80 items, only six were correlated negatively with the rest of the items in the test. These items, therefore, were excluded from the rest of the analysis.

Crocker and Algina (1986) have pointed out that in the test development stage, test developers often elect to keep every item that has an item criterion correlation significantly greater than zero. For point biserial correlations, the authors suggested that a convenient approximation for the standard error for the Pearson product moment correlation can be used to establish this level, by computing

$$
\sigma_p = \frac{1}{\sqrt{N - 1}}
$$

where N is the sample size.

Furthermore, Crocker and Algina (1986) also mentioned that the common practice is to set the critical value at two standard errors above zero. Applying this formula to the 461 student sample in the present study resulted in the elimination of 21 items due to very low item discrimination indices (and six more due to negative discrimination indices).

Table 3.6 shows the six items that negatively correlated with the total scores and the 21 items that correlated below .1 with the total scores.
Table 3.6 suggests that in terms of p-values, item types, location in the passage, and the correct answer key, the 27 "bad items" above were indistinguishable from the

<table>
<thead>
<tr>
<th>ITEM</th>
<th>Corrected-R</th>
<th>p-values</th>
<th>Item-types</th>
<th>Passages</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.08</td>
<td>.91</td>
<td>O</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>4</td>
<td>.09</td>
<td>.31</td>
<td>M</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>8</td>
<td>-.07</td>
<td>.36</td>
<td>P</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>13</td>
<td>.00</td>
<td>.82</td>
<td>O</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>17</td>
<td>.04</td>
<td>.30</td>
<td>M</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>18</td>
<td>.08</td>
<td>.88</td>
<td>O</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>22</td>
<td>-.06</td>
<td>.72</td>
<td>D</td>
<td>2</td>
<td>Y</td>
</tr>
<tr>
<td>36</td>
<td>.08</td>
<td>.96</td>
<td>O</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>37</td>
<td>-.16</td>
<td>.87</td>
<td>M</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>38</td>
<td>.06</td>
<td>.81</td>
<td>O</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>39</td>
<td>.09</td>
<td>.84</td>
<td>P</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>44</td>
<td>-.17</td>
<td>.25</td>
<td>M</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>46</td>
<td>.04</td>
<td>.54</td>
<td>P</td>
<td>3</td>
<td>Y</td>
</tr>
<tr>
<td>50</td>
<td>.06</td>
<td>.91</td>
<td>P</td>
<td>4</td>
<td>Y</td>
</tr>
<tr>
<td>53</td>
<td>.09</td>
<td>.80</td>
<td>P</td>
<td>4</td>
<td>Y</td>
</tr>
<tr>
<td>57</td>
<td>.09</td>
<td>.74</td>
<td>P</td>
<td>4</td>
<td>Y</td>
</tr>
<tr>
<td>58</td>
<td>.07</td>
<td>.57</td>
<td>P</td>
<td>4</td>
<td>Y</td>
</tr>
<tr>
<td>60</td>
<td>.01</td>
<td>.34</td>
<td>M</td>
<td>4</td>
<td>N</td>
</tr>
<tr>
<td>61</td>
<td>.02</td>
<td>.81</td>
<td>M</td>
<td>4</td>
<td>N</td>
</tr>
<tr>
<td>62</td>
<td>.02</td>
<td>.51</td>
<td>M</td>
<td>4</td>
<td>N</td>
</tr>
<tr>
<td>66</td>
<td>.09</td>
<td>.34</td>
<td>M</td>
<td>5</td>
<td>N</td>
</tr>
<tr>
<td>70</td>
<td>-.03</td>
<td>.50</td>
<td>D</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>71</td>
<td>-.03</td>
<td>.90</td>
<td>P</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>72</td>
<td>.08</td>
<td>.83</td>
<td>O</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>73</td>
<td>.02</td>
<td>.40</td>
<td>M</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>76</td>
<td>.09</td>
<td>.72</td>
<td>P</td>
<td>5</td>
<td>Y</td>
</tr>
<tr>
<td>77</td>
<td>.03</td>
<td>.77</td>
<td>O</td>
<td>5</td>
<td>Y</td>
</tr>
</tbody>
</table>
rest of the items. In term of item difficulty, for instance, the p-values values of this 27 items were ranging from .25 to .96.

To further investigate the characteristics of the above mentioned 27 “bad items” an attempt was made to compare the 27 “bad items” with 27 items that had higher correlations with the rest of the items (hereafter, collectively referred to as “good items”). Figure 3.1 shows the plot of these items in term of their p-values.

![Plot of p-values and corrected item-total correlation](image)

Figure 3.1 Plot of p-values and corrected item-total correlation

Figure 3.1 indicates that, in term of p-values, the 27 “bad items” were indistinguishable from the 27 “good items”. This result was also supported by the non significant correlation ($r = .19; p > .5$) between the category of the items ("Good" or "Bad") and their p-values.

Table 3.7 shows cross-tabulating the two groups of items with item types.
Table 3.7
Cross Tabulation of Items and Item Types

<table>
<thead>
<tr>
<th>Original</th>
<th>Distractor</th>
<th>Paraphrase</th>
<th>Meaning Change</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Good Items&quot;</td>
<td>4</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>&quot;Bad Items&quot;</td>
<td>8</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>15</td>
<td>15</td>
</tr>
</tbody>
</table>

A closer examination to the data showed that the \( \chi^2 \) statistics indicated that there was no significant association between the two groups of items with item types (\( \chi^2 (3) = 12.47; \ p > .05 \)). Applying the same procedures by cross tabulating the two groups of items with the passage where the items were derived resulted in no significant association between the two variables (\( \chi^2 (4) = 1.43; \ p > .05 \)). The cross tabulation is presented in the following Table 3.8.

Table 3.8
Cross Tabulation of Item and the Passage from which the Items were Derived

<table>
<thead>
<tr>
<th>Passage1</th>
<th>Passage2</th>
<th>Passage3</th>
<th>Passage4</th>
<th>Passage5</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Good Item&quot;</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>&quot;Bad Item&quot;</td>
<td>4</td>
<td>3</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>9</td>
<td>12</td>
<td>12</td>
</tr>
</tbody>
</table>
Results of the above comparison supported the aforementioned conclusion that in term of p-values, item types, and item location in the passage, the 27 “bad items” were indistinguishable from the rest of the items. Further investigation, therefore, were directed toward looking more closely to the item itself in term of wording and how the questions were formulated.

Results of a closer examination to each of the “bad items” indicated that there were six problems that can explain those disappointing item statistics.

1. The items were derived from sentences that were relatively unimportant to the main idea of the passage. As described earlier, SVT items was derived from sentences appeared in the reading passages. The rule is each of the sentences appear in the reading passage has to be utilized to make an item in the test. The sentence may be paraphrased, altered in meaning, or kept in its original form to make an SVT test item. Most passage, however, contain sentences or phrases that are relatively unimportant or indirectly related to the main idea of the passage. Because most reader would not pay a close attention to these unimportant sentences and phrases, it is possible that examinee’s performance on items derived from these unimportant sentences will be unrelated to his or her performance on the overall test. This supposition, however, needs to be elaborated further. One possibility is to have an expert rate each sentence in the text in terms of its importance, and use the rating to test the hypothesis that the sentence importance is related to the correlation of the item with the total score. Items that felt into this category were item 1, 8, 38, and 39.

2. The items were derived from sentences that would reflect student’s knowledge in the domain other than Biology. Student answers to these items, therefore, would reflect
more of a background knowledge outside the domain of interest, be independent of the text material, and would not related to the student performance on the overall test. Items that fell into this category were items 13 and 58.

3. The items were derived from a sentence that were too long that make each of the sentences too difficult to comprehend. As a result, the answer would not be more than random guessing. Items that fell into this category were items 4, 18, and 73.

4. Distractor items that were taken from the passage discussing the same topics as the one presented in the reading passage, and therefore positively false-alarmed examinees. Items that fell into this category were item 22, 42, and 70.

5. The items were derived from sentences that stating general background knowledge so that most examinees answered the items correctly regardless of their understanding of the text material. Items that fell into this category were items 36, 50, 53, and 71.

6. The items were simply too difficult so that examinees that correctly answer the items got it right by chance alone. In most cases items that fell into this category were “Meaning Change” items where the change was so subtle that even the higher negatively positively false-alarmed by the items. Items that fell into this category were items 17, 44, 60, and 62.

The six categories mentioned above explained all but four of the “bad items. These four items that did not belong to the above category yet correlated either zero or negatively with the rest of the items were items 37, 57, 61, and 72. The following are the exact wording of those four items.
Item 37: Living cells were first scientifically observed and reported in 1655 by the Englishman, Robert Hooke. (Correct answer: “No”; Meaning Change item).

Item 57: Also floating in the cytoplasm are bigger structures, called organelles, that perform a variety of functions. (Correct answer: “No”; Paraphrase item).

Item 61: The nucleus of a cell, or cellular area in bacteria, contains DNA, RNA, and a variety of proteins. (Correct answer: “No”; Paraphrase item).

Item 72: In the cytoplasm of virtually all cells we find a nucleus or nuclear area containing cells genetic material (DNA). (Correct answer: “Yes”; Original item).

It was not clear why the four items above did not function well in discriminating the examinees in term of their ability to comprehend the text material. Item 72, for instance, is an original item which means that it is exactly the same as the sentence that appeared on the text material. The p-value indicated that this item average in difficulty yet the correlation showed that some of the higher performers responded to the item incorrectly while some of the lower performer answered it correctly. In fact 68.8% of the higher performers (top 50%) answered the item incorrectly while 57.9% of the lower performers (bottom 50%) answered the item correctly.

3.2.3 Difference in Item Statistics

As mentioned earlier, the Bio-SVT consists of four item types distributed equally across the five passages in test. Motivated by earlier studies (e.g., Tollefson & Tripp, 1983, 1986; White & Carcelli, 1982; see also Haladyna, 1990) indicating that, in general, different item formats were found to have different effects on examinee performance, the present study was also designed to examine whether different item types in the Bio-SVT test have different effects on student performance. Furthermore, this study also investigated whether these different item types contributed differently to the total reliability of the test. In addition, since the test also consisted of five different
passages, the same analysis was also performed by considering the passage grouping as the independent variable.

After transforming the data, a general factorial analysis of variance was performed where item type was the independent variable and both item difficulty level and item-total correlation were the dependent variables. Another general factorial analysis of variance was also performed using the five passages in the test as the independent variable.

The result shows that there was a significant difference in item-total correlations among the four item types (F_{3,76} = 3.46; p < .05). In term of item difficulty, there was also a significant difference among the four item types (F_{3,76} = 19.15; p < .001) indicating that the four item types were not equal in difficulty. A more detailed examination of each of the parameter’s confidence interval shows that the “meaning change” item was more difficult than the other three item types and that the “distractor” item was, on average, correlated higher with the total score than the other three item types.

Applying the same procedures using passages as the independent variable, yielded a finding of no significant difference in item-total correlation among the five different passages (F_{4,75} = .24; p > .1). The same result was observed using item difficulty level in each passage as the dependent variable (F_{4,75} = 1.09; p > .1) suggesting that the five passages were about equal in difficulty.

3.3 Validation Study

The study to investigate the validity of the SVT as a predictor variable was divided into two studies. The first study was related to establishing criterion-related
validity by using student’s performance in the Bio-SVT to predict their performance in a college level introductory biology class. In addition to the first study, the second validation study was designed to investigate the number of latent variables being measured by the Bio-SVT. The motivation behind this particular study was to initiate a preliminary construct validation study using a factor analysis technique. Results of these two studies will be presented in the next two sections.

3.3.1 Using the Bio-SVT to Predict College Performance

The motivation for developing the Bio-SVT was to determine if it is possible to use a sentence verification technique type of test to predict a student’s class performance. Previous research showed that experts in domain specific areas perform better on SVT tests than novices. In this regard, the present study collected predictive validity evidence on the use of the SVT to predict domain specific academic performance. The questions this study tried to answer were:

1. Is Bio-SVT test performance related to a student’s performance in a regular class at the college level?

2. Does each of the item types in the Bio-SVT contribute significantly in predicting student’s class performance?

3. Does each of the passages in the Bio-SVT contribute significantly in predicting student’s class performance?

To answer these questions, correlation and regression analyses were performed on the same set of data previously used after discarding items that correlated negatively or had low positive correlations with the total test scores. In this analysis, total SVT scores, item types, and passages were considered as the independent variable or
predictors, while final grade in introductory biology class was considered as the criterion variable. The overall data used in this analysis are presented in Table 3.9.

Table 3.9
Descriptive Statistic of the Selected Data

<table>
<thead>
<tr>
<th>Statistics</th>
<th>(n = 195)</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Scale</th>
<th>N Items</th>
<th>Mean</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biology grades</td>
<td></td>
<td>78.72</td>
<td>11.98</td>
</tr>
<tr>
<td>Overall SVT</td>
<td>53</td>
<td>39.83</td>
<td>6.03</td>
</tr>
<tr>
<td>Item types:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Original</td>
<td>12</td>
<td>10.25</td>
<td>1.45</td>
</tr>
<tr>
<td>Distractor</td>
<td>18</td>
<td>14.49</td>
<td>3.20</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>11</td>
<td>8.19</td>
<td>1.64</td>
</tr>
<tr>
<td>Meaning Change</td>
<td>12</td>
<td>6.99</td>
<td>2.64</td>
</tr>
<tr>
<td>Passages:</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Passage-1</td>
<td>12</td>
<td>8.33</td>
<td>1.78</td>
</tr>
<tr>
<td>Passage-2</td>
<td>13</td>
<td>10.37</td>
<td>1.80</td>
</tr>
<tr>
<td>Passage-3</td>
<td>10</td>
<td>7.77</td>
<td>1.58</td>
</tr>
<tr>
<td>Passage-4</td>
<td>9</td>
<td>7.30</td>
<td>1.43</td>
</tr>
<tr>
<td>Passage-5</td>
<td>9</td>
<td>6.05</td>
<td>1.91</td>
</tr>
</tbody>
</table>

With respect to the first question, Table 3.10 shows results of the regression analysis indicated that there was a significant relationship between student performance on the Bio-SVT and performance in the introductory biology class (\( F_{1,193} = 102.04; \ p < .01 \) ).
Table 3.10  
Anova Table for Regression Analysis  
between SVT scores and Final Grades in Introductory Biology Class  
(n=195)  

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>1</td>
<td>9634.63</td>
<td>9634.63</td>
<td>102.04</td>
</tr>
<tr>
<td>Residual</td>
<td>193</td>
<td>18223.85</td>
<td>94.42</td>
<td></td>
</tr>
</tbody>
</table>

A detailed examination of the data showed that performance on the Bio-SVT was correlated .58 with performance in the introductory biology class, accounting for 34.5% of the total variance. In addition, it should be noted that this correlation was based on a subset of the data (n=195 out of the total n=461). An examination of the Bio-SVT scores of students who withdrew from the class showed that these students scores were significantly lower than those who remained in the class until the end of the semester (F_{1,447} = 13.03; p < .01). The actual correlation between students performance on the Bio-SVT and their performance on the introductory biology class, therefore, was likely to be higher than the one reported above.

With regard to the second and third questions, Table 3.11, Table 3.12 and Table 3.13 show results of the correlation and step-wise regression analyses between each of the item types and each of the passages to the final grades in the introductory biology class.
Table 3.11
Correlation between Passages, Item types, and Total scores with the Final Grade in Introductory Biology Class

<table>
<thead>
<tr>
<th>Item Types</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>.44</td>
</tr>
<tr>
<td>Distractor</td>
<td>.36</td>
</tr>
<tr>
<td>Paraphrase</td>
<td>.33</td>
</tr>
<tr>
<td>Meaning Change</td>
<td>.39</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Passage</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage-1</td>
<td>.40</td>
</tr>
<tr>
<td>Passage-2</td>
<td>.44</td>
</tr>
<tr>
<td>Passage-3</td>
<td>.38</td>
</tr>
<tr>
<td>Passage-4</td>
<td>.42</td>
</tr>
<tr>
<td>Passage-5</td>
<td>.41</td>
</tr>
<tr>
<td>Total SVT</td>
<td>.58</td>
</tr>
</tbody>
</table>

Table 3.12
Regression analysis of Item types and Final Grade in Introductory Biology Class (n = 195)

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>SE B</th>
<th>95% LL</th>
<th>95% UL</th>
<th>Stand. B</th>
<th>T</th>
<th>Sig. T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original</td>
<td>6.09</td>
<td>1.57</td>
<td>2.98</td>
<td>9.19</td>
<td>.27</td>
<td>3.87</td>
<td>.0002</td>
</tr>
<tr>
<td>Distract.</td>
<td>2.98</td>
<td>.84</td>
<td>1.31</td>
<td>4.64</td>
<td>.23</td>
<td>3.53</td>
<td>.0005</td>
</tr>
<tr>
<td>Paraph.</td>
<td>3.74</td>
<td>1.24</td>
<td>1.29</td>
<td>6.19</td>
<td>.20</td>
<td>3.01</td>
<td>.0029</td>
</tr>
<tr>
<td>M.Chage</td>
<td>3.13</td>
<td>.78</td>
<td>1.58</td>
<td>4.68</td>
<td>.25</td>
<td>3.98</td>
<td>.0001</td>
</tr>
</tbody>
</table>
Table 3.13
Regression analysis of Passages and Final Grade in Introductory Biology Class

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>SE B</th>
<th>95% LL</th>
<th>95% UL</th>
<th>Stand. B</th>
<th>T</th>
<th>Sig. T</th>
</tr>
</thead>
<tbody>
<tr>
<td>Passage1</td>
<td>3.04</td>
<td>1.56</td>
<td>-0.04</td>
<td>6.12</td>
<td>0.14</td>
<td>1.94</td>
<td>0.0537</td>
</tr>
<tr>
<td>Passage2</td>
<td>4.35</td>
<td>1.62</td>
<td>1.16</td>
<td>7.53</td>
<td>0.19</td>
<td>2.69</td>
<td>0.0078</td>
</tr>
<tr>
<td>Passage3</td>
<td>2.52</td>
<td>1.44</td>
<td>0.33</td>
<td>5.36</td>
<td>0.12</td>
<td>1.75</td>
<td>0.0824</td>
</tr>
<tr>
<td>Passage4</td>
<td>5.22</td>
<td>1.80</td>
<td>1.66</td>
<td>8.77</td>
<td>0.19</td>
<td>2.89</td>
<td>0.0042</td>
</tr>
<tr>
<td>Passage5</td>
<td>3.56</td>
<td>1.18</td>
<td>1.23</td>
<td>5.89</td>
<td>0.19</td>
<td>3.01</td>
<td>0.0029</td>
</tr>
</tbody>
</table>

A detailed look at the regression coefficients showed that each of the four item types contributed significantly to the prediction equation over and above the others. In term of the passages, however, even though the first passage correlated .40 with the criterion measure it did not contribute significantly to the prediction equation over and above the other passages (t_{194} = 1.94; \ p > .05). The same case was also true for the third passage (t_{194} = 1.75; \ p > .05).

3.3.2 Factor Analysis Study

The next question the present study addressed was related to the number of latent variables being measured by the Bio-SVT. To answer this question, a factor analysis technique was utilized. The data was set up by averaging items that belong to the same item types category. Because there were four item types included within each passage, this grouping resulted in a 461 by 20 data matrix to be analyzed.

Exploratory factor analysis yielded five different factors that had eigenvalues larger than 1.0. Table 3.14 shows the rotated (Varimax) factor matrix.
Table 3.14
Rotated Factor Matrix

<table>
<thead>
<tr>
<th></th>
<th>Factor 1</th>
<th>Factor 2</th>
<th>Factor 3</th>
<th>Factor 4</th>
<th>Factor 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR1N</td>
<td>.06131</td>
<td>.24890</td>
<td>.13367</td>
<td>.10368</td>
<td>.19814</td>
</tr>
<tr>
<td>OR2N</td>
<td>.19616</td>
<td>.37213</td>
<td>.08028</td>
<td>.14229</td>
<td>.15095</td>
</tr>
<tr>
<td>OR3N</td>
<td>.21698</td>
<td>.33555</td>
<td>.12018</td>
<td>.13235</td>
<td>-.01001</td>
</tr>
<tr>
<td>OR4N</td>
<td>.10680</td>
<td>.36748</td>
<td>.20322</td>
<td>.14293</td>
<td>.08785</td>
</tr>
<tr>
<td>OR5N</td>
<td>.22742</td>
<td>.51110</td>
<td>-.36250</td>
<td>-.32662</td>
<td>.16057</td>
</tr>
<tr>
<td>DR1N</td>
<td>.52832</td>
<td>-.26984</td>
<td>-.00285</td>
<td>.17589</td>
<td>.26892</td>
</tr>
<tr>
<td>DR2N</td>
<td>.62250</td>
<td>-.24443</td>
<td>-.13213</td>
<td>.18658</td>
<td>.15551</td>
</tr>
<tr>
<td>DR3N</td>
<td>.56237</td>
<td>-.19232</td>
<td>-.14775</td>
<td>.24200</td>
<td>.04194</td>
</tr>
<tr>
<td>DR4N</td>
<td>.60058</td>
<td>-.12622</td>
<td>-.04610</td>
<td>.17505</td>
<td>-.27920</td>
</tr>
<tr>
<td>DR5N</td>
<td>.61611</td>
<td>.22274</td>
<td>-.24275</td>
<td>-.15411</td>
<td>-.24810</td>
</tr>
<tr>
<td>PR1N</td>
<td>.12344</td>
<td>.29589</td>
<td>.09904</td>
<td>.02012</td>
<td>.05120</td>
</tr>
<tr>
<td>PR2N</td>
<td>.12387</td>
<td>.37215</td>
<td>.19575</td>
<td>.26242</td>
<td>-.01005</td>
</tr>
<tr>
<td>PR3N</td>
<td>.05268</td>
<td>.47509</td>
<td>.24461</td>
<td>.09363</td>
<td>-.04998</td>
</tr>
<tr>
<td>PR4N</td>
<td>-.07635</td>
<td>.49045</td>
<td>.04893</td>
<td>.33007</td>
<td>-.08757</td>
</tr>
<tr>
<td>PR5N</td>
<td>.08248</td>
<td>.42997</td>
<td>-.07221</td>
<td>.06604</td>
<td>.02043</td>
</tr>
<tr>
<td>MR1N</td>
<td>.30008</td>
<td>-.00172</td>
<td>-.39949</td>
<td>-.25432</td>
<td>-.05877</td>
</tr>
<tr>
<td>MR2N</td>
<td>.35582</td>
<td>.00110</td>
<td>-.39748</td>
<td>-.14287</td>
<td>-.05854</td>
</tr>
<tr>
<td>MR3N</td>
<td>.32571</td>
<td>-.09254</td>
<td>.23426</td>
<td>-.10634</td>
<td>.01504</td>
</tr>
<tr>
<td>MR4N</td>
<td>.38108</td>
<td>-.07475</td>
<td>.24170</td>
<td>-.18056</td>
<td>-.05452</td>
</tr>
<tr>
<td>MR5N</td>
<td>.41669</td>
<td>-.04087</td>
<td>.22560</td>
<td>-.37284</td>
<td>.15068</td>
</tr>
</tbody>
</table>

Even though Table 3.13 shows that there were five factors that had eigenvalues greater than 1.0, the Scree Plot depicted in Figure 3.2 below indicates that there were basically three dominant factors underlying students performance on the test.

![Factor Scree Plot](image-url)

Figure 3.2 Factors scree plot
The result of the preliminary factor analysis study mentioned above suggest that the Bio-SVT seems to measure two dominant latent variables accounting for 26.4% of the total variance. Figure 3.3 shows the resulted factor structures:

![Factor structures of the Bio-SVT](image)

Figure 3.3  Factor structures of the Bio-SVT

The factor structures depicted in Figure 3.3 indicated that all the four variables loaded on the first factor. This factor, accounting for 17.9% of the total variance, was interpreted as the “Comprehension” factor, which is the factor that the test purports to measure.

The second factor, accounting for 8.5% of the total variance is interpreted as some kind of general ability factor, considering that this second factor is the factor that gave rise primarily to student’s performance on “Meaning Change” and “Distractor Items” and that both the “Meaning Change” items (the most difficult items) and the “Distractor” items required a more complicated thinking process than the other two item types in the test. Furthermore, this interpretation is also consistent with the notion of
theory of the assessment of reading comprehension that a test of reading comprehension would, to some extent, also measure some kind of a general ability.

Kyllonen (1996) once argued of the close relationship between the working memory capacity and general intellectual ability. He also noted that the working memory factor is the central factor in cognitive ability (pp. 72) and might serve as sources of individual differences (pp. 54). Following his argument, the second factor then can be attributed as the working memory capacity factor. This interpretation is consistent with the observation that both “Meaning Change” and “Distractor” items require more complicated cognitive processes (i.e. comprehending the item and compare it to the memory representation of the reading material) which would also mean that student’s correct or incorrect responses to these items is a function of each individual student’s working memory capacity.
CHAPTER 4
SUMMARY, CONCLUSION, DISCUSSION, AND IMPLICATION FOR FURTHER RESEARCH

The present study had three general purposes. The first was to systematically investigate the psychometric properties of Bio-SVT, a variant of the general SVT, from a classical test theory (CTT) perspective. In particular, the study was aimed at investigating the reliability of the Bio-SVT and examining the item statistics of the test.

For these purposes, a domain specific SVT test, the Bio-SVT, was developed and, utilizing the test, empirical data were collected from a community college in Western Massachusetts. Cronbach alpha estimates of internal consistency and "Split Half" reliability estimates were used to study the test reliability. The item statistics were obtained by computing the item difficulties and the corrected item-total correlation.

The result of the study indicated that the reliability of the test was .71, which seemed to be at the lower bound considering that the test was five passage length. Previous research indicated that the average reliability for a five passage test length was in the neighborhood of 0.90.

This rather modest level of reliability was thought, in part, to be directly related to the fact that six items correlated negatively with the test as a whole, thus functioning against the rest of the items in the test.

Removing the six items that correlated negatively with the rest of the items increased the reliability to .74. The six items, therefore, were excluded from the rest of the investigation.
Furthermore, about 25% of the items were found to be correlated not significantly greater than zero with the rest of the items. Strictly speaking, these items did not contribute anything to the total reliability of the test. Revising these items would certainly increase test reliability.

Before removing the six negatively discriminating items, a detailed analysis was performed to examine if there were differences in item difficulty and item-total correlations among item types and among passages. A general factorial model of analysis of variance technique was chosen to examine the data.

The result indicated that there were no significant differences in item difficulty and item discrimination among the passages. There was, however, a significant difference in item difficulty between the item types. A detailed examination showed that the meaning change items were more difficult than the other three item types.

This result was not surprising. Meaning change sentences had the same theme as the passage, used the same sentence structure, and used almost the same words. Yet the meaning of the sentence was altered so that examinee were more likely to be false alarmed (respond “Yes” to a “No” sentence). Carlisle (1989) once argued that “poor comprehenders tended to respond on the basis of what they knew about the topic or what they thought the passage should have said, rather than what it did say”. Thus, poor comprehenders were less likely to correctly verify the meaning change items while they had higher probability to correctly verify the other item types. The result, then, was that the proportion of examinees that answered the meaning change items correctly were less than the proportion of examinees that answered the other item types correctly.
An item analysis was also performed to examine the item difficulty and item discrimination indices. The p-value, which is the proportion of examinees answering the item correctly, was used as the item difficulty index, while the corrected item-total correlation was used as the item discrimination index. Results indicated that 27 items in the test had unacceptable item discrimination indices. Further examination of these 27 items indicated that even though in term of difficulty, types, and location in the passage the differences among these items and the rest of the items in the test was imperceptible, an item by item examination indicated that the items shared the following characteristics:

1. The items were derived from sentences that were relatively unimportant to the main idea of the passage.
2. The items were derived from sentences that, if read separately from the whole passage, were related more to subjects other than biology.
3. The items were derived from sentences that were too long, thus making these sentences too difficult to comprehend.
4. Distractor items were taken from the passage addressed the same topics as the one presented in the reading passage.
5. The items were derived from sentences stating general background knowledge.
6. The items were simply too.

From a test development perspective, these findings suggest that in developing the SVT, the test developer needs to consider the following details:

1. The reading should be selected from the passage that were succinctly written where almost all sentences in the passage are about equally important. Passages that have many illustrations and multiple anecdotal examples would increase the possibility of
producing items that have a low item discrimination index. One possible way to proceed with test development is to exclude several unimportant sentences in the passage from the test through an expert rating procedure. This suggestion implies the possibility of deriving a 16 item test from an 18 or 19 sentence length passage. Further investigation, however, need to be conducted to examine if a slight difference in the number of sentences in the passage would not make the passage more difficult to comprehend.

2. Avoid using passages that contain sentences that are too long for it is also going to make items that are too long.

3. Distractor items should be taken from passages that are not too close in content to the reading passage, yet still addressing the same topics.

Another interesting finding of the study was that, on average, the distractor items correlated higher with the total score than the other three item types. These distractor items were also the most reliable items. As described earlier, distractor items are items containing true statements. They, however, were not derived from the reading passage. Distractor items were sentences that had a similar syntactic structure yet differed in meaning from any original sentence in the reading passage. To be able to correctly verify distractor items, therefore, one should be able to (1) comprehend the text material and preserved the meaning in to some kind of memory representation, (2) comprehend the distractor sentence and also preserve its meaning, and (3) compare the two representations to see that the distractor sentence was not belonging to the reading passage. Clearly, only the higher comprehenders would be able to perform these more complicated tasks. This more complicated process involved in solving distractor items
explained why the items had higher correlations with the total scores than the other item types for the reason that higher comprehenders would also tend to get a higher score on the overall test.

The second purpose of the study was to investigate the validity of the use of student performance in the Bio-SVT to predict their Biology class performance. The criterion measure used in this study was final grades in an introductory Biology class. For the purpose of this study, the Springfield Technical Community College located in Western Massachusetts provided a total score for each student before the score was transformed to letter grade. The test scores earned by students in the introductory Biology class ranged from 1 to 100 which enhanced score variability over score variability available with letter grades.

The results of the study showed that student performance on the Bio-SVT predict significantly performance in the introductory Biology class. A detailed examination of the data showed that performance on the Bio-SVT accounted for about 35% of the variance of the final grades in introductory Biology.

This result was as expected. Royer et al.(1987) showed that knowledge specific SVT tests could be used to predict specific learning performance. It was reported that relevant SVT tests were significant predictors of course performance while irrelevant SVT tests were not. In this regard, the present study replicated the Royer et al.(1987) study, based on different empirical data. The two findings, therefore, can be taken as confirming the used of a domain specific SVT tests for predicting relevant future learning performance.
This finding suggests what the Bio-SVT test are measuring. In the introductory chapter it was mentioned that the SVT test was intended to provide a measure of the extent to which a reader could extract meaning from a text. Previous literature reviewed earlier suggests that this reader comprehension ability was affected by the individual’s background knowledge. The finding that a domain specific SVT test can be used to predict relevant domain specific learning performance is consistent with this notion, suggesting that the Bio-SVT test was also measuring the individual’s background knowledge in Biology.

The third question this study addressed was concerning the underlying latent variable measured by the test. A factor analysis study was conducted to answer this question. Items were grouped into its item types within each paragraph resulting in 20 observed variables to be analyzed.

The results showed that The Bio-SVT was measuring two dominant factors. The factor structures were that all the item types were primarily loading on the first factor and, in addition to that, the “Meaning Change” and “Distractor Items” were loading on the second factor.

The first factor, accounting for 17.9% of the total variance, was interpreted as the “Comprehension” factor, which is the factor that the Bio-SVT test purported to measure. The second factor, accounted for 8.5% of the total variance, was interpreted as the working memory capacity factor. This interpretation of the second factor was derived by adopting Kyllonen (1996) notion concerning the concept working memory and its closeness to the general ability construct. Moreover, this interpretation of the second factor was also supported by another finding in the present study which was that
both “Meaning Change” and “Distractor” items require more complicated cognitive processes (i.e. comprehending the item and compare it to the memory representation of the reading material) meaning that student’s correct or incorrect responses to these items also depend on each individual student’s working memory capacity.

There appear to be at least two implications for further research. The first relates to the development of the Bio-SVT. The present study indicated that more that 20% of the items contained in the Bio-SVT test had either low correlation or correlated negatively with total scores. Looking into each individual item, it was hypothesized that there are at least six factors that would make a “bad” Bio-SVT test item. Further research might be directed toward investigating each of these six factors.

The second future research agenda is related to the latent variables measured by the Bio-SVT. Using an exploratory factor analysis technique, the present study found that there were at least two latent variables underlying student performance on the Bio-SVT test. The measurement model of the relationship between the observed variables and the underlying latent variables, however, was not exhaustedly investigated in the present study so that further research is warranted to comprehensively investigate the nature and structure of the underlying latent variables measured by the Bio-SVT.

One of the limitations of the present study was that by using an exploratory factor analysis technique it was assumed that the relationship between the observed variable and the latent variables is linear, which is, in most cases, not quite accurate. The well known “item characteristic curve” within the framework of item response theory, for instance, shows that the relationship between items and the latent ability is not linear. If the measurement model underlying the Bio-SVT is to be investigated further,
therefore, this non-linear relationship between the observed and latent variables would also need to be considered.
REFERENCES


Assessment of Reading Comprehension

This test examines how well you understand material that you read.

You will read a passage and then you will turn the page and evaluate 16 test sentences. Your task is to evaluate whether each of the 16 test sentences have the same or a different meaning than a sentence that was contained in the passage you have read.

Mark "A" for "YES" on your answer sheet if the test sentence has the same meaning as the sentence in the passage.

Mark "C" for "NO" on your answer sheet if the sentence has a different meaning than the sentence in the passage.

Let's try a couple to make sure you have the idea.

Read the short paragraph below, slowly and carefully.

**Portrait of a Teacher**

She was thirty-four. She wore a white skirt and yellow sweater and thin gold necklace, which she held in her fingers as if holding her own reins, while waiting for children to answer. Her hair was black with a hint of Irish red. It was cut short to the tops of her ears, and swept back like a pair of folded wings.

Now turn the page and answer the test sentences.
Answer the following test sentences on your answer sheet. Mark "A" (for Yes) on your answer sheet if you believe that the test sentence and a passage sentence have the same meaning. Mark "C" (for No) on your answer sheet if you believe they have a different meaning. Do not turn back to the passage.

Sample Test Sentences

1. Her hair was black, with a hint of Irish red.
2. It was cut long, to the tops of her shoulders, and swept back like a pair of wings.
3. The woman was thirty-four years old.
4. She was a fifth-grade teacher.

Answers

1. Sentence 1 should have been answered YES. It is identical to the second sentence in the paragraph. Therefore you should mark "A" on number 1 on your answer sheet.
2. Sentence 2 should have been answered NO. A sentence in the paragraph said that her hair was cut short to the top of her ears. Therefore you should mark "C" on number 2 on your answer sheet.
3. Sentence 3 should have been answered YES. It is a paraphrase of a sentence appearing in the paragraph. Therefore you should mark "A" on number 3 on your answer sheet.
4. Sentence 4 should be answered NO. The paragraph said nothing about what grade the teacher taught. Therefore you should mark "C" on number 4 on your answer sheet.

If you are uncertain about how to complete the tests, or if you have any other questions, please raise your hand.

If you have no questions you may continue. You will read five passages and take tests based on those passages. Read each passage carefully, and then turn the page and respond to the test questions. Once you have read the passage and turned to the test, do not turn back to the passage.

Turn the page and begin.
Atoms: The Basic Building Block of Matter

An atom is the smallest unit of an element that retains all of the element's properties. Using the extraordinary physical forces of nuclear accelerators, atoms can be split into three main kinds of particles: protons, neutrons, and electrons. Imagine an atom the size of a football field with an orange on the center of the fifty-yard line. The orange represents the atom's nucleus, where positively charged protons and uncharged neutrons cluster together. The much tinier, negatively charged electrons whiz around the nucleus at nearly the speed of light. They move so fast that, just as the whirring blades of a helicopter appear as a solid blur, the zipping electrons would appear to occupy most of the football field's space if we could see them. Electrons move in orbits or spaces called electron shells, and each electron shell can hold only a certain number of electrons. Because electrons have negative charges, they are drawn to the positively charged protons in the nucleus, and this attraction holds the atom together. An atom contains equal numbers of electrons and protons, so it's net electric charges is zero.

The number of protons in an atom determines what element it is and forms the element's atomic number. An element's atomic mass is the sum of its numbers of protons and neutrons. Electron's are so light that their mass isn't counted.

WHEN YOU HAVE FINISHED, TURN THE PAGE AND ANSWER THE TEST QUESTIONS. DO NOT TURN BACK TO THE STORY.
Read carefully each of the test sentences below. Mark "A" for "YES" on your answer sheet if the sentence has the same meaning with the sentence in the story. Mark "C" for "NO" on your answer sheet if the test sentence has a different meaning than a sentence in the story.

Start your answer with number 5 on your answer sheet.

5. Imagine an atom the size of a football field with an orange on the center of the fifty yard line.
6. A stable or chemically inert atom does not react with other elements.
7. The much tinier, negatively charged electrons whiz around the nucleus at nearly the speed of light.
8. The electrons are so dense that, just as the whirring blades of a helicopter appear as a solid blur, the zipping electrons would appear to occupy most of the football field's space if we could see them.
9. The smallest part of an element that has all of the properties of that element is called an atom.
10. Using the extraordinary physical forces of nuclear explosions, atoms can be split into three main kinds of particles: protons, neutrons, and electrons.
11. The number of electrons orbiting the nucleus of an atom influences how readily it will react with other atoms.
12. Imagine that the orange is the nucleus of the atom which consists of positively charged protons and neutrally charged neutrons clustered together.
13. The number of electrons in an atom determines what element it is and forms the element's atomic number.
14. Because electrons have negative charges, they are drawn to the positively charged protons in the nucleus, and this attraction holds the atom together.
15. Electrons move in orbits or spaces called electron shells, and each electron shell can hold only a single electron.
16. Oxygen and nitrogen are much more electronegative than hydrogen.
17. An element's atomic mass is the sum of its numbers of protons and neutrons.
18. An atom consist of an equal number of negatively charged electrons and positively charged protons meaning that the sum of the electric charges in the atom is zero.
19. Elements are least reactive when their outermost electron shell contains a stable number of electrons.

20. Electrons are so small that they contribute nothing to an element's atomic mass.
Lipids: Nature’s Miracle Compound

Lipids are organic compounds that share one distinguishing property: They are nonpolar and so do not dissolve appreciably in water. Lipids contain mostly carbon and hydrogen, with a very small portion of oxygen. Some also incorporate phosphorus and nitrogen. Because they repel water, lipids such as oils and waxes are found as waterproof coatings on the outer surfaces of many organisms.

Because lipids are insoluble in water, they are vital components of the membranes that separate living cells from one another and from their surroundings. Lipids also offer a unique way to store energy. They contain a high proportion of energy-rich, carbon-hydrogen bonds. This allows lipids to contain six times as much energy as the same mass of carbohydrates, another source of energy for organisms. This is undoubtedly why lipids have become increasingly important food reserves for plants and animals during the course of evolution. Before making their fall nonstop migrations to warmer climates, many songbirds gain twice their mass in fat reserves. Some pre-migratory birds are so fat that they can hardly open their eyelids. If birds had to rely upon carbohydrates plus the necessary storage water for this same amount of stored energy, they would be so heavy that they couldn't fly.

------------------------
WHEN YOU HAVE FINISHED, TURN THE PAGE AND ANSWER THE TEST QUESTIONS. DO NOT TURN BACK TO THE STORY.
Read carefully each of the test sentences below. Mark "A" for "YES" on your answer sheet if the sentence has the same meaning with the sentence in the story. Mark "C" for "NO" on your answer sheet if the test sentence has a different meaning than a sentence in the story.

Start your answer with number 21 on your answer sheet.

21. Most lipids do not incorporate phosphorus and nitrogen.
22. Because lipids are insoluble in water, they are vital components of the membranes that separate living cells from one another and from their surroundings.
23. Fatty acids are the simplest lipids.
24. Lipids also provide an important way to store energy.
25. Lipids are organic compounds that share one distinguishing property: They are nonpolar and easily soluble in water.
26. Fat and oils are lipids used to store energy reserves in plants and animals.
27. Lipids repel water, and therefore form waterproof barriers on the exterior surface of many organisms.
28. Lipids contain mostly carbon and hydrogen, with a very small portion of oxygen.
29. Lipids contain a high proportion of energy-rich, carbon-hydrogen bonds.
30. Before they head South for the winter, some birds become so fat they can barely open their eyes.
31. This means that lipids contain six times as much energy as the same mass of protein, another source of energy for organisms.
32. Proteins make up more than 50% of the dry mass of animals and bacteria, and perform many important functions in living organism.
33. Before making their fall nonstop migrations to warmer climates, many songbirds gain twice their mass in added plumage.
34. If the pre-migratory birds had to use carbohydrates rather than fats as a source of stored energy, they would be so bulky they wouldn’t be able to get off of the ground.
35. Proteins are long, unbranched chains of amino acids, but they may fold up into complex shapes.

36. This is undoubtedly why lipids have become increasingly important food reserves for plants and animals during the course of evolution.

WHEN YOU HAVE FINISHED, TURN THE PAGE AND READ THE NEXT STORY.
Cells: The Building Blocks of Living Organisms

Cells were first scientifically observed and reported in 1655 by the Englishman, Robert Hooke. What Hooke actually saw in the bark of the cork oak tree were empty, dead cell walls, lacking the living matter they once contained. Other early microscopists soon observed cells in all kinds of plants. They found similar structures in animals, too, but animal cells were harder to distinguish because they lack the thick cell walls that surround plant cells. Observers also reported the existence of many tiny organisms consisting of only one cell each.

Eventually, biologists recognized the main features of the cell theory:
1. All organisms consist of one or more cells.
2. Cells are the fundamental units of life - the smallest entities that can be called "living."
3. Cells arise only by division of existing cells.

Many organisms are unicellular (consisting of only one cell). However, a cell's size is limited. The biochemical reactions of its metabolism require raw materials from outside the cell and generate waste products that must be expelled. So, the cell is constantly trading chemicals with its environment.

----------------------------------
WHEN YOU HAVE FINISHED, TURN THE PAGE AND ANSWER THE TEST QUESTIONS. DO NOT TURN BACK TO THE STORY.
Read carefully each of the test sentences below. Mark "A" for "YES" on your answer sheet if the sentence has the same meaning with the sentence in the story. Mark "C" for "NO" on your answer sheet if the test sentence has a different meaning than a sentence in the story.

Start your answer with number 37 on your answer sheet.

37. A living cell keeps its chemical composition steady within narrow limits.
38. Observers also reported the existence of many tiny organisms consisting of only a few cells.
39. All biological membranes have similar structures and functions, whether they are plasma membranes or membranes of organelles inside the cell.
40. All organisms consist of one or more cells.
41. Living cells were first scientifically observed and reported in 1655 by the Englishman, Robert Hooke.
42. Other early microscopists soon observed cells in all kinds of plants.
43. What Hooke saw was not living cells, but rather dead and empty cell walls forming the bark of the cork tree.
44. The early microscopists also found cells in animals, but they were harder to see because animal cells did not have the thick walls that were present in plant cells.
45. A cell's size is limited.
46. Many kinds of animal cells have direct cytoplasm-to-cytoplasm connections at gap junctions.
47. Cells are the fundamental unit of matter - the smallest entities that can be called "living or non-living."
48. The biochemical reactions of cellular reproduction require raw materials from outside the cell and generate waste products that must be expelled.
49. Cells are created when other cells divide.
50. Chemicals are constantly moving back and forth through the cell walls.
51. Many organisms are unicellular (consisting of only one cell).
52. All cells take in macromolecules or particles by endocytosis.

WHEN YOU HAVE FINISHED, TURN THE PAGE AND READ THE NEXT STORY.
Cell Functioning and Structure

All cells must perform certain basic tasks. In addition, each cell of a multicellular organism makes a specialized contribution to the body as a whole. For example, a muscle cell in the heart is specialized to contract and help pump blood. Because it is deep inside the body, it cannot capture its own food or obtain oxygen from the air, but must rely on other specialized cells, such as those of the lungs and blood, to provide the food and oxygen it needs. Thus, there is a division of labor among the cells of a multicellular organism.

Most cells have three main parts:

1. The plasma membrane, covering the outside of the cell and controlling what enters and leaves. (In plants, this is surrounded by a tough cell wall.)

2. The cytoplasm ("cell fluid"), containing water, various salts, and organic molecules, including many metabolic enzymes. The cytoplasm also contains a variety of larger structures, collectively called organelles, which perform various tasks. Many of these "little organs" are surrounded by membranes very similar to the plasma membrane.

3. The cell nucleus (in bacteria, the nuclear area), housing the cell's genetic material (DNA and associated RNA and proteins). The genetic material contains directions for making the cell's proteins.

---------------------------

WHEN YOU HAVE FINISHED, TURN THE PAGE AND ANSWER THE TEST QUESTIONS. DO NOT TURN BACK TO THE STORY.
Read carefully each of the test sentences below. Mark "A" for "YES" on your answer sheet if the sentence has the same meaning with the sentence in the story. Mark "C" for "NO" on your answer sheet if the test sentence has a different meaning than a sentence in the story.

Start your answer with number 53 on your answer sheet.

53. Because it is deep inside the body, a muscle cell in the heart cannot capture its own food or obtain oxygen from the air, but must rely on other specialized cells, such as those of the lungs and blood, to provide the food and oxygen it needs.

54. Cells have three major structures.

55. The plasma membrane, present in all cells, is the only membrane found in many prokaryotes.

56. A muscle cell in the heart is specialized to contract and help pump blood.

57. Every cell accomplishes certain fundamental activities.

58. Most of the cell's general metabolism and protein synthesis occur in the cytosol, so it is not surprising that the cytosol is about 20 percent protein - largely metabolic enzymes.

59. There is no division of labor among the cells of a multicellular organism.

60. In addition, each cell of a multicellular organism makes the same contribution to the body as a whole.

61. Also floating in the cytoplasm are bigger structures, called organelles, that perform a variety of functions.

62. The genetic material contains directions for making the cell's proteins.

63. The cytoplasm ("cell fluid"), containing water, various salts, and organic molecules, including many metabolic enzymes.

64. Many of these enzymes are surrounded by membranes very similar to the plasma membrane.

65. The nucleus of a cell, or cellular area in bacteria, contains DNA, RNA and a variety of proteins.

66. The plasma membrane covering the outside of the cell and controlling metabolic processes.
67. A plant cell is surrounded by a thick but porous cell wall, lying just outside the plasma membrane.

68. Some bacteria produce a polysaccharide or polypeptide capsule outside the cell wall.

WHEN YOU HAVE FINISHED, TURN THE PAGE AND READ THE NEXT STORY.
Two Types of Cells

The viscous, fluid cytoplasm forms the ground substance in a living cell. It contains food molecules, metabolic enzymes, and so on. In the cytoplasm of virtually all cells we find a nucleus or nuclear area containing the cell's genetic material (DNA). In the cytoplasm we also see many ribosomes, structures that make the cell's protein according to instructions from the DNA.

There are two basic cell types that differ in their elementary structure: prokaryotic and eukaryotic. Prokaryotic cells have a simpler structure. In particular they lack a discrete nucleus that is separated from the cytoplasm by a membrane and also lack any membrane-bounded organelles. In addition, prokaryotic cells are usually smaller than eukaryotic cells, and they usually have a rigid cell wall outside the plasma membrane. Most prokaryotes are also unicellular - in function if not in structure - because when a cell divides, its offspring may stick together and form clumps or strings of attached but independent cells. For these reasons, prokaryotic cells are believed to have evolved earlier than eukaryotic cells.

An eukaryotic cell contains a nucleus bounded by membrane, as well as other membrane-bounded organelles. Eukaryotic cells make up the bodies of all organisms other than bacteria, which are made of prokaryotic cells.

-----------------------

WHEN YOU HAVE FINISHED, TURN THE PAGE AND ANSWER THE TEST QUESTIONS. DO NOT TURN BACK TO THE STORY.
Read carefully each of the test sentences below. Mark "A" for "YES" on your answer sheet if the sentence has the same meaning with the sentence in the story. Mark "C" for "NO" on your answer sheet if the test sentence has a different meaning than a sentence in the story.

Start your answer with number 69 on your answer sheet.

69. Another type of body we find in the cytoplasm are numerous ribosomes which manufacture cellular protein from a blueprint contained in DNA.
70. The viscous, solid cytoplasm forms the ground substance in a living cell.
71. Some bacteria produce a polysaccharide or polypeptide capsule outside the cell wall.
72. Prokaryotic cells have a more complex structure.
73. Within the cytoplasm one finds bits of cellular food, enzymes enabling energy production, and other cellular material.
74. The plasma membrane, present in all cells, is the only membrane found in many prokaryotes.
75. There are two kinds of cells: prokaryotic and eukaryotic that differ from one another in basic structure.
76. In the cytoplasm of virtually all cells we find a nucleus or nuclear area containing the cell's genetic material (DNA).
77. Most prokaryotes are also parasitic - in function if not in structure - because when a cell divides, its offspring may stick together and form clumps or strings of attached cells.
78. Prokaryote cell walls contain unique polymers of sugars and amino acids.
79. Eukaryotic cell contains a nucleus bounded by membrane, as well as other membrane-bounded organelles.
80. Prokaryotic differ from eukaryotic cells in that they do not have a discrete nucleus and they do not have organelles that are surrounded by membrane.
81. Prokaryotic cells are usually smaller than eukaryotic cells and they usually have a rigid cell wall outside the plasma membrane.
82. Eukaryotic cells make up the bodies of all organisms other than plants, which are made of prokaryotic cells.

83. Prokaryotic cells are believed to have evolved earlier than eukaryotic cells.

84. Cells also contain other components too small to be seen even with the most powerful electron microscope.
YOU HAVE FINISHED THE TEST! PLEASE RAISE YOUR HAND AND SOMEONE WILL PICK UP YOUR PAPER. THANK YOU FOR YOUR COOPERATION.