Uses for the Sentence Verification Technique for Measuring Language Comprehension

James M. Royer

University of Massachusetts, Amherst

Running Head: Measuring Comprehension

I would like to thank Maria Carlo, Cheryl Durwin, Barbara Greene, CarolAnne Kardash, Horace Marchant and Gale Sinatra for their comments on an earlier version of this article, and for their contributions to the research reviewed in the article. Requests for information about this article should be sent to James M. Royer, Department of Psychology, University of Massachusetts, Amherst, MA 01003, or via email to: royer@psych.umass.edu

Abstract

The Sentence Verification Techniques (SVT) is a procedure that non-psychometricians can use to develop reading and listening comprehension tests that can be based on a wide variety of text materials. This article begins with a discussion of how SVT tests are developed, administered, scored and interpreted. The article then briefly reviews research on the reliability and validity of SVT tests. The major section of the article is concerned with purposes that SVT tests can serve. Research is reviewed that uses SVT tests as diagnostic tools, as measures of comprehension in language research, as a means of predicting future learning performance, as a measure of progress in bilingual education programs, and as a means of assessing language skills in developing countries.

Uses for the Sentence Verification Technique of Measuring Comprehension

Educators and educational researchers often have assessment needs that standardized tests cannot fulfill. A number of these needs center around the question of whether a student or participant in a research study has understood a particular text. As instances, a teacher seeking to place a new student in the curriculum would like to zero in on the right level of reading material for the student. A researcher evaluating competing procedures for teaching reading comprehension would like to know if one group can read a criterion text better than another. Researchers preparing a brochure describing the risk factors of a particular medication would like to know if recent immigrants can understand the text when it is read to them. An educational diagnostician evaluating whether a student has a reading disability would like to know if a student's listening comprehension ability exceeds his or her reading comprehension ability.

Each of the situations described above address the question of whether a reader or listener has understood a particular text. This is a different question than determining a reader or listener's general comprehension aptitude. Questions about comprehension aptitude can best be addressed using measures such as standardized tests that have been shown to predict future performance. This article reviews evidence regarding the utility of the Sentence Verification Technique (SVT) in situations where the question being addressed is how well a reader or listener has understood a particular text. The intent of the article is not to compare SVT tests with any other technique for measuring comprehension, and in particular, to not argue for the merits of the SVT relative to other assessment procedures. Rather, the purpose of the article is to review available evidence on uses for SVT tests.

The interest in uses for tests derives from Messick's (1980) argument that test validity is ultimately a question of whether a test is useful for a particular purpose. Much of the research on the SVT technique has been conducted with Messick's argument in mind. That is, a need for an assessment procedure was identified, and research was subsequently conducted to determine if it was possible that SVT tests could fill that need. It should be emphasized that the research reviewed in this article should be viewed with Messick's (1980) admonition in mind that establishing the validity of an assessment procedure for a particular purpose is, "...a continuous, never-ending process developing an ever-expanding mosaic of research evidence." (Pg. 1019) In short, the research to be reviewed does not indicate that SVT tests are valid for a particular purpose; it merely indicates that the evidence collected thus far is consistent with the interpretation of validity for that purpose.

The initial section of the article describes the rationale for using SVT tests as a measure of comprehension, and how SVT tests are developed, administered, scored, and interpreted. The article then briefly examines the reliability of SVT tests before turning to research examining the validity of SVT tests for a variety of purposes.

Developing, Administering and Scoring SVT Tests

Research reviewed in later sections of this article documents the fact that SVT tests can be based on virtually any natural text material, and even some linguistic material that is not text. Tests have been based on excerpts from elementary, middle school, high school and college level textbooks, from newspaper articles, from passages developed for use in basic research, from medical information materials, from abstracts of psychology journal articles, from book reviews published in a weekly magazine, and from manuals used for

military training. Non-text sources for SVT tests have been radio scripts and video tapes of classroom interactions. In short it appears that SVT tests can be based on virtually any natural language source.

Developing SVT Tests

The construction of SVT tests involves developing one of four types of test sentences from each sentence appearing in a text passage. The first type of test sentence is called an original and it is a copy of a sentence as it appeared in the passage.

The second type of test sentence, called a paraphrase, is constructed by changing as many words as possible in an original sentence without altering the meaning of the sentence. The general guidelines for the construction of paraphrase items are: 1) change as many words as possible; 2) make sure the paraphrase has the same meaning as the original sentence; and 3) make sure the meaning of the paraphrase sentence fits with the meaning of the passage as a whole. This last guideline is particularly important since many sentences have a number of acceptable paraphrases that vary in meaning. However, if the item writer writes a paraphrase with both the meaning of the sentence and the meaning of the passage in mind, the number of acceptable paraphrases is greatly constrained.

The third type of test sentence is called a meaning change, and is constructed by changing one or two words in the sentence so that the meaning of the sentence is altered. The general guidelines for constructing meaning change items are: 1) substitute one or two words in the original sentence so that the meaning of the new sentence is distinctly different from the original; 2) make sure the meaning of the sentence is inconsistent with the meaning of the passage as a whole; and 3) make sure the meaning of the test sentence is not bizarre in any way.

Meaning change sentences are the most difficult of the test sentences to write because the nature of a meaning change can vary from being highly obvious to very subtle. The approach that the author has found most fruitful is to attempt to develop meaning change sentences that subjectively occupy a mid-point between these two extremes. In practice, writing a meaning change sentence involves identifying the main idea of a sentence and then changing the words in the sentence so that the main idea is different from the main idea in the original sentence. An attempt is also made to make changes in a manner that would make it very difficult for the examinee to identify the altered sentence without reading the passage, thereby requiring the examinee to read and understand the passage in order to differentiate between original and meaning changed sentences.

The importance of developing meaning change test sentences at the right level was demonstrated in an unpublished study that entailed either changing the main idea of a sentence or changing a minor idea in the sentence. Meaning change sentences having minor idea changes actually had negative discrimination indices. That is, the better examinees performed on the test as a whole, the more likely they were to get the meaning change item wrong (i.e., they said it had the same meaning as an original sentence). It should be noted that meaning change sentences that involve too big of a change also have problems--they are so easy to identify they have no discriminative power.

The final kind of test sentence is called a distractor. A distractor is a sentence that has a syntactic structure that is similar to a sentence in the passage and is consistent with the overall theme of the text passage, but is unrelated in meaning to any sentence that appeared in the passage. The general guidelines for writing distractors are: 1) make sure the distractor is roughly comparable to original sentences in terms of length, syntactic

structure, and vocabulary level; 2) make sure the distractor sentence is consistent with the thematic content of the passage; and 3) make sure the distractor sentence is different in meaning from any original sentence in the passage. One technique that has been useful when developing distractor sentences is to search the text surrounding the target material for suitable distractors. Selection of surrounding sentences frequently satisfies the requirement that the distractor sentences are comparable to original sentences in vocabulary, theme, and syntax.

Meaning Identification Technique (MIT) A variant of the SVT technique, called the Meaning Identification Technique (MIT), has been reported by Marchant, Royer and Greene (1988). An MIT test consists of a set of passages and a test consisting of only two item types: paraphrases and paraphrases that have one or two words changed so that the meaning of the sentence no longer matches the meaning of an original sentence. This type of test sentence is referred to as a meaning change paraphrase. An MIT test is constructed by converting each sentence in an original passage into a paraphrase and then changing the meaning of half of the paraphrases by changing one or two words in the sentence. Research has indicated that MIT tests have somewhat better reliability and validity than SVT tests (Marchant, et al., 1988). This improvement in reliability and validity is associated with the fact that the bulk of the discriminatory power of SVT test sentence types resides in the paraphrase and meaning change items. The down side of using MIT tests is that some of the diagnostic potential of the test is lost when two, rather than four, test item types are used. The reason why this is so will be explained later in the article. Examples of SVT item types and MIT item types are presented in Table 1.

Insert Table 1 About Here

Content of SVT tests. An SVT test generally consists of a set of three to six passages, each followed by a set of test sentences. For school-based assessment activities, the author has used passages that "straddle" the reading ability level of examinees. For example, a six-passage test for fourth grade students might consist of two third-grade passages, two fourth-grade passages, and two fifth-grade passages. The test for each passage consists of equal numbers of each of the test sentence types. So, for example, most of the research using the SVT has used 12 sentence passages and either 12 sentence SVT tests (3 of each of the test sentence types) or 16 sentence SVT tests that are constructed by selecting equal numbers of original, paraphrase and meaning change sentences to represent each of the 12 passage sentences, and then adding four distractor sentences to form a 16 sentence test.

The test sentences in an SVT test are arranged randomly in the test with the restriction that test sentences assessing the first half of the original passage appear first in the test. The purpose of this restriction is to prevent a situation where the first test sentence the examinee encounters is one that had just been read, thereby allowing the examinee to respond to the test sentence using the contents of short-term memory.

The decision to construct tests based on equal numbers of each of the item types was made with several concerns in mind. First, it was thought desirable to have tests constructed with equal numbers of YES and NO items where sentences that mean the same as a passage sentence are YES sentences (originals and paraphrases) and sentences that have a meaning different from a passage sentence are NO sentences

(meaning change and distractors). This rule was followed so that examinees would be unlikely to develop a response bias favoring one response over another. Another consideration was to attempt to prevent an examinee from scoring well using any technique other than evaluating the meaning of the passages and test sentences. Imagine, for instance, a test consisting entirely of meaning changes and paraphrases. A very observant examinee might notice that every time the wording of a test sentence deviates markedly from the wording of a passage sentence (i.e., a paraphrase sentence) the correct response is YES, and every time the wording of the test sentence is very similar to a passage sentence (i.e., a meaning change sentence) the correct response is NO. Thereafter, the examinee could perform very well by simply responding on the basis of remembering similarities in the wording of passage sentences and test sentences. Notice, however, that this strategy does not work if the test also contains original and distractor test sentences. When originals and distractors are included in the test, test sentences that are similar in wording sometimes are YES sentences (originals) and sometimes are NO sentences (meaning changes), and test sentences that are different in wording are sometimes YES sentences (paraphrases) and sometimes NO sentences (distractors). With all four test sentence types in the test, the only way the examinee can reliably differentiate YES and NO sentences is to process the meaning of both the passage and the test sentences.

Having presented the rationale for a balanced SVT test, it should be admitted that it may not be necessary to have <u>equal</u> numbers of each of the test item types. It could very well be, for example, that a 16 sentence SVT test consisting of three originals, three distractors, five meaning changes and five paraphrases would have better reliability and validity than the balanced version of the test because paraphrases and meaning change

sentences have better discriminatory properties than originals and distractors, which are more easily identified. This is an issue for future research to decide.

There is also probably nothing crucial about basing SVT tests on original passages that are 12 sentences in length. Twelve sentence passages were chosen initially because they allow the development of tests that are easily balanced with respect to item type. The author has, on occasion, used passages that were shorter and longer than 12 sentences. For instance, 9 sentence original passages have been used very successfully with very young children (e.g., grade 1 students) who may have difficulty attending to longer passages, and one study has been completed that used passages that were several thousand words long and SVT tests were developed by sampling sentences from the passage to serve as SVT test sentences (Royer, Tirre, Sinatra & Greene, 1989). The SVT tests developed in the study using long passages had relatively poor reliabilities and the author believes that if one wants to test the comprehension of a long passage it is better to divide the passage into segments, have an examinee read a segment and then take a test on the segment, read the next segment and so on.

A team approach to developing SVT tests. The author and his colleagues have trained numerous individuals over the years in SVT test development and have found that virtually any highly literate individual can develop SVT tests having good reliability and validity after approximately 16 hours of training. Examples of individuals trained to develop SVT tests include teachers in bilingual education programs, mainstream classroom teachers, educators from Grenada, Agency for International Development personnel in Guatemala and Belize, graduate students in psychology and education, and undergraduate college students enrolled in an educational research seminar. In addition, the research

reported in this article completed by researchers other than the author suggests that other investigators have been able to develop SVT tests after reading published descriptions of the test development process.

A training and test development process that appears to work well involves a group (four is a good size for groups) of test developers. The process, as the author and his colleagues have used it, begins with a description of the theoretical and practical rationale underlying the SVT procedure, and contrasts this rationale with the rationale underlying traditional comprehension assessment techniques. The purpose of this part of the training sequence is to develop the idea that there are sound empirical and theoretical reasons why the SVT technique measures comprehension.

The next part of the training process entails describing how to select passages upon which to base SVT tests. Emphasis is placed on the idea that passages should be fairly "self-contained" and coherent in that they can be understood in isolation from other text and they should contain a sense of beginning, middle and end. This part of the training process also develops the idea that it is all right to edit passages that don't have these properties and examples are provided of how minor editing can improve the assessment properties of passages occurring in natural text.

The next stage in the training process introduces the four SVT item types and the rules for developing each item type, and provides a rationale for why each type of item is contained in an SVT test. At this point trainees are ready to begin practicing test development. The practice process begins with passage selection and editing. The author has found that the most effective source of passage material for training purposes comes from sources that ultimately will serve as the basis for the SVT tests (e.g., textbooks,

newspaper articles, etc.), or from sources in the trainees' surrounding environment such as excerpts from magazines, or even biblical passages. Using material from the trainees' environment emphasizes the point that the SVT technique is general and can be used to generate tests from a wide variety of source material.

When a passage has been selected to base an SVT test on, each trainee is asked to edit the passage and then share their edited version with the other trainees. An effective way to do this is to copy a passage on an overhead transparency and then write each editorial change in marking ink on the transparency as the trainees suggest them. As editorial recommendations are made, the group comments on the appropriateness of the suggested changes. After an edited version of the passage has been accepted by the group, the trainees copy the first sentence in the passage and each trainee and the trainer create paraphrase and meaning change versions of that sentence. After the sentences have been written, each trainee and the trainer reads his or her paraphrase version of the sentence and it is copied onto an overhead transparency. After all of the sentences have been copied, the group critiques each of the sentences and then selects one sentence or a combination of sentences that they believe is the best paraphrase for the original sentence.

Each member of the group then offers their meaning change version of the sentence and the critique process is repeated. After both paraphrase and meaning change sentences for the first original sentence has been completed, the group then does the same thing for the second sentence, and the process continues until all of the sentences in the original passage have been reviewed. This process generally results in a paraphrase and meaning change version of each original passage sentence that the group as a whole has agreed represents the best test version of the sentences they can collectively develop.

Note that the previous sentence says "generally results in ....." There are occasions where a sentence cannot sensibly be converted into a paraphrase or a meaning change version. This happens often in text for very young children which abounds with sentences like, "No, he said!" When these sentences are encountered, they most often become original test items in the SVT test.

The process described to this point results in the development of three of the four item types (originals, paraphrases and meaning changes) for each sentence in an original passage. The next task is to create distractor sentences. Trainees are given practice in two procedures for developing distractors. The first entails creating a sentence that has the same general syntactic structure as sentences in the original passage and that is consistent in theme with the passage. The second procedure involves searching the text surrounding the original passage (i.e., text before or after the passage segment) for sentences that could serve as suitable distractor sentences. Four distractor sentences per 12 sentence original passage are created using one or both of these procedures.

The trainees are now ready to create the actual test. They are given the rules for test construction (an original, paraphrase or meaning change to represent each original sentence in a passage; 4 added distractors; an equal mix of each test sentence type, and so on) and then told to construct a 16 sentence test. When a test for one passage has been completed, the trainees move on to a second passage, and the process is repeated.

After the trainees have collectively developed tests for at least two passages, they then select two new text segments to base two other passages on and each trainee goes through the process separately, generally as a homework assignment. When the trainees return to the group setting, each displays his or her passage and test on an overhead

transparency and the group critiques both the passage editing and the development of the test sentences. After completing this process, trainees are now ready to develop their own SVT tests.

It should be noted that an advantage of creating paraphrase and meaning change sentences from each original sentence in a passage is that it makes it easy to develop multiple tests from the same passage. If the test developer creates an original, meaning change and paraphrase test sentence for each passage sentence, and then develops three sets of four distractor sentences, he or she will have created the makings for three alternate SVT tests for that passage. That is, in one test version, a sentence might be represented by an original sentence, in a second version by a meaning change, and in a third by a paraphrase. This property of making it easy to create multiple forms of a test is particularly attractive in a situation where one wanted to measure comprehension gains on a particular passage, but was worried about whether examinees would remember their responses from a previous occasion. With multiple SVT forms, the same passage could be administered several times and comprehension each time would be assessed with a different test.

A computer program to assist in the development of SVT tests. Walczyk (1992) has written a computer program that greatly simplifies the process of developing SVT tests. The program, which is written in BASIC, parses passage into sentences and prompts the test developer to produce a test sentence. After the test sentences have been written the program automatically formats the test into printable copy and produces an answer key.

Test Administration

An examinee taking an SVT test reads or listens to successive passages and then, in the absence of the text, judges each of the test sentences to be "yes" or "no" sentences.

Yes sentences are defined as sentences that are the same as or mean the same as passage sentences (originals and paraphrases) and no sentences have a different meaning than passage sentences (meaning changes and distractors).

Another research issue that could be addressed regarding SVT test administration involves the question of whether it makes a difference if examinees are able to freely inspect the test passages while taking and SVT test. As the above paragraph indicates, the common practice is to have the examinee read the passage and then turn a page and respond to the SVT test sentences without being able to return to the passage. The author's intuition is that it would not make a difference in the measurement properties of SVT tests if examinees were allowed to return to the passage. This idea is, however, speculative and should be examined empirically before changing the manner in which tests are administered.

When the author and his colleagues administer listening SVT tests, they make audio tapes of the instructions, the passages, and the test sentences. These tapes are stopped at various points in the instruction phase of the test and the examiner asks questions designed to assure that examinees understand what they are supposed to do. Following the instruction phase, the tape goes nonstop through each passages and set of test questions. It should be noted that one attractive feature of SVT listening comprehension tests is that the entire test process occurs aurally. This contrasts to some listening comprehension tests where examinees listen to a text passage and then answer written multiple-choice questions. This results in a situation where poor performance could be the result of not understanding the aurally presented material, or it could be the result of being unable to read the test questions.

Scoring SVT Tests

Several procedures have been developed for scoring SVT tests. The simplest is to compute proportion correct scores. These scores can be calculated for overall performance, for performance on separate passages, for performance on particular sentences within a passage and for particular test sentence types (e.g., originals, paraphrases, etc.) within or across passages. There are occasions when it is very useful to keep track of scores on individual item types. As an example, Royer, Lynch, Hambleton & Bulgareli (1984) completed Kintschian text analyses (Turner & Greene, 1977) of passages and then separately scored items that were revealed by the text analysis to be important and unimportant. The important items had better discrimination indices than did the unimportant items. It is also useful, on occasion, to separately score the four item types. Later in the article it will be described how differential performance on the item types can be used for diagnostic purposes.

The author and his colleagues have also experimented with asking examinees to rate their confidence in their yes/no judgments as a means of possibly enhancing the reliability of the tests (e.g., Royer, Kulhavy, Lee & Peterson, 1986), though they found little difference between tests where confidence scores were collected and those where they were not.

Another scoring technique that has been examined is to score performance using theory of signal detection parameters (e.g., Swets, Tanner & Birdsall, 1961). Signal detection theory suggests that responding in two-choice discrimination situations is based on two characteristics of the person performing the task: 1) the ability to detect a signal when it is present (in our case, the ability to recognize that a test sentence has the same

meaning as a passage sentence); and 2) the criteria (cutoff) the subject establishes in order to judge a signal is present (the willingness to say a test sentence is a YES sentence).

As an example of the difference between these two parameters, imagine a situation where two examinees had exactly the same ability to detect when a test sentence had the same meaning as a passage sentence, but differed in percent correct performance because one examinee was willing to say YES when he or she "suspected" that the sentence was a yes sentence whereas the second examinee would say YES only when he or she was "certain" the meaning was the same. The willingness to say a signal is present is the cutoff parameter in signal detection analysis and research has shown that the cutoff can be manipulated through payoffs for errors and correct "hits" (saying a signal is present when it is). As an instance, payments for hits with no penalty for false alarms will shift a criterion downward (meaning more trials are judged to have a signal present), whereas penalties for false alarms will shift it upward.

Signal detection theory provides a means of separating the sensitivity and the cutoff components of performance. The two characteristics are expressed by the parameters, d' (signal detection ability) and c (the criteria or cutoff parameter). One study has found that younger children had more stringent cutoff criteria than did older children (Royer, Hastings & Hook, 1979). Thus, d' was a better index of student performance in that study than was proportion correct.

Interpreting SVT performance. As is the case with any measurement procedure, the interpretation of SVT performance is clearest when the performance of an individual student can be compared to the performance of a sample of comparable students. The

author and his colleagues have, however, developed a "feel" for evaluating performance as a function of developing and administering SVT tests in a wide variety of situations. We have generally found that if tests are based on materials that are appropriate for the population to be tested, the average examinee gets about 75% of the items correct, with poor comprehenders scoring in the 70% and lower range (50% correct is chance if all of the items are answered) and good comprehenders scoring in the 80% and above range. As an anecdotal note, in our college student studies we have occasionally seen percent correct scores that are under 20%. Given the care we take to make sure examinees know how to complete the test, these are almost certain to come from students who purposefully answer questions wrong.

If SVT tests are scored using signal detection theory, there is the possibility that test performance could have a criterion-referenced interpretation that is based on an objective standard. The signal detection theory parameter d' can potentially provide a rational basis for establishing criterion-referenced cut-scores. Imagine, for instance, that an examiner wanted to make a yes/no decision about whether an examinee had understood a text. Imagine further that the examiner wanted to express the cut-point in terms of the probability that an examinee would attain a specific score by chance. More specifically, the examiner wanted to set a cut- score that was high enough so that the examinee would attain that score by chance only about 5 times out of 100. A nice property of d' scores is that they have distributional properties similar to Z scores. Hence, a d' score of 1.96 (the .05 level for Z scores) would only be attained by chance approximately 5 times out of 100. At this point, the author does not know of any research that examines criterion-referenced interpretations of SVT performance based on d' scores, but it certainly seems a viable

possibility.

Reliability of SVT Tests

The reliability of SVT tests has been reported in a number of published sources (Greene, Royer & Anzalone, 1990; Marchant, et al, 1988; Royer & Carlo, 1991a; Royer, Tirre, Sinatra & Greene, 1989) and in one dissertation (Sinatra, 1989). In addition, the author and his colleagues have, on numerous occasions, calculated reliability coefficients for tests but, since reliability was not the central thrust of the research, the reliabilities did not appear in the printed article. Finally, reliabilities of tests prepared for use by school systems are routinely calculated. In all cases the author is aware of, the reliability of SVT tests has been evaluated using Chronbach's alpha.

In general, it has been found that SVT tests consisting of three passages and their accompanying 16 sentence tests (48 test sentences) have reliabilities in the .5 to .6 range, SVT tests based on four passages (64 test sentences) generally have reliabilities in the .7 to .8 range, and tests based on 6 passages (96 test sentences) have reliabilities in the .8 to .9 range. The largest study examining the reliability of SVT tests administered 96 item SVT tests to over 1000 students enrolled in grades 3 through 7 (Royer & Hambleton, 1983). The 50 passages used in the study were divided into 24 booklets of 6 passages and each booklet was read by a minimum of 25 and a maximum of 90 students. The reliabilities of the booklets ranged from .84 to .98 with an average reliability of .92.

Generally speaking, the author and his colleagues have found that SVT listening comprehension tests have lower reliabilities that reading tests. This is unlikely to be an attribute of the tests themselves since some studies have used one version of a test as a listening test with one group and as a reading test with another group, and have still found

slightly lower reliabilities when the tests are listened to. One probable reason for the lower reliabilities has to do with the fact that almost all of our research examining reliability has involved group administration of both the listening and reading tests. When young students take a group administered listening test they sometimes signal to one another what they believe the answers to be. This undoubtedly reduces the reliability of the tests. It also happens far less frequently with reading tests because students have to attend to the printed page and differential reading speeds quickly spreads examinees to different parts of the tests. Another reason for the lower reliabilities of listening tests has to do with the nature of the test administration process. An examinee has one opportunity to understand the text presented aurally, but when they read they can return to material they do not understand. Thus, the reading index is likely to be closer to an examinees "true score" than is the listening index.

It should be noted that the reliabilities in the studies described above were obtained with tests that had not gone through a try-out and revision process. There were unquestionably bad test sentences in the tests that could have been identified and revised if the goal of the research was to develop commercial tests. The identification and revision of these bad items would presumably increase the reliability of the tests.

On two occasions the author has been contacted by researchers who were concerned that the SVT tests they had developed were not behaving as a comprehension test should and they wanted his advice on how they might be improved. On one of the occasions the caller had developed a test consisting of a single passage and a 12 sentence test; on the other occasion, the caller had developed a two passage test consisting of 32 test questions. On neither occasion had the caller calculated the reliability of the test, but a

safe assumption would be that in both cases the reliabilities of the tests were very low. The important point is that SVT tests should be long enough to provide reliable assessment, particularly if comparisons between individuals are to be made.

Base Level Validity and Uses for SVT Tests

The initial research examined in this section reviews evidence that SVT tests have "base-level" validity, and the section to follow (Uses for SVT tests) discusses research relevant to the issue of whether SVT tests are valid when used for a particular practical purpose. The seven "base-level" validity questions addressed in this section are: 1) Do readers who differ in reading skill perform differently when they read the same passage? 2) Do readers perform differently on SVT tests based on text that varies in difficulty? 3) Do SVT tests measure passage comprehension, rather than sentence comprehension? 4) Does performance on SVT tests improve as a function of instruction? 5) Does SVT performance vary in accordance with working memory capacity? 6) Does SVT listening and reading comprehension performance covary in a sensible manner? and 7) Do SVT tests have appropriate patterns of relationships with other tests? This last issue addresses the question of whether SVT tests have convergent and divergent validity.

The Sensitivity of SVT Tests to Differences in Reading Skill

Any technique that purports to measure reading comprehension should be sensitive to differences in reading skill. Sensitivity to differences in reading skill has been examined in two ways in SVT research. First, examinees who vary in age, and presumably in reading skill, have taken the same SVT tests and the research examines whether older readers perform better than younger readers. Second, examinees who vary in reading skill as indexed by an outside criterion have taken the same SVT tests and the research examined

whether SVT performance varied in accordance with outside indices of reading skill. The outside indices used in this research were teacher evaluations of reading competence, performance on standardized reading tests, and varying subject matter expertise in situations where the SVT tests were based on text drawn from a subject matter area (e.g., having novices and experts in physics take SVT tests based on excerpts from a college level physics text). Research using both methods of evaluating reading skill provided evidence indicating that students having superior reading skills perform better on SVT tests that did students with lesser skills (Greene, et al., 1990; Rasool & Royer, 1986; Royer & Carlo, 1991a; Royer, Carlo, Carlisle and Furman, 1991; Royer, Carlo, Dufrense & Mestre, 1996; Royer, et al, 1979; Royer, et al, 1986; Royer, et al, 1984; Royer, Sinatra & Schumer, 1990).

SVT Tests are Sensitive to Differences in Text Difficulty

A second quality that language comprehension tests should have is that they should be to be sensitive to variation in text difficulty. Whereas it may seem obvious that comprehension tests should have this property, not all do, as shown by Drum, Calfee and Cook's (1981) study that examined the extent to which text characteristics (including readability indices) predicted the p value (the proportion of examinees getting an item correct) of test items contained in 18 standardized reading comprehension tests. They found that only 12% of the variance in p values was associated with characteristics of the passages that the items were based on. In contrast, 50% of the variance was associated with properties of the test questions themselves. Taken literally, this result suggests that performance on individual test items was far more dependent on an examinee's ability to read and understand the question than it was on an examinee's ability to read and

understand the passage.

The sensitivity of SVT tests to text difficulty has been assessed in a variety of ways. The first way entailed basing SVT tests on passages drawn from texts used in different grades (e.g., reading texts used in grades 4, 6, and 8) using the assumption that text difficulty would increase with advancing grade. Research using this method showed that SVT performance varied systematically with the grade level of the text (Greene, et al., 1990; Royer et al., 1979; Royer & Carlo, 1991a; Royer, et al., 1986).

A second method of assessing sensitivity to text difficulty involved basing SVT tests on passages that had been written to have different readability levels. For instance, using the Dale-Chall (1948) readability formula, Royer and Hambleton (1983) wrote 50 passages that had readabilities ranging from grade 3 to grade 7. These passages were incorporated into SVT tests that were then administered to approximately 1100 students in grade 3 to grade 7 in a manner such that students at varying grades read passages of varying difficulty. This research, and other research using the same method (Royer, et al, 1986), showed that SVT test scores varied with text readability.

A third demonstration that SVT tests are sensitive to text difficulty involved the use of a formal text analysis method developed by Walter Kintsch and his associates (e.g., Kintsch & Vipond, 1977) and documented by Turner and Greene (1977). Kintsch and Vipond (1977) had shown that text analysis characteristics such as propositional density, the redundancy of arguments, the extent of intersentential connections, the number of levels associated with the graphical representation of a text structure, and the ordinal position a sentence occupied in a text passage were much better predictors of passage comprehension than were readability formulas. Royer et al., (1984) and Lynch (1984)

demonstrated that students scored better on SVT tests based on passages with good Kintschian text characteristics than they did on tests based on passages with poor Kintschian text characteristics.

SVT Tests Measure Passage Comprehension

One universally accepted assumption about the nature of prose is that coherent text is more than a concatenation of sentences. Each sentence in a coherent text contributes to a "theme" that emerges as the reader proceeds through the text. This emergent meaning eases the processing of subsequent sentences and ties together text ideas that are separated physically. SVT testing occurs at the level of the individual sentence. This raises the question of whether SVT tests measure the comprehension of individual sentences, or the comprehension of coherent text.

This issue has been shown to be very important. Research conducted in the early 1980s (Kibby, 1980; Shanahan & Kamil, 1982; Shanahan & Kamil, 1983; Shanahan, Kamil & Tobin, 1982) demonstrated that cloze tests often measured sentence comprehension rather than passage comprehension. The method for demonstrating this was to scramble passages so that there was no coherent connection between sentences and to then develop cloze tests based on the scrambled passages and on the original (coherent) passages. The research cited above demonstrated that there were often no differences in performance on the cloze tests based on the coherent passage and the cloze tests based on the scrambled passage. The inescapable conclusion from this research was that the tests were measuring sentence comprehension rather than passage comprehension, thereby discrediting the tests as a measure of reading comprehension.[1]

Royer et al., (1984, Experiment 4) evaluated the possibility that SVT tests were only

measuring sentence comprehension by selecting sets of four sentences from three different passages and then scrambling all of the sentences in a manner that resulted in the construction of a 12 sentence list of unrelated sentences. An SVT test was then constructed from these scrambled passages and participants then read the scrambled passage and took the SVT test based on the passage. Another group of participants took an SVT test consisting of coherent passages that contained the sentences making up the scrambled passages. After both groups completed their tests, the performance on sentences contained in scrambled passages was compared to performance on the same sentences presented in the context of a coherent passage. The results indicated that performance on the sentences contained in the coherent passages was significantly better than performance on the same sentences presented in scrambled order. These results suggest that SVT tests measure passage comprehension rather than sentence comprehension.

SVT Tests are Sensitive to Relevant Instruction

One of the concerns that led to the development of the SVT procedure was the possibility that standardized reading comprehension tests were not sensitive to reading instruction, and hence failed to pick up reading comprehension gains in that could occur as the result of reading comprehension instruction (Royer & Cunningham, 1981). This possibility was reinforced by several types of evidence available at the time the SVT procedure was developed that indicated that standardized reading comprehension tests measured general cognitive abilities rather than reading ability. If this were true, it would make it unlikely that the standardized tests would be sensitive to gains in reading competence resulting from reading instruction.

One item of evidence suggesting that standardized reading comprehension tests measure general cognitive ability, rather than reading performance per se, is the strong relationship between IQ test performance--which supposedly is relatively impervious to instruction--and performance on standardized reading comprehension tests (see, for example, data presented in the next section of this article).

A second relevant item of evidence regarding the issue of whether standardized tests measure general cognitive ability or reading ability is provided by studies showing that students do not have to read the passages contained in standardized reading comprehension tests in order to correctly answer the questions. For example, Tuinman and Gray (1972) had students read and take tests on passages that were either complete, missing 30% of the words, or missing 50% of the words. They found that performance in all of the treatments was above chance, and that the 30% group scored only 13% below the group reading the original text and the 50% group scored only 23% lower than the complete passage group. In an even more striking study, Tuinman (1973-1974) had 600 students take five major reading comprehension tests either with or without the passages. He found that in all cases performance was well above chance when students answered the questions without reading the passages. Since ability to correctly guess the answer to questions without reading the passage is likely to be related to general cognitive abilities such as problem solving and extent of prior knowledge, these results raise the question of whether the tests would be sensitive to gains in reading competence. Several years after the SVT procedure had been developed, Johnston (1984) provided an illustration of how extent of prior knowledge can bias standardized reading comprehension test performance in a manner that disguises an examinee's true level of reading competence.

Concerns about sensitivity to relevant instruction led to one study that evaluated this property of SVT tests (Royer, et al., 1984, Experiment 2). The study involved college undergraduates enrolled in a psychology course that devoted a considerable amount of time to reading and learning to interpret psychology journal articles. These students took SVT tests at the beginning and end of the semester that were based on material drawn from journal articles and material drawn from the New York Times Sunday Book Review. The results of the study indicated that the students made significant gains in SVT performance on the journal article test, but did not improve on the non-psychology test. This result is consistent with the interpretation that SVT tests are sensitive to relevant instruction in that the participants were instructed in a manner designed to enhance their ability to understand psychology journal articles, whereas they were not instructed in a manner that would lead to an enhancement of their understanding of book reviews.

A second study that seems to demonstrate that SVT performance was sensitive to relevant instruction was reported by Royer, et al (1990). They administered listening and reading SVT tests to 151 3rd and 4th grade students on three occasions during a year: in the Spring at the end of the school year, in the Fall at the beginning of the next school year, and again in the Spring at the end of the school year. The results of the study indicated that poorer readers tended to gain steadily in listening comprehension performance during the entire year of the study. However, reading SVT performance for the poor readers actually declined slightly over the summer and then improved considerably during the school year. The authors interpreted this pattern as indicating that listening comprehension skills were in steady use during the entire year, and therefore listening SVT performance showed steady gain, whereas reading activities for the poor readers dropped off during the

summer but increased when school was in session. Correspondingly, SVT reading performance declined during the summer and improved during the school year.

SVT Performance Covaries with Working Memory Capacity

A number of researchers have provided empirical and theoretical support for the hypothesis that working memory capacity is related to reading comprehension (e.g., Baddeley, Logie, Nimmo-Smith & Brereton, 1985; Daneman & Carpenter, 1980; Kintsch & van Dijk, 1978). In general, this hypothesis is based on the idea that an important aspect of text comprehension is the ability to retain a sufficient number of linguistic units (e.g., ideas, propositions) in working memory to allow the interpretation of a meaningful segment. Readers with an inadequate working memory capacity are thought to have comprehension difficulties because they cannot accumulate the number of linguistic units in working memory necessary to constitute a meaningful segment.

Lynch (1986) provided evidence that performance on SVT tests is related to working memory capacity. He developed a verbal working memory task similar to the one used by Perfetti and Goldman (1976) and administered that task and SVT tests to 57 fourth grade students. He reported a significant correlation between working memory capacity and SVT performance ($r = .59$, $p < .001$). Moreover, when he divided the students into poor, average and good working memory capacity groups, he found very large differences in SVT performance between the poor (proportion correct = .61) and good (proportion correct = .80) working memory capacity groups. Lynch (1987) replicated this result in a study that entailed assessing the working memory capacity of 79 fifth grade students and then having them take SVT tests after they had listened to passages, after they had read passages silently, or after they read passages under a round-robin reading condition. Collapsing over

the different groups, Lynch found a significant correlation between working memory capacity and SVT performance ($r = .67$, $p < .001$).

SVT Reading and Listening Comprehension Performance Covary in a Sensible Manner

Most theories of reading comprehension assume either explicitly or implicitly that reading comprehension and listening comprehension are highly interrelated processes (e.g., Danks, 1980, Horowitz & Samuels, 1987; Klienman & Schallert, 1978). A strong version of the theory relating reading and listening comprehension suggests that they are the same process with the exception of the word decoding component involved in reading (e.g., Carroll, 1977). A somewhat weaker version suggests that the variety of available cues to assist the comprehension process is far greater in listening than in reading, but that the underlying psychological processes are nonetheless heavily interrelated (e.g., Kleinman & Schallert, 1978).

Empirical data also support the relationship between reading and listening comprehension. Sticht, Beck, Hauke, Kleinman, & James (1974) reviewed 31 studies that compared listening to reading comprehension. These studies indicated that the strength of the relationship between listening and reading comprehension increased steadily with age. The authors concluded that variability in reading comprehension at an early age was largely dependent on variability in decoding skills plus variability in prior knowledge, intelligence, and other general cognitive factors that could influence listening comprehension. As students aged, and the variability between students in word decoding skills lessened, the influence of factors that affected both listening and reading comprehension performance became increasingly important, thereby increasing the relationship between listening and reading comprehension.

Sinatra (1990) has also provided evidence suggesting that reading and listening are the same process once a printed word has been identified. She had subjects listen to linguistic messages and then judge whether pairs of visually presented verbal strings (presented on a computer screen) were identical or different. The messages her subjects heard and read consisted of complete sentences, non-grammatical sentences, random strings of words, and strings of pronounceable non-words. She was able to show that her participants were significantly faster at identifying the visual strings when they heard sentences, ungrammatical strings, and random words, but not when they heard pronounceable non-words. She interpreted these results as supporting the position that the processing system for both oral and written language is identical once a printed word has been decoded.

The views described above suggest that listening comprehension competence develops before reading comprehension and places a "ceiling" on reading comprehension performance. The developmental (listening develops before reading) and ceiling hypotheses are supported by theory and are consistent with correlational evidence, but are very hard to test directly because of the difficulty in arranging an experiment in which presentation conditions for listening and reading tasks are equated while at the same time assuring an aspect of "normalcy" associated with the tasks. For instance, in a normal reading task, readers frequently go back and re-inspect text while reading. It is much more difficult to arrange a listening test where students could control the input of the listening materials.

It should be possible to roughly equate the conditions for listening and reading comprehension by presenting participants with a text aurally presented at a certain rate,

and then presenting the same text at the same rate, one sentence at a time on a computer screen. The problem with this procedure is that the presentation of text one-sentence-at-a-time on a computer screen at a specific rate is not a normal reading condition.

The issues described above provide a backdrop for an attempt to determine if SVT listening and reading comprehension tests behave in a manner consistent with theory and with previous research. Royer et al (1986), recognizing that it would be very difficult to directly equate conditions of presentation for listening and reading tests, decided to test predictions about how listening and reading performance would interact with other variables in a study. They had 77 fourth and 89 sixth grade students take listening and reading SVT tests that were based on text material that varied in difficulty. The SVT tests contained six passages that had been developed for use in the Royer and Hambleton (1983) study. Two of the passages had Dale-Chall (1948) grade level readability indices at the third grade level (called easy passages), two at the fifth grade level (called moderate passages), and two at the seventh grade level (called difficult passages). The six passages were divided into two sets of three passages with each set containing an easy, moderate and difficult passage. One of the sets was then converted into a listening SVT test and the other into a reading SVT test. After completing the first set of tests, another set was created through a counterbalancing process in which the passages that were listening tests in the first set became reading tests, and the passages that were reading tests in the first set became listening tests.

The predictions tested in the study were based on several assumptions. The first assumption was that 4th and 6th grade students were still developing their reading skills, thereby creating a situation where variability in decoding performance could contribute to

the relationship between listening and reading comprehension. However, the extent to which decoding skills contributed to the relationship between listening and reading comprehension should be dependent on the difficulty level of the text and the grade level of the student. Fourth grade students, and especially sixth grade students, should show less intragroup variability in the decoding of easy text than in the decoding of the moderate and difficult text. This reasoning lead to the prediction that there would be an interaction between difficulty of the material and listening and reading SVT performance.

The specific nature of this interaction should be that with easy material both listening and reading SVT performance should be parallel and high. As the readability of the passages becomes more difficult, there should be a point where reading performance drops more precipitously than listening performance. This prediction was based on the assumption that the developing reading ability of students lags behind their listening ability with difficult text.

A corollary prediction in the study involved student grade level. The logic for this prediction was that if reading ability declines sooner than listening performance as passage difficulty increases, it should be the case that fourth-grade students reach the point of decline sooner than sixth grade students. This hypothesis was based on the assumption that sixth-grade students have superior listening and reading comprehension abilities compared to fourth-grade students, and this should result in their reaching the point of decline later than fourth-grade students. The data from the Royer et al. (1986) study supported the predictions. Both the predicted interaction between SVT test modality and passage difficulty and the predicted interaction between SVT test modality, passage difficulty and grade level were significant sources of variance.

In addition to providing support for the predictions, there were several other interesting aspects of the data from the Royer et al. (1986) study. First, performance on the reading SVT test was higher than performance on the listening SVT test with easier material (grade 3 passages for the 4th grade students and grade 3 and 5 passages for the sixth grade students), but listening performance exceeds reading performance with more difficult material. As noted earlier, reading performance should not exceed listening performance if listening comprehension places a ceiling on reading comprehension. However, as the authors of the study noted, the test administration situation favors the reading comprehension modality, if the examinee is able to read the passage with understanding. That is, students taking a reading SVT test can reread a sentence that had not been understood, and even read a passage multiple times if they want to. In contrast, a listening SVT test (typically) is tape-recorded and the listener has no control over passage presentation.

A second interesting aspect of the data from the Royer et al., (1986) study was the tendency for the cross-over point between listening and reading comprehension to occur at about the student's grade level: at grade 4 readability for the fourth grade students and at grade 6 readability for the 6th grade students. That is, for grade 4 students, reading comprehension was better that listening comprehension with easy (grade 3) material, but listening performance was slightly better than reading performance on the grade 5 material, and quite a bit better than reading performance on the grade 7 material. For the sixth grade students, reading was better than listening on grade 3 material, slightly better than listening on grade 5 material, and listening was superior to reading on grade 7 material. In other words, the cross over point (where listening became better than reading) occurred

between grade 3 and 5 material for the fourth grade students and between grades 5 and 7 for the sixth grade students. A discussion of the practical implications of this finding will occur in the section of the article concerned with diagnostic uses for SVT tests.

The Royer et al., (1986) study indicated that SVT reading comprehension performance on easy material exceeded SVT listening comprehension on the same material, but that as difficulty of the material increased, listening performance exceeded reading performance. This pattern of results suggested the possibility that if text difficulty was matched with average reading ability, patterns of listening and reading comprehension performance might change as a function of the reading skill of individual examinees. This hypothesis, tested by Royer, et al. (1990), suggested specifically that good readers might be able to take advantage of their reading skills in a manner that resulted in their reading performance being superior to their listening performance. In contrast, poor readers would be challenged by the text and their listening performance should exceed their reading performance. A study conducted with 75 third and fourth grade students who had been sorted into good and poor reading categories on the basis of their performance on the California Test of Basic Skills confirmed this expectation. Again, the implications of this finding for reading diagnostics will be discussed in a later section of the article.

Convergent and Divergent Validity

If SVT tests are measuring comprehension there should be at least a moderate positive relationship between SVT reading performance and performance on other reading comprehension tests. Moreover, reading SVT performance should relate more positively to reading comprehension performance on standardized tests than would listening SVT performance. The logic for this expectation is that two tests that measure an aspect of

reading comprehension should relate more positively than a test that measures reading comprehension and one that measures listening comprehension. Evidence relevant to these expectations is presented in Table 2.

Insert Table 2 About Here

As can be seen in Table 2, the correlations between reading SVT performance and reading comprehension performance on standardized tests are in the moderate to moderately high range. In addition, in the single case where listening comprehension was also assessed, it related less positively to reading comprehension performance on standardized tests that did reading SVT performance. The correlation between reading and listening SVT performance in the study where both were collected was .59.

The modest correlation between SVT performance and performance on standardized tests of reading comprehension would be expected if the tests were measuring somewhat different things. Specifically, it was suggested earlier in the article that standardized tests may measure general cognitive attributes that contribute to reading comprehension whereas SVT tests are designed to measure the comprehension of a specific text. If one test is measuring a general characteristic whereas the other is measuring comprehension of a specific text, a high correlation between performance on the two types of tests would not be expected. The issue of general versus specific measures of performance will be returned to later in the article in the context of a presentation of evidence showing that SVT tests can be used to predict future learning performance.

In addition to SVT performance being positively related to reading comprehension

performance on standardized tests, SVT performance should also be related to performance on other measures that require the student to read and understand coherent text. Examples of such measures include the science and social studies section of the Iowa Test of Basic Skills. In addition, since vocabulary knowledge is presumably an important component of reading competence, there should be a positive relationship between SVT performance and standardized measures of reading vocabulary. Correlations between SVT performance and these measures are presented in Table 3. For comparison purposes, Table 3 also contains the correlations between the other measures and scores from the reading comprehension component of the standardized tests.

Insert Table 3 About Here

Table 3 indicates there is a pattern of positive relationships between SVT performance and other measures requiring reading, and between SVT performance and standardized measures of reading vocabulary. The relationships were generally in the moderate to moderately high range. However, the correlations involving SVT were lower than those involving the standardized reading comprehension measure. One interpretation for this is that there is a certain amount of "common task" variability (i.e., answering multiple-choice questions) shared by all measures on standardized tests. The SVT does not utilize the same task and thus has somewhat lower relationships with the other measures than the standardized reading comprehension measures.

Another item of data that is relevant to the convergent validity of SVT tests is the relationship between teacher judgments of reading competence and SVT performance. On

a number of occasions we have asked teachers to rate the reading and listening comprehension ability of the students in their class. In the Royer, et al. (1979, Experiment 1) study, for instance, teachers were asked to rank order the students in their class in terms of judged reading comprehension ability. The correlation between teacher judgments and reading SVT performance was .55.

Royer & Carlo (1991a) used a slightly different technique to collect teacher judgments of the reading competence of students enrolled in bilingual education classes. They asked teachers to select the student with the best reading comprehension performance in the class and assign that student a score of 9. They were then instructed to pick the student who had the poorest reading skills in the class and assign that student a score of 1. They were then instructed to give the other students in the class scores ranging from 2 to 8 in a manner that resulted in a relatively normal and symmetrical distribution of scores. These teacher judgment scores were determined subsequently to be significantly related to SVT performance. Greene, et al., (1990) used the same technique with a very different population (elementary students in Grenada) and found essentially the same result.

A final piece of data that is relevant to the convergent validity of SVT tests was collected in the Royer, et al., (1984, Experiment 2) study. This study entailed asking college students to free-recall a passage they had just read before taking a reading SVT test on the passage. The correlation between free-recall performance and SVT performance was significant ($r = .43$, $p < .01$).

The previous paragraphs in this section have documented the fact that SVT test performance is related positively to performance on other tests or measures that also

purportedly involve reading comprehension ability. Evidence will now be considered that indicates that SVT performance is not strongly related to measures that it should not be related to (divergent validity).

Table 4 shows the relationship between SVT performance and a number of other measures that are not strongly dependent on reading skill. Again, for the purpose of comparison, the table also contains correlations between the other measures and standardized reading comprehension test performance. Performance on the mathematics sections of standardized tests provide an example of a measure that should not be strongly correlated with reading comprehension ability. Actually, there are generally two measures of mathematics competence provided on standardized tests, one of which requires more reading skill than the other. The mathematical concepts measure on standardized tests typically requires the examinee to read a question probing his or her mathematical knowledge. In comparison, mathematical computation questions require no reading at all. Notice that the correlations reported in Table 4 range from low to moderately low, and that the relationship between the mathematical concepts measure and reading comprehension performance as indexed by both the standardized and SVT tests is higher than the relationship between reading comprehension and the mathematical computation measure. Finally, notice that in every case, save one, the relationship between SVT performance and the other measure is lower than that between standardized reading comprehension performance and the other measure. This pattern is consistent with the "common task" interpretation mentioned earlier, but it is also consistent with the hypothesis that standardized reading comprehension tests measure general cognitive attributes rather than abilities specifically related to reading performance.

Insert Table 4 About Here

The final set of relationships to be reported involve IQ and reading comprehension performance. Earlier in the paper a concern was raised that standardized reading comprehension tests were measuring much the same properties as verbal IQ tests, and to the extent this was true, it could decrease the sensitivity of standardized tests to reading skill gains caused by instructional exposure. Table 5 presents the results of studies that involved the collection of IQ, SVT performance, and standardized reading comprehension test performance. Note that the relationships between verbal IQ and SVT performance are less positive than those between standardized reading comprehension performance and verbal IQ.

Insert Table 5 About Here

Uses for SVT Tests

The research reviewed in the preceding sections reviewed evidence for the base-level validity of SVT tests as a measure of language comprehension. The research reviewed in the section to follow reviews evidence suggesting that SVT tests are valid when used for a particular purpose.

Using SVT tests as a measure of comprehension in basic research. A number of studies have used SVT tests as a dependent variable in language research. As an instance, Kardash, Royer, & Greene (1988) conducted a study that took advantage of the

SVT's presumed ability to measure surface comprehension of text. They had college students read Pichert and Anderson's (1977) home buyer/burglar passage under perspective conditions (i.e., read the passage from the perspective of a burglar or a home buyer) given before reading the passage, or at the time of recall. Participants then either free-recalled the passage or took an SVT test based on the passage. The results showed that the impact of the perspective manipulation differed as a function of the type of test the participants took. The authors suggested this indicated that the schema activation effects associated with perspective instructions had an impact on information retrieval (as measured by free-recall tests) but not on information encoding (as measured by SVT tests).

Walczyk (1990) also used the ability of SVT tests to measure the surface comprehension of text (which he called literal comprehension) in a profitable way in a study involving 4th grade children. He was interested in examining the relationship between surface level comprehension, strategic comprehension (as indexed by performance on an error detection task) and low-level reading skill as indexed by measures such as lexical access, and verbal working memory. He found that SVT performance was correlated with the measures of low-level reading skill and that neither low level reading skills nor SVT performance were correlated with error detection ability. Walczyk (1990) interpreted these results as indicating that strategic reading competence is an ability that depends on the satisfactory accomplishment of low level activities, but beyond this level of accomplishment, variation in low level reading skill has no impact on strategic reading performance.

Several studies have utilized SVT tests as a means of examining the impact of prior beliefs on the comprehension of information presented in text form. Kardash and Scholes (1995), for example, collected information concerning the beliefs that college participants

had about how AIDS was transmitted, and then examined the impact of those beliefs on the comprehension of AIDS information presented in text form. Kardash and Scholes (1995) used both SVT tests and free recall as measures of text comprehension and found that prior beliefs about AIDs transmission influenced performance on free recall tests but not on SVT tests. This result is consistent with the Kardash, et al. (1988) study that indicated that information that is actually encoded (as measured by SVT tests) can be distorted during free recall.

Dole, Sinatra, & Reynolds (1991) reported similar SVT results in a study that examined the influence of strong prior beliefs on the comprehension of text concerned with creationist and evolutionary biology views on the origin of species. Using questionnaires, they identified participants who believed in either the creationist or evolutionary biology position, and then asked those participants to read texts espousing one or the other views. Dole et al (1991) found no evidence that prior beliefs influenced performance on SVT tests that measured comprehension of the passages.

A different finding has been reported by Quick and Andre (1992; 1993). Like Kardash & Scholes (1995), they were interested in the influence of prior beliefs on the comprehension of AIDs information. They used a novel and interesting variant of the SVT procedure in a study that tested the hypothesis that prior beliefs about a controversial topic like AIDS served as a "filtering system" that actually could alter the nature of the information that participants acquired from text. They tested this hypothesis in two studies that first involved administering questionnaires about attitudes toward sexually responsible behavior. On the basis of the responses to these questionnaires, participants were classified as being high or low in sexual responsibility, with low sexual responsibility subjects expressing

greater willingness to engage in risky sexual behaviors than high sexual responsibility participants. All subjects then read information about how AIDS is acquired and took a test that they believed to be a measure of how much of the text information they had understood.

The comprehension test the participants took was a variant of an SVT test that was designed to be sensitive to the extent to which previous attitudes influenced text comprehension. The test consisted of original and paraphrase SVT items and meaning change items that had been biased to pick up the influence of sexual attitudes on what the participants understood the passages to say. Some of the meaning change items were biased in a manner to be more consistent with the belief structure of someone with low sexual responsibility, and some were biased to be consistent with the belief structure of someone with high sexual responsibility. The studies found that false alarm rates on the meaning change items (calling the item a "yes" item, indicating that the participant thought the sentence meant the same as a passage sentence) varied in accordance with attitudes about sexual responsibility, thereby providing support for the hypothesis that the acquisition of the meaning of informative text can be distorted by the belief system of the reader.

On the surface, there is a contradiction between the Kardash & Scholes (1995) and Dole et al. (1991) studies and the Quick and Andre (1992; 1993) studies. Kardash and Scholes and Dole et al. found no evidence that prior beliefs influenced SVT performance (but Kardash & Scholes did find that prior beliefs influenced free recall), whereas Quick and Andre did find evidence suggesting that the prior beliefs influenced the nature of the information participants acquired from text. One likely explanation for the differences in the studies is that Quick & Andre's biased item technique is likely to be more sensitive to the

influences of prior beliefs than traditional SVT techniques. The degree to which one of their biased meaning change items differed from an original sentence was relatively subtle and changes as small as the ones exhibited in the items would probably be difficult to detect under most circumstances. The interesting aspect of Quick & Andre's studies, though, was that the ability to detect meaning changes was a joint function of prior beliefs and the slant of the biased item. That is, subjects that expressed opinions interpreted as showing low sexual responsibility tended to correctly reject items inconsistent with this attitude and false alarm (say that the sentence meant the same as a passage sentence) to sentences consistent with the attitude. In contrast, the subjects expressing attitudes interpreted as indicating high sexual responsibility tended to correctly reject the items the low sexual responsibility subjects had false alarmed to, but false alarm to the items the low group had correctly rejected. These results suggest that influences of prior beliefs on newly learned information can be detected if a highly sensitive assessment technique is used.

Another study that involved the use of a variant of the SVT procedure has been reported by Clark (1994). Clark was interested in examining the extent to which novice and experienced classroom teachers would differ in their ability to comprehend classroom interactions. Clark (1994) prepared a video tape that depicted classroom interactions and then isolated significant verbal interactions displayed in the tape and constructed SVT tests from those interactions. She used original, paraphrase and meaning change items to construct her tests, but substituted an inferential item type for the normal distractor item. After viewing the tape, Clark's novice and experienced teachers took the modified SVT tests. She found that the experienced teachers displayed overall better performance on her modified SVT tests and they displayed different patterns of responding than did the

novices. Specifically, the novices were more inclined to choose originals and meaning changes to represent what they viewed on the video tapes, whereas the experienced teachers were more inclined to select paraphrases and inferential items. Clark (1994) suggested that these different patterns were associated with the novices' tendency to encode a literal representation of the activities on the video tape, whereas the experienced teachers were more likely to process the interactions in terms of prior experiences.

Using SVT tests as a diagnostic tool. Research examining the diagnostic potential of SVT tests has identified two ways in which the tests might be used. The first involves examining the ratio of SVT listening to reading performance as a diagnostic index, and the second involves relating patterns of performance on SVT item types to diagnostic categories.

The idea that ratios of listening to reading performance would have diagnostic properties is not new. Venezky and Calfee (1970) described a w-o ratio (written-oral) in their model of the reading process and Carroll (1977) elaborated the attractiveness of a W-O comprehension scale as a means of indexing reading competence. The essence of Carroll's W-O scale was the idea that if there was a minimal discrepancy between a student's written and oral comprehension then the student was reading as well as possible, independent of the absolute level of performance. Alternatively, if the student's oral comprehension was considerably higher than his or her written comprehension, there was a need for instructional effort designed to bridge the gap.

For reasons that are undoubtedly related to the difficulty of developing parallel listening and reading comprehension tests, the idea of a W-O index of reading competence has never taken hold. However, there is a recent increase in interest in the idea because

of concerns about the identification of readers with specific learning disabilities. The U.S. Congress passed PL 94-142 in the late 1970s that dictated that students with disabilities were entitled to special individual services. Among the disabilities targeted by the law was learning disabilities, which according to the law, involved students who:

> ....exhibit a disorder in one or more of the basic psychological processes involved in the understanding or using spoken or written language. These may be manifested in writing, spelling, or arithmetic. They include conditions which have been referred to as perceptual handicaps, brain injury, minimal brain dysfunction, dyslexia, developmental aphasia, etc. They do not include learning problems which are due primarily to visual, hearing, or motor handicaps, to mental retardation, emotional disturbance, or to environmental disadvantage. (Federal Register, 1978, p. 42478).

The Federal description of a specific learning disability led to defining the syndrome as a discrepancy between observed and expected achievement, and this, in turn, gave rise to the common practice of identifying learning disabled students through discrepancies between IQ performance and a test that measured specific academic performance (Fletcher, Shaywitz, Shankweiler, Katz, Liberman, Stuebing, Francis, Fowler, & Shaywitz, 1994). This practice has proven to be controversial because of uncertain support in the research literature (Fletcher, Francis, Rourke, Shaywitz & Shaywitz, 1993; Siegel, 1992; Stanovich, 1991; Stanovich & Siegel, 1994) and because of practical issues like varying assumptions about how large the IQ-reading discrepancy should be before a student is considered learning disabled (Reynolds, 1981; 1985; Shepard, 1980).

Researchers interested in learning disabilities have called for alternative ways of defining the syndrome, and one procedure involves defining it in terms of differences

between listening and reading comprehension (Fletcher, et al., 1994; Stanovich, 1991). The idea is essentially the same as proposed by Carroll (1977): students with a specific reading disability would be characterized by normal to above normal listening comprehension and sub-normal reading comprehension. In contrast, "garden variety" poor readers (Gough & Tunmer, 1986) might be characterized by relatively low performance on both listening and reading tests.

One line of research concerned with using ratios of SVT listening and reading performance as diagnostic indicators began with the Royer et al., (1986) study that found that the reading performance of students reading easy text was generally superior to their listening performance on easy text, but that the pattern was reversed when they read and listened to difficult text. This finding led to the hypothesis that good readers might be characterized by the reading superior pattern whereas poor readers would be characterized by listening performance being superior to reading performance. This hypothesis was evaluated by Royer, et al, (1990).

Royer, et al., (1990) formed groups of 3rd and 4th grade students into low, medium, or high reading comprehension groups based on their performance on the California Test of Basic Skills. These students took SVT reading and listening tests that were matched in difficulty on three occasions during a year: at the end of the first school year, at the beginning of the next school year, and at the end of the second school year. The typical pattern for students in the good reading group was that their reading comprehension performance was superior to their listening comprehension performance. In contrast, the pattern for the poor reading group was that their listening performance was superior to their reading performance. These results confirmed the predictions derived from the Royer, et

al., (1986) study.

There was, however, substantial variability within the poor reading group and Royer, et al., (1990) took a detailed look at the pattern of development on the listening and reading tests during the year of the study. Three distinct patterns seemed to be present in the poor reading group. The first pattern was very similar to the one found for good readers. Specifically, these readers showed gains in both listening and reading performance during the entire year of the study and their reading performance was superior to their listening performance. These appeared to be readers who were reading as well as their listening comprehension performance enabled them to read, and who were making satisfactory progress in the attainment of reading skill, even though the absolute level of that skill might be low.

A second pattern of performance involved students who overall had very comparable levels of listening and reading performance, but markedly different patterns of development for the two skills. This group showed steady gain on the listening tests during the year, but a drop in reading performance during the summer followed by a substantial gain in reading performance during the school year. One interpretation of this pattern is that the group consisted of readers who were losing some of their skill during the summer months due to reading inactivity, thereby falling increasingly behind peers who read over the summer.

The final group was characterized by substantially better listening performance than reading performance, slight gains in listening performance over the year, and no gains on reading performance during the year. In short, the reading instruction the students received during the school year had no demonstrable impact on the reading SVT tests. This last group of students appeared to meet the spirit of the definition of a reading

disability in that their oral language performance was superior to their reading performance and they were not receiving benefit from regular reading instruction.

Royer (1996) continued the line of research examining the ratio of listening to reading comprehension on SVT tests. He reports research in which SVT listening and reading comprehension tests are administered to students suspected of having a specific learning disability. The results of these assessments are then used for diagnostic purposes and for the purpose of establishing baseline levels of performance that can then be used to evaluate performance gains attributable to educational interventions.

Carlisle's use of SVT tests for reading diagnosis. Whereas Royer and his colleagues have concentrated on using differences between listening and reading performance to identify students' reading problems, Carlisle (1989a, 1989b; 1990; 1991; Carlisle & Felbinger, 1991) has taken the additional step of using patterns of performance on the different SVT item types as a reading diagnostic. Carlisle examined the patterns of errors made by good readers and poor readers on listening and reading SVT tests. Among other findings she has reported that good comprehenders exhibited essentially the same pattern of errors on both the listening and reading tests. In contrast, poor comprehenders had the same pattern of errors on original and meaning changes across modalities, but made more errors on paraphrase items when they were part of the reading test and more errors on distractor items when they were part of the listening test. Carlisle (1989b) interpreted this finding as a possible indication 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" (p. 40). Thus, poor comprehenders were more likely to suffer a false alarm with sentences that did not appear in the passage but

had the same theme as the passage.

Assessing the language skills of non-native speakers of English. There are substantial numbers of students enrolled in U.S. schools who have a native language other than English and many of these students are enrolled in bilingual education classes where they receive varying amounts of instruction in their native language. School systems must make several important decisions with children who are candidates for enrollment in bilingual education classes (Royer & Carlo, 1991). First, the school system must decide whether the student's competence in English is high enough to warrant placement in a mainstream classroom. If the decision is made that the student's English skills are limited, and that placement in a bilingual education classroom is called for, additional decisions must be made about curriculum placement within the program (e.g., what reading text to start with) and movement through the program.

The most common form of bilingual education in the U.S. is Transitional Bilingual Education (TBE) (Hornberger, 1989). In the TBE program the author is most familiar with, students judged to have virtually no competence in English (Level I students) receive English as a Second Language (ESL) training and virtually all of their remaining instruction in the native language. When the student acquires some competence in English, he or she graduates to Level IIA of the program where instruction in math is provided in English. As English competency increases, the student progresses to Level IIB where formal reading instruction and science instruction are provided in English. In Level IIC, social science instruction is added to the other subjects taught in English. The student then graduates to Level III, where almost all subject matter instruction takes place in English. Finally, when the student is judged to have academic English skills equivalent to those possessed by the

average native English speaker, the student is transferred to the mainstream classroom. The school system sets the goal of completing the process from initial placement to mainstreaming in three years, though Cummins (1984) argues that attainment of academic proficiency in a second language typically takes six years.

Each of the critical decisions involving initial placement, movement of a child through a bilingual program, and graduation of a child from the TBE program could be facilitated using reliable and valid tests of a child's ability to understand spoken and written material in both the native language and in English. One research strand that has been pursued has involved examining the extent to which SVT tests might fulfill this need (Mestre & Royer, 1991; Royer & Carlo, 1991a; Royer & Carlo, 1991b; Royer, Carlo, Carlisle, & Furman, 1991). This research has examined the reliability, validity, and practicality of SVT tests for the purpose of assessing the language skills of students enrolled in bilingual education programs.

Royer and Carlo (1991a) and Royer, et al. (1991) conducted longitudinal studies that involved administering Spanish and English SVT listening and reading tests to Spanish speaking students enrolled in 3rd, 4th, 5th and 6th grade TBE classrooms. The studies examined the validity of the tests by examining three outside indices of language competence: level of placement within the TBE program, teacher judgments of listening and reading competence in both English and Spanish, and the ability to comprehend text of varying difficulty.

Students involved in the study were placed in the levels of the TBE program described earlier in the section (i.e., Level I, Level IIA and so on). If SVT tests were valid measures of language competence it should be the case that performance on the English

SVT tests would vary systematically in accordance with program placement. That is, Level I students should have the lowest performance, Level II the next lowest, and so on. Performance on the Spanish tests should not, however, vary systematically in accordance with TBE level because English competence, not Spanish competence, determined program placement.

Teachers in the studies rated the listening and reading competence of each student in both English and Spanish. The second validity test involved determining the extent to which SVT performance on both the English and the Spanish tests varied in accordance with teacher ratings.

Finally, the teachers were asked to provide the research team with text materials in both English and Spanish that were below the grade level competence of the average student, that were at the grade level competence of the average student, and that were above the grade level competence of the average student. These materials were then converted into listening and reading SVT tests. Evidence consistent with the interpretation that the tests were valid would be present if performance on the tests varied in accordance with the difficulty level of the material.

The results of the studies indicated that the SVT tests were particularly sensitive to program placement and passage difficulty. In addition, reading SVT performance in both English and Spanish varied in accordance with teacher judgments of reading competence. However, listening SVT performance was only consistent with teacher judgments of listening competence in the native language (Spanish). The authors suggested that the insensitivity of the English listening tests to teacher judgments of English listening competence could have been due to the fact that the teachers making the judgments rarely

worked with students on English listening skills (an activity undertaken by the ESL teacher) and therefore had little basis for making the evaluations.

Another demonstration of the utility of SVT tests for tracking progress in bilingual education programs has been reported by Beeman (1993). She administered listening and reading SVT tests and vocabulary tests in both English and Spanish to native Spanish speaking fifth-grade students in the Fall and Spring of a school year. She reported finding general improvement on all tasks during the school year and she found evidence that early competence in Spanish made a significant contribution to subsequent competence in English reading comprehension.

Measuring language skills in countries other than the U.S. The practical features of SVT test development make SVT tests an attractive measurement procedure for countries that lack the monetary and human resources required to develop traditional tests. Several studies have examined the use of SVT tests in developing countries, and Royer and Carlo (1993) provide a review of general issues involving the testing of language skills in cross-cultural settings.

One research activity (Bosch, 1994; Greene, et al. 1990; Royer, Greene & Anzalone, 1994) involved developing reading SVT tests that were used to evaluate the impact of a computer-assisted-instruction (CAI) system that had been donated to the Grenada Ministry of Education by a U.S. computer manufacturer. Two parallel forms of SVT reading tests based on text materials in use in Grenada and suitable for administration to students enrolled in Standards 1, 2, 3, 4 and 5 (roughly equivalent to U.S. grades 3, 4, 5, 6, and 7) were developed by Grenadian teachers who had participated in an SVT test development workshop. The tests were administered to over 1000 children per year at the beginning

and end of the school year starting in 1986/1987 and ending in 1991/1992. The validity of these tests was evaluated by determining if test performance varied in accordance with reading skill of the student as indexed by grade placement, teacher evaluations of student reading competence, and difficulty of the text the tests were based on. Analyses involving these indices provided evidence consistent with the interpretation that the tests were valid measures of reading competence (Greene, et al. 1990).

Another project involved evaluating the impact of interactive radio lessons in reading and math that are being broadcast to Haitian children. The impact of the radio programs in the area of reading was evaluated using SVT tests in Haitian Creole that were developed by Haitian educators who participated in an SVT test development workshop. As was the case in the Grenada research, the validity of the tests was positively evaluated by using passages of varying difficulty and by collecting teacher ratings of reading competence.

Another project (Anzalone & Mathima, 1989) involved listening SVT tests that were based on radio scripts that were to be broadcast over an educational radio network in Nepal. The developers of the radio programs were concerned that the targeted audience might not be able to understand the educational content of the scripts. To evaluate this possibility, researchers in Nepal developed SVT tests based on the scripts and administered those tests to samples of the targeted audience. Poor performance on the tests was taken as evidence that the scripts needed revision.

A fourth example of the use of SVT tests in a foreign country involved a project conducted in Belize (Royer & Greene, 1990). The project had two goals: first to evaluate the basic reading skills (e.g., word recognition) of Belize students using a computer-based assessment procedure, and second, to determine whether reading materials in use in

Belize schools appropriately matched the reading skills of students. This second purpose was accomplished by developing reading SVT tests from samples of text used in the schools and then administering those tests to Belize students. The study found that the average student had a very poor understanding of the texts they were required to read, thereby suggesting that the assigned materials were much too difficult for the average student.

A Czechoslovakian researcher (Zdenka, 1986) has reported another study involving the use of SVT tests in a foreign country. She developed SVT tests to be used in elementary placement decisions and provided evidence of their validity in the form of data showing that test performance varied in accordance with student grade level.

A final example of using SVT tests in a country other than the U.S. occurred in Guatemala (Maria Carlo, personal communication). The context for the use was an adult literacy program designed to teach literacy skills to Guatemalan peasant farmers. Participants in the program completed a literacy course and thereafter periodically received a newspaper designed to provide the participants with practice reading material that would be interesting and informative. A question arose, however as to whether the difficulty of the material in the newspaper exceeded the developing reading capacity of the peasants. The researchers addressed the issue using both listening and reading SVT tests based on articles appearing in the newspaper. SVT reading tests were used to evaluate the extent to which the articles could be read and understood, and the listening tests were used to evaluate the possibility that the articles could not be understood even when they were listened to. If listening comprehension were poor, it would suggest that the participants did not have the prior knowledge necessary to understand the material they were listening to.

The results of the study indicated that the Guatemalan peasants typically did not have levels of reading skills that allowed them to understand the passages. Moreover, the researchers found that the peasants often did not have the background knowledge that would enable them to understand the articles even when they were listened to.

The Guatemalan study also provided evidence of the hazards of conducting research in cultural settings very different from the researcher's cultural background. Maria Carlo (personal communication) reported that the Guatemalan peasants had a reverence for the printed word that translated into the following proposition: "if it is written down, it is surely true." She believed that this reverence came from the fact that the most common source of printed material available to the peasants was the Bible. The belief that all printed material was true led to considerable dissonance when the peasants were asked to indicate whether a written test sentence matched a passage sentence in meaning. Many of the peasants seemed very reluctant to respond "no" to a sentence even when there were obvious indications that they recognized that a test sentence did not match a passage sentence in meaning.

Measuring the comprehensibility of text. There are many situations in which one would like to know if text is comprehensible to a targeted population. One approach to dealing with this issue is to calculate readability indices for the text and if the readability level matches the presumed reading skill level of the targeted population, then the text is assumed to be comprehensible. This procedure has been heavily criticized, however, because readability indices describe characteristics of text rather than describing whether a reader can understand a text (e.g., Davidson & Green, 1988). SVT tests have been used to circumvent this difficulty in that they can be used to measure how well a text can be

comprehended by an individual reader or by a group of readers.

Examples of this type of use were described in the previous section where SVT tests were used to measure the comprehensibility of radio scripts in Nepal and reading texts in Belize. Another example of this type of use has been reported by Ramos and Bayona (1991) who were interested in developing medical education materials intended for use in a low-income (largely Spanish-speaking) community in Rochester, New York. The authors developed SVT tests based on Spanish and English educational materials currently in use in the community. The intent of the study was to evaluate the comprehensibility of the materials, and if they proved difficult to comprehend, to revise them so they would be more comprehensible.

Predicting future learning. The studies examining whether SVT performance could be used to predict future learning performance had their origin in the realization that theories of verbal learning and theories of comprehension had, by the early 1980s, become nearly indistinguishable (e.g., Brown, Bransford, Ferrara & Campione, 1983; diSibio, 1982; Jenkins, 1979). If learning from text and the comprehension of text are very similar cognitive processes, it should be possible to predict learning performance at time 2 from comprehension performance at time 1.

Royer, Abranovic and Sinatra (1987) described two ways in which reading comprehension could predict future learning performance. In the first, the comprehension test could be sensitive to cognitive abilities that are relevant to comprehending a wide variety of text material. Examples of abilities of this type include general knowledge, inferential reasoning ability, working memory capacity, and so on. A test that was sensitive to variation in cognitive abilities of this type might be thought of as a general test of

comprehension in that performance on the test should be representative of performance on texts covering a broad range of subject matter. A reasonable hypothesis would be that a comprehension test tapping general cognitive abilities should be able to predict general learning performance. This is hardly a novel idea in that many tests that are used to predict general learning performance include comprehension tests. For instance, both the Scholastic Achievement Test and the Graduate Record Examination contain reading comprehension sections.

In addition to identifying a general route whereby comprehension could predict learning, Royer, et al. (1987) identified a second, more specific prediction route. The second prediction route would involve a situation where the comprehension test was sensitive to specific knowledge that was relevant to the comprehension of a narrow range of text and to the learning of particular subject matter material. For example, a physician should be able to comprehend the content of a medical journal in his or her specialty area better than an historian even though the two did not differ in their general reading capabilities. Likewise, the physician should be able to learn the contents of that journal with greater ease and efficiency than would the historian, even if the two did not differ in general learning ability. This suggests the possibility that a comprehension test that was sensitive to the specific knowledge required to understand the medical text would only predict the learning of medical material.

The Royer, et al. (1987) study designed to assess these possibilities had two purposes. First to test the general hypothesis that SVT performance could predict learning performance, and second, if SVT performance did predict learning performance, to identify whether the prediction was derived from the general or the specific route. The study

involved administering reading SVT tests based on excerpts from a textbook used in a psychology course and a textbook used in a business statistics course to 184 college students enrolled in one or the other course. The SVT tests were administered the first week of the semester and performance on the tests was subsequently correlated with indices of overall course performance obtained at the end of the course. The study also obtained each student's overall GPA (grade point average) and SAT scores from central administration records.

The results of the study showed that relevant SVT tests were significant predictors of course performance, but irrelevant SVT tests were not. That is, the psychology SVT tests predicted psychology course performance and the business statistics SVT tests predicted performance in the business statistics course, but the psychology test did not predict business statistic performance and the business statistic test did not predict psychology performance. This pattern of results is consistent with the interpretation that SVT tests can be used to predict future learning performance and that the source of this prediction is specific subject matter knowledge that facilitates both the comprehension of subject matter text and the learning of subject matter knowledge.

Further evidence that the predictive power of the SVT tests was mediated through a specific subject matter route came from analyses including the GPA and SAT measures. These analyses showed that SAT scores (presumably measures of general learning ability) were significant predictors of both course performance and GPA (a measure of general learning performance). In contrast, SVT performance predicted course performance, but not GPA. A final set of analyses showed relevant SVT performance and SAT performance accounted for about equal amounts of variance in separate analyses where course

performance was the criterion variable. When relevant SVT performance and SAT scores were both included as predictor variables in an analysis where course performance was the criterion variable, the amount of variance accounted for was about 60% greater than when the two variables were separate predictor variables. Again, these results indicate that the SVT tests used in the study obtained their predictive power by tapping into specific subject matter knowledge.

The Royer et al., (1987) study showed that knowledge specific SVT tests could be used to predict specific learning performance, but a remaining interesting question is whether SVT tests could be constructed that predicted general learning performance. Royer, Marchant, Sinatra & Lovejoy (1990) examined this issue in a study that involved administering the SVT tests used in the Royer et. al., (1984) studies to college students enrolled in an introductory psychology course. The SVT tests were based on New York Times Sunday Book Review articles and on abstracts from psychology journal articles. Both of these text sources were viewed as being challenging in verbal and conceptual content, especially for the introductory psychology students in the study who were generally first or second semester freshman. The SVT tests were administered at the beginning of the semester, and at the end of the semester indices of course performance were obtained. In addition, student SAT scores and student GPAs over a three year period were obtained from central administration records.

Analyses of this data indicated that both the book review and the journal article SVTs were highly significant predictors of course performance and GPA. The finding that SVT performance predicted GPA differed from the Royer et al., (1987) study where SVT performance did not predict GPA. Analyses also indicated that SAT scores were a

significant predictor of both course performance and GPA, but unlike the Royer et al., (1987) study, when SAT performance and SVT performance were included as predictors in the same regression equation, the combined predictors accounted for no additional variance above what each accounted for independently.

Royer et al., (1990) then went on to replicate the Royer et al., (1987) finding that SVT tests could be used to predict specific learning performance. They constructed SVT tests from textbooks to be used in an environmental biology course and an introductory psychology course and administered both types of tests to students enrolled in both courses. As was the case in the Royer et al., (1987) study, the relevant SVT tests (psychology SVT in the psychology course and biology SVT in the biology course) were significant predictors of course performance but the irrelevant SVT tests were not significant predictors of course performance. Again, consistent with the interpretation that the SVTs were specific predictors of learning, they did not predict overall GPA, but when combined with SAT scores in the same prediction equation, more variance in course performance was accounted for than when SVT and SAT scores were entered as predictors in separate prediction equations.

The results from the Royer et al., (1987) and the Royer et al., (1990) studies suggest that SVT tests can be either general or specific predictors of learning performance. If the tests are constructed from general and challenging material, they provide indices of general comprehension performance and can be used to predict general learning performance. If the tests are constructed from materials that are specific to a course of study, they can provide specific indices of ability to comprehend text in that course of study and can be used as subject-specific predictors of learning performance. An interesting possibility

suggested by the analyses combining both SAT scores and specific SVT scores as predictor variables is that SVT tests could possibly serve as both specific and general predictors of learning performance. Analyses in both studies (Royer, et al., 1987; Royer, et al., 1990) indicated that when SAT scores (a general measure) and SVT scores (a specific measure) were entered as predictors in the same equation, more total variance in course performance was accounted for than in analyses where the two variables were separate predictors. Given that Royer et al., (1990) showed that SVT tests could be both general and specific predictors, a particularly potent prediction combination might be to administer both specific and general SVTs to the same population. Future research could, perhaps, examine this possibility.

The prediction research reviewed in this section offers the possibility that SVT tests could serve a variety of useful functions in practical settings. For instance, one implication of the research is that SVT tests based on text material to be studied in a course or training program could be used to select participants for the educational or training program. Purwono (1997) examined this use in a community college setting where open-entry students wanting to be nurses or medical technicians take a biology course that feeds both programs. In the past, course instructors reported bimodal distributions of student performance in the course; some of the students did fine, and others, despite effort from many of them, simply could not seem to learn the material. For some time the college had been looking, without success, for a way to identify the students who were likely to do poorly so they could take a remedial course that would enhance the possibility that they could pass the feeder course. For instance, several years of administering Nelson Denny reading tests yielded a correlation of less than .1 with biology course performance. An SVT

test based on excerpts from the text used in the course proved to be a much better predictor and the college is now using the test as a screening procedure to sort students into either the regular or remedial course.

Another possible use for SVT tests is as an indicator of educational or training success. One goal of most educational and training experiences is to prepare students/trainees to be more accomplished learners. One indicator of this ability would be a student's ability to read (or listen to) and understand material that is similar to the material that the student will encounter once the educational or training experience has ended. SVT tests could be based on material of this type and provide an indication of whether this goal of education and training has been accomplished.

An implication of the prediction research discussed in this section that is noteworthy concerns the issue of what SVT tests are measuring. In the introduction to this article it was mentioned that the SVT procedure was intended to provide a measure of the extent to which a reader could extract the surface meaning of a text. The vast majority of the research reviewed in this article seems consistent with that intention. The finding, however, that SVT tests appear to be able to measure general reading comprehension which in turn can be used to predict general learning performance, would seem to hint that SVT tests can possibly measure more than the extraction of the surface meaning of a text. While there is little evidence supporting this possibility, it remains an interesting possibility that could also be a topic for further research. Later in this article the author will discuss SVT-like procedures that are specifically designed to measure higher level abilities.

Assisting in placement decisions. At the request of a school system in Western Massachusetts, a group of college undergraduates enrolled in an educational research

seminar that the author taught developed a complete battery of SVT tests for grades 1-8 that the school system uses in educational placement decisions. The test development process entailed asking teachers at each grade level to select text material that they believed to be typical of the material they would expect the average students in their grade to be able to read with understanding. Tests were then constructed from passages that "straddled" a grade level. For instance, a 2nd grade test would consist of grade 1 material, grade 2 material and grade 3 material. These tests will be administered to students coming into the system to provide information about where to place the student in the curriculum. Joanne Carlisle (personal communication), formally of Northwestern University, has indicated involvement in a similar project in the Evanston, Illinois area, and Salvia and Hughes in their text (1991) on classroom assessment recommend using SVT tests as one means of providing objective information that could be used in curriculum placement decisions.

Beyond SVT Tests

In a 1972 book chapter Carroll suggested that reading comprehension could be broken into two stages: 1) the apprehension of the linguistic message contained in the text, and 2) relating that information to a broader context. The intent of the author's SVT research has been to develop a measure of the first of Carroll's stages, and the evidence reviewed in this article seems consistent with the interpretation that SVT tests do measure a reader's ability to apprehend the linguistic message contained in a text. There are, however, other important things to know about a reader's interaction with text, and recent research efforts have turned to developing measures of other important aspects of the reading process. These efforts examine processes occurring both below and above the

cognitive level presumably tapped by SVT tests.

When readers perform poorly on an SVT reading test, the examiner does not know the cause of that poor performance. Some additional information can be gained by administering listening SVT tests along with reading tests; an activity that would be helpful to the examiner in determining whether the reader has a specific reading disability or a general linguistic deficit. Knowing that a reader has a specific deficit may not informative, however, with respect to identifying the exact nature of the problem being experienced or in prescribing instructional approaches that might prove beneficial (though the reader should keep Carlisle's research in mind which indicates that SVT tests may provide specific diagnostic information). For this reason, the author and his associates (Cisero, Royer, Marchant & Jackson, 1994; Royer, 1996; Royer & Sinatra, 1994; Sinatra & Royer, 1993) have developed a computer-based diagnostic system (called the Computer-based Academic Assessment System-CAAS) to be used in conjunction with SVT tests. The rationale underlying the development of the CAAS system comes from cognitive developmental theories of reading (e.g., Perfetti, 1992; Stanovich, 1990) and from component processing theories of reading performance (e.g., Carr & Levy, 1990; Royer & Sinatra, 1994). The CAAS system is designed to identify component reading skills that could be blocking reading development and that could be targeted by instruction. Research using the CAAS system has shown that both elementary (Royer, 1996) and college students (Cisero, Royer, Marchant & Jackson, 1994; Cisero, Royer, Marchant & Wint, 1994) with reading problems can be divided into categories based on CAAS profiles. These profiles characterize generally poor readers, readers with a specific learning deficit and readers with general cognitive processing deficits. Each profile has a distinctly different

signature pattern and each identifies specific component skills that could become the target of instructional intervention.

Another line of research has examined the assessment of reading activities occurring at a cognitive level higher than that assessed by SVT tests. Royer, et al (1996), conducted a study that was motivated by Anderson's (1982; 1987) theory of cognitive skill development and by research on expert/novice differences in cognitive functioning (Chi, Glaser & Rees, 1982; Hardiman, Dufrense & Mestre, 1989). An analysis of Anderson's theory suggested that learners in his initial stage of skill development in a subject matter domain should be able to read and understand a text from that domain (i.e., apprehend the linguistic content) but would have difficulty in relating what they had extracted from the text to a broader domain. A reading of the expert/knowledge literature provided indices of what it meant to be able to relate information to a broader domain. In particular, Royer, et al., (1996) focused on the ability to make inferences when working with domain material and the ability to recognize underlying principles embedded in domain problems; both of these capabilities had been found to distinguish experts and novices in a domain (Chi, et al., 1982).

Anderson's theory and the expert/knowledge research led Royer, et al. (1996) to the hypothesis that an early capability that learners develop when beginning study within a domain is the ability to extract the surface meaning from a domain text. Later, as cognitive skill develops, the learner acquires the ability to efficiently relate text information to prior knowledge in a manner that results in the generation of inferences that are not directly stated in the text. Later still, as expertise continues to develop, the learner develops the ability to recognize principles, or "big ideas" that underlie text content. These hypotheses

were evaluated by generating tests in both physics and psychology that were designed to assess the ability to extract the surface meaning of a text (SVT tests), to measure ability to draw inferences from text, and to recognize "big ideas" in text. These tests were then administered to experts and novices in physics and psychology in a study that tested the hypothesis that the smallest difference in performance between experts and novices would occur with the SVT tests, that the largest difference would occur with the "big idea" tests, and that an intermediate difference would occur with the inference tests. The results of the study confirmed these predictions. The study suggests that it is possible to develop practical reading assessment procedures (i.e., suitable for administration in actual instructional settings) that measure a range of cognitive capabilities, and the results were consistent with the interpretation that SVT tests measure a reader's ability to extract the surface meaning of a text.

Concluding Comments

The research reviewed in this article indicates that SVT tests can provide a reliable means of assessing the reading and listening comprehension of a particular text, and that the tests possess seven base-level qualities that tests of comprehension should have. Specifically, the reviewed research indicates that SVT tests are sensitive to variation in reading skill, they are sensitive to variation in the difficulty of text, they measure passage comprehension rather than sentence comprehension, they are sensitive to instruction that should boost comprehension ability, performance on the SVT tests varies as a function of working memory capacity, performance on reading and listening SVT tests varies in a manner consistent with theoretical expectations, and they display good convergent and divergent validity attributes in that performance on SVT tests correlates moderately high

and positive with other tests that also measure comprehension, and they have much smaller relationships with tests that measure attributes other than comprehension.

SVT tests have also been shown to be useful for a variety of purposes. Research has shown that SVT tests can be used as a dependent variable in language comprehension research, as a reading diagnostic tool, as a means of assessing linguistic competence in populations who are non-native speakers of English, as a means of assessing linguistic skills in developing countries, as a procedure for measuring the comprehensibility of text, as a means of predicting future learning performance, and as a tool for assisting curriculum placement decisions. An attractive property of SVT tests for these purposes is that they are cheap and easy to develop.

Having pointed out some of the uses that SVT tests can serve, it is also important to mention some that they are probably ill-suited to serve. Foremost among this category of purposes not well served are uses where one wants a measure of general comprehension aptitude. An example of such a use are aptitude tests such as the SAT or GRE where the intent is to measure comprehension of “text in general.” SVT tests, with their focus on comprehension of a particular text, are probably not appropriate for this purpose. Another example of a use where broad comprehension measurement is called for are large assessment instruments like the NAEP assessment where one wants to make broad generalizations about the reading capability of large numbers of students. These assessments are necessarily based on very general text samples.

The research reviewed in the article certainly does not exhaust the realm of research issues involving SVT testing. There are a variety of purely psychometric questions that could be asked such as whether reliability and validity could be enhanced by changing the

density of SVT item types from a balanced state (equal numbers of each item type) to an unbalanced state, or whether criterion referenced interpretations could be based on signal detection parameters (see section on SVT scoring). Other research questions involve uses for SVT tests. Could, for example, SVT testing provide a means of assessing the comprehensibility of important messages such as medical instructions, academic lectures, or textbooks. Or as another example, could SVT tests be used as a screening procedure for selecting students unprepared for a course of study, or as a means of selecting promising candidates for industrial positions or military training. Future research on these issues could contribute to expanded uses for SVT testing and it could identify areas where SVT tests cannot be used with benefit.

Footnotes

[1]This does not mean that all cloze tests have this problem. For example, Degrees of Reading Power, a cloze based test, appears to have good psychometric properties. The research cited in this section showed that cloze tests that were not carefully constructed could have the limitation of being a sentence measure rather than a passage measure.

References

Anderson, J. R. (1982). Acquisition of cognitive skill. Psychological Review, 89, 369-406.

Anderson, J. R. (1987). Skill acquisition: Compilation of weak-method problem solutions. Psychological Review, 94, 192-210.

Anzalone, S. A., & Mathima, S. S. B. (1989). Final external evaluation: Radio technology training II. (Project RETT II). Arlington, VA: Institute for International Research.

Baddeley, A., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. Journal of Memory and Language, 24, 119-131.

Beeman, M. M. (1993). A longitudinal study of academic language proficiency and crosslinguistic transfer in Hispanic students in a dual language program. Unpublished Ph.D. dissertation, Northwestern University.

Bosch, A. (1994). Computer-assisted instruction in Grenada: High-tech success and sustainability against the odds. Learn Tech Case Study Series No. 3. Washington, DC: Educational Development Center.

Brown, A. L., Bransford, J. D., Ferrara, R. A., & Campione, J. C. (1983). Learning, remembering and understanding. In J. H. Flavell & E. M. Markman (Eds.), Carmichael's manual of child psychology (Vol 1 pp.77-166). New York: John Wiley.

Carlisle, J. F. (1989a). Diagnosing comprehension deficits through listening and reading. Annals of Dyslexia. 39, 159-176.

Carlisle, J. F. (1989b). The use of the Sentence Verification Technique in diagnostic assessment of listening and reading comprehension. Learning Disability Research, 5, 33-44.

Carlisle, J. F. (1990). Diagnostic assessment of listening and reading comprehension. In H. L. Swanson and B. Keogh (Eds.), Learning disabilities: Theoretical and research issues. Hillsdale, NJ: Erlbaum.

Carlisle, J. F. (1991). Language comprehension and text structure. In J. F. Kavanagh (Ed.), The language continuum: From infancy to literacy. Parkington, MD: York Press.

Carlisle, J. F., & Felbinger, L. (1991). Profiles of listening and reading comprehension. Journal of Educational Research, 84, 345-354.

Carroll, J. B. (1972). Defining language comprehension: Some speculations. In J. B. Carroll & R. O. Freedle (Eds.), Language comprehension and the acquisition of knowledge. New York: John Wiley & sons.

Carroll, J. B. (1977). Developing parameters of reading comprehension. In J. T. Guthrie (Ed.), Cognition, curriculum, & comprehension. (pp. 1-15) Newark, DE: International Reading Association.

Chi, M. T. H., Glaser, R., & Rees, E. (1982). Expertise in problem solving. In R. J. Sternberg Ed., Advances in the Psychology of Human Intelligence, (Vol. 1). Hillsdale, NJ: Lawrence Erlbaum Associates.

Cisero, C. A., Royer, J. M., Marchant, H. G., & Jackson, S. (1994). Can the Computer-based Academic Assessment System (CAAS) be used to identify reading disability?: A look at CAAS profiles of reading-disabled college students. Journal of Educational Psychology, 89, 599-620.

Cisero, C. A., Royer, J. M., Marchant, H. G., & Wint, F. (1994). Diagnostic identification of dyslexic and learning disabled college students using the Computer-based

Academic Assessment System (CAAS). Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, LA.

Clark, C. L. (1994). An experimental study of the ability of novice and experienced teachers to comprehend classroom interaction. Unpublished M.A. Thesis: Arizona State University.

Cummins, J. (1984). Bilingualism and special education: Issues in assessment and pedagogy. San Diego, CA: College-Hill.

Dale, E., & Chall, J. S. (1948). A formula for predicting readability. Educational Research Bulletin, January 21 and February 17, 27, 11-20, 37-54.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19, 450-466.

Danks, J. H. (1980). Comprehension in listening and reading: Same or different? In J, H, Danks and K. Pezdek, Reading and understanding. Newark, DE: International Reading Association.

Davidson, A., & Green, G. M. (Eds.) (1988). Linguistic complexity and text comprehension: Readability issues reconsidered. Hillsdale, NJ: Erlbaum.

diSibio, M. (1982). Memory for connected discourse: A constructivist view. Review of Educational Research, 52, 149-174.

Dole, J., Sinatra, G. M., & Reynolds, R. (1991). The effects of strong beliefs on text processing. Paper presented at the Annual Meeting of the National Reading Conference, Palm Springs, CA.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. Reading Research Quarterly,

16, 149-174.

Fletcher, J. M., Francis, D. J., Rourke, B. P., Shaywitz, B., & Shaywitz, S. E. (1993). Classification of learning disabilities: Relationships with other childhood disorders. In G. R. Lyon, D. Gray, J. Kavanagh, & N. Krasnegor (Eds.), Better understanding learning disabilities. (pp. 27-56) Baltimore, MD: Paul H. Brookes.

Fletcher J. M., Shaywitz, S. E., Shankweiler, D. P., Katz, L., Liberman, I. Y., Stuebing, K. K., Francis, D. J., Fowler, A. E., & Shaywitz, B. A. (1994). Cognitive profiles of reading disability: Comparisons of discrepancy and low achievement definitions. Journal of Educational Psychology, 86, 6-23.

Gough, P. B., & Tunmer, W. (1986). Decoding, reading and reading disability. Remedial and Special Education, 7, 6-10.

Greene, B. A., Royer, J. M., & Anzalone, S. J. (1990). A new technique for measuring listening and reading literacy in developing countries. International Review of Education, 36, 57-68.

Hardiman, P. T., Dufresne, R., & Mestre, J. P. (1989). The relation between problem categorization and problem solving among experts and novices. Memory & Cognition, 17(5), 627-638.

Hornberger, N. H. (1989). Continua of biliteracy. Review of Educational Research, 59, 297-314.

Horowitz, R., & Samuels, S. J. (1987) (Eds.) Comprehending oral and written language. San Diego, CA: Academic Press.

Jenkins, J. J. (1979). Four points to remember. A tetrahedral model and memory experiments. In L. S. Cernak & F. I. M. Craik (Eds.), Levels of processing in human

memory. Hillsdale, N. J.: Erlbaum.

Johnston, P. (1984). Prior knowledge and reading comprehension test bias. Reading Research Quarterly, 19, 219-239.

Kardash, C. M., Royer, J. M., & Greene, B. A. (1988). Effects of schemata on both encoding and retrieval of information from prose. Journal of Educational Psychology, 80, 324-329.

Kardash, C. M., & Scholes, R. J. (1995). Effects of pre-existing beliefs and repeated reading on belief change, comprehension, and recall of persuasive text. Contemporary Educational Psychology, 20, 201-221.

Kibby, M. (1980). Intersentential processes in reading comprehension. Journal of Reading Behavior, 12, 299-312.

Kintsch, W. & van Dijk, T. (1978). Toward a model of text comprehension and production. Psychological Review, 85, 363-394.

Kintsch, W., & Vipond, D. (1977). Reading comprehension and readability in educational practice and psychological theory. In L. G. Nilsson (Ed.), Proceedings of the University of Upsala conference on memory. Hillsdale, NJ: Erlbaum.

Kleinman, G., & Schallert, D. L. (1978). Some thinks the reader needs to know that the listener doesn't. In P. D. Pearson & J. Hansen (Eds.), Reading: Disciplined inquiry in process in practice. Twenty-Seventh Yearbook of the National Reading Conference.

Lynch, D. J. (1984). Reading comprehension performance as a function of individual differences in working memory for texts of varying reading difficulty. Unpublished Ph.D. dissertation, University of Massachusetts, Amherst, MA.

Lynch, D. J. (1986). Reading comprehension performance as a function of individual

differences in working memory. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Lynch, D. J. (1987). Reading comprehension under listening, silent and round-robin reading conditions as a function of text difficulty and working memory. Paper presented at the Annual Meeting of the American Educational Research Association, Washington, DC.

Marchant, H. G., Royer, J. M., & Greene, B. A. (1988). Superior reliability and validity for a new form of the Sentence Verification Technique for measuring comprehension. Educational and Psychological Measurement, 48, 827-834.

Mestre, J. P., & Royer, J. M. (1991). Cultural and linguistic influences on Latino testing. In G. Keller, J. Deneen, & R. Magallan (Eds.), Assessment and access: Hispanics in higher education. Albany, N. Y.: State University of New York Press. (pp. 39-66)

Messick, S. (1980). Test validity and the ethics of assessment. American Psychologist, 35, 1012-1027.

Perfetti, C. A. (1992). The representation problem in reading acquisition. In P. B. Gough, L. Ehri, and R. Trieman (Eds), Reading acquisition. (pp. 145-174). Hillsdale, NJ: Erlbaum.

Perfetti, C. A., & Goldman, S. R. (1976). Discourse memory and reading comprehension skill. Journal of Verbal Learning and Verbal Behavior, 15, 33-42.

Pichert, J. W., & Anderson, R. C. (1977). Taking different perspectives on a story. Journal of Educational Psychology, 69, 309-315.

Purwono, U. (1997). Using SVT biology tests as a screening device for community college students enrolling in biology courses. Unpublished M. A. Thesis. Amherst, MA: University of Massachusetts.

Quick, K., & Andre, T. (1992). Schematic influences on the comprehension of AIDS education materials. Paper presented at the Annual Meeting of the Midwestern Educational Research Association, Chicago, Illinois.

Quick, K., & Andre, T. (1993). Sexual schema and the comprehension of AIDS education materials. Paper presented at the Annual Meeting of the American Educational Research Association, Atlanta, Georgia.

Ramos, D., & Bayona, J. (1991). Testing patient comprehension and the comprehensibility of educational materials using the Sentence Verification Technique. (Technical Report). Rochester, NY: University of Rochester Medical School.

Rasool, J. M., & Royer, J. M. (1986). Assessment of reading comprehension using the sentence verification technique: Evidence from narrative and descriptive texts. Journal of Educational Research, 79, 180-184.

Reynolds, C. R. (1981). The fallacy of :two years below grade level for age" as a diagnostic criteria for reading disorders. Journal of School Psychology, 19, 350-358.

Reynolds, C. R. (1985). Measuring the aptitude-achievement discrepancy in learning disability diagnosis. Remedial and Special Education, 6, 37-55.

Royer, J. M. (1996). A cognitive perspective on the assessment, diagnosis, and remediation of reading skills. In G. Phye (Ed.), Handbook of Academic Learning. San Diego: Academic Press.

Royer, J. M., Abranovic, W. A., & Sinatra, G. (1987). Using entering reading performance as a predictor of course performance in college classes. Journal of Educational Psychology, 79, 19-26.

Royer, J. M., & Carlo, M. S. (1991a). Assessing the language acquisition progress

of Limited-English-Proficient Students: Problems and a new alternative. Applied Measurement in Education, 4, 85-113.

Royer, J. M., & Carlo, M. S. (1991b). Transfer of comprehension skills from native to second language. Journal of Reading, 34, 450-455.

Royer, J. M., & Carlo, M. S. (1993). Assessing language comprehension skills in cross-cultural settings. In J. Altarriba (Ed.), Cognition and culture: A cross-cultural approach to psychology. (pp. 157-175). Amsterdam: Elsevier Science Publishers.

Royer, J. M., Carlo, M. S., Carlisle, J. F., & Furman, G. A. (1991). A new procedure for assessing progress in transitional bilingual education programs. Bilingual Review, 16, 3-14.

Royer, J. M., Carlo, M. S., Dufrense, R., & Mestre, J. (1996). The assessment of levels of domain expertise while reading. Cognition and Instruction, 14, 373-408.

Royer, J. M. & Cunningham, D. J. (1981). On the theory and measurement of reading comprehension. Contemporary Educational Psychology, 6, 187-216.

Royer, J. M., & Greene, B. A. (1990). The computer-based assessment of cognitive reading skills in Belieze: Final Report. Arlington, VA: Institute for International Research.

Royer, J. M., Greene, B. A., & Anzalone, S. J. (1994c). Can U.S. developed CAI work effectively in a developing country? Journal of Educational Computing Research, 10, 41-61.

Royer, J. M., & Hambleton, R. K. (1983). Normative study of 50 reading comprehension passages that use the Sentence Verification Technique. Unpublished Study.

Royer, J. M., Hastings, C. N., & Hook, C. (1979). A sentence verification technique

for measuring reading comprehension. Journal of Reading Behavior, 11, 355-363.

Royer, J. M., Kulhavy, R. W., Lee, J. B., & Peterson, S. E. (1986). The sentence verification technique as a measure of listening and reading comprehension. Educational and Psychological Research, 6, 299-314.

Royer, J. M., Lynch, D. J., Hambleton, R. K., & Bulgareli, C. (1984). Using the sentence verification technique to assess the comprehension of technical text as a function of level of expertise. American Educational Research Journal, 21, 839-869.

Royer, J. M., Marchant, H. G., Sinatra, G. M., & Lovejoy, D. A. (1990). The prediction of college course performance from reading comprehension performance: Evidence for general and specific prediction factors. American Educational Research Journal, 27, 158-179.

Royer, J. M., & Sinatra, G. M. (1994). A cognitive theoretical approach to reading diagnostics. Educational Psychology Review, 6, 81-113.

Royer, J. M., Sinatra, G. M., & Schumer, H. (1990). Patterns of individual differences in the development of listening and reading comprehension. Contemporary Educational Psychology, 15, 183-196.

Royer, J. M., Tirre, W. C., Sinatra, G. M., & Greene, B. A. (1989). The assessment of on-line comprehension of computer-presented text. Journal of Educational Research, 82, 348-355.

Salvia, J., & Hughes, C. (1991). Curriculum-based assessment: Testing what is taught. New York: Macmillan.

Shanahan, T., & Kamil, M.L. (1982). The sensitivity of cloze to passage organization. In J. A. Niles & L. A. Harris (Eds.), New inquiries in reading research and

instruction: Thirty-first yearbook of the National Reading Conference. (pp. 204-208) Rochester, NY: National Reading Conference.

Shanahan, T., & Kamil, M.L. (1983). A further comparison of sensitivity to cloze and recall to passage organization. In J. A. Niles & L. A. Harris (Eds.), Search for meaning in reading/language processing and instruction: Thirty-second yearbook of the National Reading Conference. (pp. 123-128) Rochester , NY: National Reading Conference.

Shanahan, T., & Kamil, M.L., & Tobin, (1982). Cloze as a measure of intersentential comprehension. Reading Research Quarterly, 17, 229-255.

Shepard, L. (1980). An evaluation of the regression discrepancy method for identifying children with learning disabilities. The Journal of Special Education, 14, 79-91.

Siegel, L. S. (1992). Dyslexics vs. poor readers: Is there a difference? Journal of Learning Disabilities. 25, 618-629.

Sinatra, G. M. (1989). Implementation and initial validation of a computer-based system for the assessment of reading competencies. Unpublished Ph.D. dissertation, University of Massachusetts.

Sinatra, G. M. (1990). Convergence of listening and reading processing. Reading Research Quarterly, 25, 115-130.

Sinatra, G. M., & Royer, J. M. (1993). The development of cognitive component processing skills that support skilled reading. Journal of Educational Psychology, 85, 509-519.

Stanovich, K. E. (1990). Concepts in developmental theories of reading skill: Cognitive resources, automaticity, and modularity. Developmental Review, 10, 72-100.

Stanovich, K. E. (1991). Discrepancy definitions of reading ability: Has intelligence

led us astray? Reading Research Quarterly, 26, 1-29.

Stanovich, K. E., & Siegel, L. S. (1994). Phenotypic performance profile of children with reading disabilities: A regression-based test of the phonological-core variable-difference model. Journal of Educational Psychology, 86, 24-53.

Sticht, T. G., Beck, L. J., Hauke, R. N., Kleinman, G. M., & James, J. H. (1974). Auding and reading: A developmental model. Alexandra, VA: Human Resources Research Organization.

Swets, J. A., Tanner, W. P., & Birdsall, T. G. (1961). Decision processes in perception. Psychological Review, 68, 301-340.

Tuinman, J. J. (1973-1974). Determining the passage dependency of comprehension questions in five major tests. Reading Research Quarterly, 9, 206-223.

Tuinman, J. J., & Gray, G. (1972). The effect of reducing the redundancy of written messages by the deletion of function words. Journal of Psychology, 82, 299-306.

Turner, A., & Greene, E. (1977). The construction and use of a propositional text base. Institute for the study of intellectual behavior, University of Colorado.

Venezky, R. L., & Calfee, R. C. (1970). The reading competency model. In H. Singer and R. B. Ruddell (Eds.), Theoretical models and processes of reading. (pp. 273-291) Newark DE: International Reading Association.

Walczyk, J. J. (1990). The relationship between error detection, sentence verification, and low-level reading skills in fourth graders. Journal of Educational Psychology, 82, 491-497.

Walczyk, J. J. (1992). A computer program for constructing language comprehension tests. Computers and Human Behavior, 9, 113-116.

Zdenka, M. (1986). Using the Sentence Verification Technique to assess text comprehension. Paper presented at the 5th International Conference in Prague, Prague Czechoslovakia.

Table 1
Examples of SVT and MIT Items

SVT Item Types

Original
A growing body of evidence suggests that the gap in reading achievement between boys and girls may be larger than the gap in math achievement.

Paraphrase
Evidence is increasingly becoming available showing that differences in reading performance between boys and girls is even greater than differences in math performance.

Meaning Change
A growing body of evidence suggests that the gap in reading achievement between boys and girls is smaller that the gap in math achievement.

Distractor
The differences in academic performance are larger in the upper grades than they are in the lower grades.

MIT Item Types

Paraphrase
Evidence is increasingly becoming available showing that differences in reading performance between boys and girls is even greater than differences in math performance.

Meaning Change Paraphrase
Evidence is increasingly becoming available showing that differences in reading performance between boys and girls is smaller than differences in math performance.

Table 2

The Relationship Between Reading and Listening SVT Performance and Several Measures of Reading Comprehension

Standardized Test	Reading SVT and Reading Comp	Listening SVT and Reading Comp
Iowa Test of Basic Skills	.73[a] (N=54)	---
California Achievement Test	.52[b] (N=166)	.23 (N=166)
Stanford Achievement Test	.50[c] (N=46)	---

[a] 5th and 6th grade students
[b] 4th and 6th grade students
[c] 4th and 6th grade students

Table 3

The Interrelationships Between Standardized Reading Comprehension Performance (Reading Comp), Vocabulary Performance, Reading SVT Performance, and Other Measures Requiring Skilled Reading

Standardized Test	Reading SVT and other Measures	Reading Comp and other Measures	Reading SVT and Vocab	Reading Comp and Vocab
Iowa Test of Basic Skills	-	-	.61 (N=54)	.79 (N=54)
Science	.72[a] (N=54)	.89 (N=54)	-	-
Social Studies	.65 (N=54)	.88 (N=54)	-	-
California Ach. Test	-	-	.53. (N=166)	62 (N=166)
Grammatical Expression	.58[b] (N=166)	.73 (N=166)	-	-
Grammatical Mechanics	.58 (N=166)	.60 (N=166)	-	-
Stanford Achievement Test	-	-	.42[c]. (N=46)	70 (N=46)

[a] 5th and 6th grade students
[b] 4th and 6th grade students
[c] 4th and 6th grade students

Table 4

Correlations Between Reading SVT Performance, and Measures Requiring Minimal Reading Skills

Standardized Test	SVT and Other Measures	Standardized Reading Comp and Other Measures
Iowa Test of Basic Skills[a]		
Math Concepts	.33 (N=54)	.69 (N=54)
Math Computation	.28 (N=54)	.56 (N=54)
Spelling	.07 (N=54)	.54 (N=54)
California Achievement Test[b]		
Math Concepts	.34 (N=166)	.37 (N=166)
Math Computation	.15 (N=166)	.12 (N=166)

[a] 5th and 6th grade students
[b] 4th and 6th grade students

Table 5

The Intercorrelations Between Verbal IQ, Standardized Reading Comprehension Test Performance and SVT Reading Performance

Standardized Test	Standardized Reading Comp and Verbal IQ	SVT and Verbal IQ
California Achievement Test[a]	.59 (N=166)	.47 (N=166)
Iowa Test of Basic Skills[b]	.79 (N=54)	.53 (N=54)

[a] 4th and 6th grade students
[b] 5th and 6th grade students