DOCUMENT RESUME

ED 357 036 TM 019 742

AUTHOR	Butler, Frances A.; And Others
TITLE	Benchmarking Text Understanding Systems to Human Performance: An Exploration.
INSTITUTION	National Center for Research on Evaluation, Standards, and Student Testing, Los Angeles, CA.
SPONS AGENCY	Office of Naval Research, Arlington, Va.
REPORT NO	CSE-TR-347
PUB DATE	Sep 90
CONTRACT	N00014-86-K-0395
NOTE	76p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC04 Plus Postage.
DESCRIPTORS	*Adults; *Artificial Intelligence; Comparative Analysis; *Computer System Design; Elementary School Students; Elementary Secondary Education; *Evaluation Methods; Junior High School Students; Man Machine Systems; Performance; Pilot Projects; *Reading Comprehension; Reading Tests; Sample Size; Scores; Systems Development
IDENTIFIERS	*Benchmarking; Comprehensive Tests of Basic Skills; Natural Language Processing; *Performance Based Evaluation; Text Processing (Reading)

ABSTRACT

This study, part of a larger effort to develop a methodology for evaluating intelligent computer systems (Artificial Intelligence Systems), explores the use of benchmarking as an evaluation technique. Benchmarking means comparing the performance of intelligent computer systems with human performance on the same task. Benchmarking in evaluation has been concentrated in the areas of natural language understanding and expert systems. A criterion reading measure was used so that national grade level norms for reading could be established and would provide the anchor for benchmarking the text understanding systems. Eleven text understanding systems were considered, and 6 were finally chosen for a pilot test with 13 adults and 3 school-age students. The refined reading comprehension test was administered to 74 sixth graders, 273 eighth graders and 58 eleventh graders. Comprehensive Test of Basic Skills scores were available for all of the students in the sample. Due to the relatively small and clustered subject sample, it was possible to neither benchmark on a continuous scale nor determine the statistical significance of many of the results. Nevertheless, general descriptive results indicate that a human benchmark methodology can distinguish certain kinds of natural language processing abilities of intelligent computer systems. Six tables present study results. Seven appendixes present supplemental information about the study and the computer systems used. (SLD)

CRESST

National Center for Research
on Evaluation, Standards,
and Student Testing

Benchmarking Text Understanding Systems to Human Performance: An Exploration

CSE Technical Report 347

Frances A. Butler, Eva L. Baker, Tine Falk,
Howard Herl, Younghee Jang, and Patricia Mutch

▸ UCLA Center for the
Study of Evaluation

in collaboration with:

▸ University of Colorado

▸ NORC, University
of Chicago

▸ LRDC, University
of Pittsburgh

▸ The RAND
Corporation

Benchmarking Text Understanding Systems to Human Performm.ance: An Exploration

CSE Technical Report 347

Frances A. Butler, Eva L. Baker, Tine Falk,
Howard Herl, Younghee Jang, and Patricia Mutch

September 1990

Center for Technology Assessment/
Center for the Study of Evaluation
Graduate School of Education
University of California, Los Angeles
Los Angeles, CA 90024-1522
(310) 206-1532

This research report (Govt. Report Accession No. AD M000 093) was supported by contract number N00014-86-K-0395 from the Defense Advanced Research Projects Agency (DARPA), administered by the Office of Naval Research (ONR), to the UCLA Center for the Study of Evaluation. However, the opinions expressed do not necessarily reflect the positions of DARPA or ONR, and no official endorsement by either organization should be inferred.

Acknowledgement

The authors wish to thank Anat Jacoby for her assistance in identifying the text understanding systems reviewed for this study and also for her comments on an earlier draft of the paper.

In addition, special thanks to Carol Lord, Vice-President, Intelligent Text Processing, Inc., Santa Monica, CA, for acting as linguistic consultant to the project. Her input was invaluable both on a conceptual and an analytic level.

Introduction

The present study is part of a larger effort by the Center for Technology Assessment at UCLA to develop a methodology for evaluating intelligent computer systems, also referred to as Artificial Intelligence (AI) Systems. Four broad areas of inquiry—vision, natural language understanding, expert system shells, and expert systems—have all figured into the development of a methodological approach which focuses largely, though not exclusively, on benchmarking intelligent computer systems to human performance. At UCLA the use of benchmarking as an evaluation technique has been concentrated in the areas of natural language (NL) understanding (Baker & Lindheim, 1988; Baker, Turner, & Butler, 1990) and expert systems (O'Neil, Ni, & Jacoby, 1990; O'Neil, Ni, Jacoby, & Swigger, 1990).

The research reported here continues to explore the use of NL in intelligent computer systems specifically with regard to text understanding systems. The specific goal of this research is to benchmark selected text understanding systems to human performance in reading comprehension. The research is exploratory, but our findings promise to contribute to the development of an innovative methodology for assessing intelligent computer systems.

We first consider the concept of benchmarking and how it is applied in this study. Next, we discuss various aspects of the development of the reading comprehension test, the test used to measure both human and computer performance in reading comprehension. This section is followed by a section detailing the ımplementation and scoring of the test. We then present our methods of analysis and end by presenting and discussing the results.

Benchmarking

Benchmarking in the present context means quite simply comparing the performances of intelligent computer systems to the performances of humans on the same task. In this study, computer responses to questions based on specific reading texts are referenced back to human responses to the same questions about the same texts.

While intelligent computer systems can clearly be evaluated on various levels ranging from speed in accomplishing a task to effectiveness at accomplishing the task to sophistication of programming techniques, it is beyond the scope of this study to consider all such levels. Rather, our goal is to focus exclusively on the answers a system provides for specific types of questions and to use the answers to benchmark the system. By taking this approach, we are looking at the output of a system irrespective of the processes it might go through to produce that output.

Eleven intelligent computer systems were reviewed (Jacoby, 1989). Texts from six of the systems that answered specific questions about the texts they read were selected for use in the study. (The text selection process is discussed in detail below.) The texts and questions were used to form a reading comprehension test which was administered to the subjects in this study along with a criterion reading measure.

In an earlier study at UCLA (Baker et al., 1990), a natural language query system was referenced to the performance of kindergartners and first graders. A criterion measure of language ability was used to allow for the grouping of students by national grade equivalent norms. This grouping provided a means for benchmarking the NL query system. The study yielded the beginning of a continuum of difficulty for the NL understood by the query system.

For the present study, a similar approach was taken. A criterion reading measure was used so that national grade level norms for reading could be established and would thus provide the necessary anchor for benchmarking the text understanding systems.

Development of the Reading Comprehension Test

This section describes (a) the selection of texts and questions for the reading comprehension test, (b) pilot testing, and (c) answer key development.

Selection of Texts From AI Systems

We reviewed a variety of text understanding systems and selected from them the texts used in our reading comprehension test. We looked mainly at research systems rather than commercial systems (e.g., the commercial system CONSTRUE, Carnegie Group, Inc.) because research systems, although limited since they focus on specific problems, employ the most current knowledge and state-of-the-art technology.

We considered in detail 11 research systems: 7 from Yale University, 3 from UCLA, and 1 from the IBM Los Angeles Scientific Center. All these systems, with the exception of the IBM system, address specific concerns such as the understanding of irony or the implementation of knowledge structures, for instance, scripts or goals. The domains of the systems vary in size and type: Some domains are as small as one text (e.g., Reeves & Dyer, 1986); others are larger in size but restricted to a certain type of text, for example, news stories, editorials, or melodramatic divorce stories. In the following paragraphs we briefly discuss each of the 11 systems in order to familiarize the reader with them and to provide an explanation of the constraints these systems impose on the texts they can process.

We looked at seven Yale-based systems: SAM (Cullingford, 1978), PAM (Wilensky, 1978), BORIS (Dyer, 1983), POLITICS (Carbonell, 1979), FRUMP (DeJong, 1979), IPP (Lebowitz, 1980), and CYRUS (Kolodner, 1980). SAM understands stories by identifying an appropriate script from a programmed collection of scripts. A script is a prepackaged set of expectations, inferences, and knowledge that is applied in a common situation, for example, a doctor's visit or eating at a restaurant. A script is analogous to a blueprint for action without the details filled in. SAM understands a text by identifying an appropriate script and then filling in the blanks with information from the text. SAM's domain is a variety of stories including actual newspaper stories.

PAM was built to test the idea of goals and plans. PAM functions using a theory of goal interaction: rules we have for resolving goal conflicts, achieving temporarily blocked goals, etc. Unlike SAM, PAM can understand stories which do not match stereotypical scripts. But PAM has other limitations including its inability to recognize counterplans. PAM's domain is the same as SAM's: a variety of stories including actual newspaper stories.

BORIS is a system which combines scripts with plans and goals, uses additional methods such as interpersonal relations, role themes, and affect in understanding a text, and also uses a new set of memory structures. The domain of BORIS is melodramatic divorce stories, a different domain from that of SAM and PAM.

FRUMP, IPP, POLITICS, and CYRUS were eliminated from our considerati~~n because they do not understand texts in the sense of reading a specific text and answering questions about it. FRUMP and IPP both read input texts and give summaries of the texts. They do not answer specific questions about the texts they read. FRUMP is script-based and IPP uses plans and goals as well as scripts. Both systems sacrifice careful understanding for speed. The domain for both systems is newspaper articles taken from the UPI news wire.

POLITICS answers questions about a subject presented in an input text but draws exclusively on information not provided in the text to answer these questions; it draws on information in a database. CYRUS answers questions about a particular individual, Cyrus Vance, asked in the absence of an input text, from information in its memory. The memory in CYRUS changes constantly as the program receives and processes all the stories from FRUMP about Cyrus Vance. Although POLITICS and CYRUS do not read and understand texts in a way appropriate to a reading comprehension test such as ours, their ability to draw on information in memory when answering a question does mimic the ability of human subjects to draw on information in memory. The human subjects in our test population most likely used this ability to some extent in answering questions on our reading comprehension test.

The three UCLA-based systems we looked at were JULIP (August & Dyer, 1985, 1986), OpEd (Alvarado, 1990), and Reeves (Reeves & Dyer, 1986). JULIP and OpEd are part of a larger project which has as its goal the understanding of letters to and from the editor. JULIP can understand only one text and can answer only three questions about this text. JULIP deals with the role of analogy in arguments. It recognizes and understands analogy by recourse to lexical clues and the comparison of conceptual similarities.

OpEd understands editorial texts by focusing on argument structure. It considers goals and plans, recognizes belief relationships, and determines argument units and the structure of these units. The editorials that OpEd understands have been edited to remove reference to issues outside the scope of OpEd's process model. Reeves takes as its goal the understanding of irony in stories. The system understands one ironic story and can answer three questions about that story.

We also looked at a text understanding system developed by the NewSelector Project at the IBM Los Angeles Scientific Center, the Kind Types system (Dahlgren, 1988). The Kind Types system is, unlike the other research systems discussed so far, a full-scale text understanding system (i.e., it is not restricted by its method of understanding to a limited type of English language texts). The Kind Types system uses a parser, logic translator, and Naive Semantics representations interactively in its understanding of a given text. The Naive Semantics Lexicon is divided into four databases—ontological schema, generic information, typing information, and Kind Types—which each require specific types of reasoning.

As mentioned previously, the requirement that systems be able to understand and answer questions about an input text eliminated four of the eleven systems described above as sources for texts for our reading comprehension test. In addition, the Reeves system was eliminated because the content of the one story it understands (killing a rabbit with dynamite) was inappropriate for the subject population. The six remaining systems represent a variety of approaches to the problems of text understanding: scripts, reasoning about goals and plans, analogy, argument structure, full-scale parsing and semantic interpretation, and various combinations of these approaches (e.g., scripts and goals/plans in BORIS).

In selecting texts for our reading comprehension test, we limited the number of texts we took from each of the six systems to one. When it was necessary to choose among texts from one system, we evaluated the texts according to (a) the appropriateness of the subject content and (b) our desire that the final group of texts represent a range of subject contents.

The six texts remaining after this selection process was completed were combined to form the reading comprehension test.[1] The only change we made to the test was to omit two questions that provided answers to other questions

[1] We had originally planned to include analogous texts and questions on the reading comprehension text, but had to forgo doing so because of limitations on testing time and hence on test length. Further research could include the use of analogous texts—texts similar in structure to the original texts but different in content domain—to investigate the effect of content knowledge on reading. In addition, two types of analogous questions might be used: One would involve paraphrasing an original AI question omitting ambiguous phrasing and difficult vocabulary. The second would alter the syntax of a question but retain the intent. It would be useful to see how computer systems and human subjects of different ages handle the variations which analogous texts and questions introduce. Doing this would help reduce the potential error that could result from the test takers not being familiar with a particular topic or not understanding a particular syntactic structure.

asked about the SAM text. This resulted in a test with 48 questions. The questions were not equally distributed among the texts. The texts from BORIS and the Kind Types system had 24 and 11 questions, respectively, associated with them. The remaining four texts were each followed by either three or four questions. See Appendix A for a list of the systems considered and those chosen. Appendix B provides the reading comprehension test with answer key.

Pilot Testing

Once the AI texts and questions were selected and the reading test assembled, we began an informal pilot testing procedure. The goals of the procedure were to establish necessary testing time and to identify problems with the test directions or problems with test passages or individual questions. In addition, we planned to use the responses of the adult subjects to provide the first draft of the answer key.

The procedure for administering the pilot test was as follows: Directions written on the test explained that the test was a reading comprehension exercise that required the examinee to read each passage and write a short answer in the space provided. If there was a question the examinee could not answer, he or she was to indicate why in the answer space. The test was administered on an individual basis. Either the examinee was timed by an examiner or, where an examiner was absent, the examinee was asked to note on the test the amount of time necessary to complete the exercise.

The subjects for pilot testing included 3 school-age students, one each from the 5th, 8th and 10th grades, and 13 adults, all of whom were college graduates and most of whom were graduate students in education. Because the subject pool for pilot testing was limited, we felt that testing a 5th, an 8th, and a 10th grader along with the adults would provide a reasonable grade increment for detecting p.oblems with the test.

Since actual testing would need to be completed in a 45- to 55-minute class period, time necessary to complete the test was a critical consideration. The 5th grader in the pilot sample did not complete the test. The 8th grader and the 10th grader took 20 and 40 minutes, respectively. The adults completed the test in 15 to 30 minutes. The range of time for the pilot sample was 15 to 40 minutes; thus, we concluded that the reading test without length modification could be comfortably completed in a 45- to 55-minute class period. Because the 5th grader in the pilot sample did not complete the test, we decided to test 6th graders along with 8th graders. Eleventh graders rather than 10th graders were ultimately available to us. We felt the 6th/8th/11th-grader range would provide the information needed for the benchmarking effort.

No major problems regarding directions, passage content, or questions emerged from the pilot testing. Some subjects commented that a particular text sounded stilted because few pronouns were used. Others felt the differences in passage difficulty and topics seemed haphazard and strange. Finally, some subjects thought some questions too simple and thus felt they might be trick questions. These problems were alleviated by having the test administrator give the students an explanation of where the text and questions originated. That is, the modified directions explained that the texts are ones that computer systems can read, and the questions are those they can answer. Therefore, some of the content may seem unusual or strange to humans.

Answer Key Development

The development of the answer key took place in stages beginning with the pilot testing. As mentioned above, we used the adult responses from pilot testing to produce an initial draft of the answer key. The initial draft was then modified based on issues that arose during practice scoring sessions designed to train scorers. Since the subjects in the study were not adults, there were occasionally peculiarities in responses that did not occur with the adult answers. The steps below describe the specifics of the process which led to the final answer key. The final answer key is given in Appendix B.

Step 1. Answers from the 13 adults who took the pilot test provided the basis for the answer key. These answers were reviewed by a committee of six researchers and a few obviously incorrect responses were eliminated. Then answers with greatest consensus, that is, answers given by the largest number of the adult examinees, were regarded as correct answers and awarded 2 points. Answers given by only one or two examinees were discussed among the six researchers and given 2 points if similar to the high consensus answers, 1 point if somewhat similar, and no points if not similar at all.

Step 2. After completing the answer key draft, six researchers scored 25 randomly selected tests from the actual subject pool, noting any difficulties or unclear cases in the process. Interscorer reliabilities were calculated and problems that had arisen with the answer key were discussed. Additional correct answers and guidelines for scoring the responses to specific questions were added to the answer key.

Step 3. Another 10 randomly selected tests were scored by the six researchers. Interscorer reliabilities were again calculated. Answers with low agreement usually involved issues not covered by the current answer key. Those answers were discussed and incorporated into a revised answer key.

Step 4. Ten more randomly selected tests were scored by the six researchers. Interscorer reliability was sufficiently high, .96, to allow us to begin scoring the test.

Implementation of the Reading Comprehension Test

The reading comprehension test was administered to more than 300 students. To rank scores on the reading comprehension test against grade level equivalencies for reading comprehension, we used the students' reading comprehension scores on the Comprehensive Test of Basic Skills (CTBS). The CTBS was selected in part because it was being used in one of the school districts where we would be testing. CTBS scores from that district were made available to us; in the other districts, the CTBS was administered as part of the study.

Subjects

The subjects for the study came from three different school districts and five different schools in southern California. From the first district there were 74 sixth graders and 110 eighth graders. Those subjects were almost entirely native English speakers. From the second district, 163 eighth graders were tested. Approximately 80% of those students were native English speakers. Students from both of the above districts came from families of middle socioeconomic status. There were 58 eleventh graders tested in the third district. Those subjects came from families of high socioeconomic status, were primarily native English speaking, and were members of an honors class in English.

Test Administration

Approximately half the subjects were tested by the project researchers and the other half tested by their teachers. Where researchers tested the students, students were given the reading comprehension test during one period, and then one day to one week later, given the CTBS reading comprehension section. Where teachers tested their students, all sections of the CTBS test were given one to two weeks prior to the reading comprehension test. In the first district, the testing was conducted entirely by the classroom teachers. In the second and third districts, the researchers administered the tests.[2]

For the reading comprehension test, students were directed to read each passage and question and write a short answer to the question in the space provided. If the student could not answer a question, he/she was to indicate why in the answer space. In addition, the students were told that the text passages and questions on the test were items that computer systems could read and answer. The researchers, they were told, were interested in seeing how humans answer those same questions. Directions for the CTBS were taken directly from the *Test Administrator's Guide*.

Scoring Procedure for the Reading Comprehension Test

When testing and answer key development were complete and actual scoring was to begin, the tests used for the answer key formulation were mixed back in with the original tests, which were then randomly divided in half. There were two scorers; each scored one half of the tests. A total of 305 tests were scored.

On the reading comprehension test there were a total of 48 questions and a possible score range of 0 to 96 points. As mentioned above, the subjects' answers to each question were graded on a 0 to 2 point scale depending on the degree of correctness. If an answer to a question was totally incorrect, a score of zero was given. One point was given when the subject's answer reflected some understanding of the passage but was not sufficiently informative to be considered totally correct. For example if the subject understood part of the

[2] All students were given Form U of the CTBS. District 1 sixth graders were given Level G, District 1 eighth graders were given Level H, District 2 eighth graders were given Level J, and District 3 eleventh graders were given Level K.

story only, if his/her response was incomplete or was not specific enough, a partial credit of 1 point was given. Also, when a subject brought his/her world knowledge to bear but did not provide a totally correct answer to a question, 1 point credit was given. Subjects who showed complete understanding of the passage in their answers to a question were given the maximum of 2 points. If an answer was given that included the correct responses plus some additional information, the answer received 2 points if the additional information was plausible. Otherwise, zero to 1 point was given. Answers with incorrect spelling or awkward sentence structure which did not indicate misunderstanding of content were not penalized. For example, 6th graders often misspelled "divorce" and wrote it as "deforce" or wrote "too" as "to."

To obtain a high degree of consistency in grading, the following method was adopted: Each question was graded separately for all subjects within a class, that is, the first question was graded for all subjects within a given class, then the second question was graded, etc. This approach was preferable to grading an entire test for a subject and then going on to the next subject and test because it allowed the scorers to focus on answer variation for a specific question.

Twenty-five tests were scored by both scorers in order to establish interscorer reliability. The overall interscorer reliability of the two graders for the answers of these 25 subjects was .97. The individual item agreement between the two scorers ranged from .83 to 1.00. When the scorers encountered an ambiguous answer, they discussed the response and came to agreement about the score.

Development of Analysis

When a reader correctly answers a question about a text, this indicates that he or she has understood the question and the information in the text which is relevant to the question. Therefore, we focused on questions rather than entire texts in our efforts to describe and categorize the grammatical and conceptual structures and information necessary to answer a question correctly. We first looked at the Iowa Test of Basic Skills (ITBS) classification of question types (Hieronymus et al., 1986, pp. 50-51), but because this classification did not describe adequately the variables we were interested in, we decided to devise a classification scheme tailored to the needs of our study.

The classification scheme developed for the purposes of our study specifies (a) type of knowledge required to answer a question and (b) the linguistic/ textual domains addressed by a question.[3] We discuss classification schemes for questions more fully in the next section.

We also needed to establish an objective measure of text difficulty to describe differences that might occur at a given grade level equivalent for a given question type across systems. We explored two methods of determining objective difficulty of texts. First, we attempted to derive a measure of difficulty based on syntactic structures. Second, we looked at established readability measures. These two methods and our decision to use the latter are discussed in the section below entitled "Methods to Determine Objective Difficulty of Texts."

Categorization of Question Types

Answering questions about a text can require accessing different types of knowledge, including world knowledge from firsthand or vicarious experience of physical and socio-cultural phenomena, as well as linguistic knowledge of the particular language, comprising vocabulary, syntactic patterns, cohesive devices, and discourse structure. Answering a question may require information that is stated directly in the text; other questions may require the answerer to make inferences based on his/her world knowledge.

Following the texts below are similar questions:

"Twas brillig, and the slithy toves did gyre and gimbal in the wabe."[4]

Q: Where did the slithy toves gimbal?

Joe and Fred drove in Joe's car to Fred's cabin on Lake Minnetonka. They went for a swim before dinner.

Q: Where did Joe and Fred go for a swim?

A reader who knows English can tell where the slithy toves probably gimballed even without world knowledge of toves or gimballing. However, unless the reader knows that lakes are typical swimming places (or that people swim in water, and that lakes typically contain water), he/she could fail to infer that the

[3] The type of knowledge and linguistic domain classfications were suggested by Carol Lord and were adopted by the project team.

[4] Taken from "Jabberwocky" (Carroll, 1928, p. 178).

lake was probably where Fred went for a swim. Answering the second question correctly requires world knowledge.

Answering the second question also requires more linguistic reasoning. The answer to the first question is found in a single clause. To answer the second question, the reader must integrate information from two clauses to identify "Joe and Fred" as the antecedent of the pronoun "they."

We wanted to recognize differences such as these in our study. To reflect these differences, we used a classification based on the type of knowledge required to answer a given question in light of the particular text being queried. Within the limited scope of this study, it was not possible to address and control for the large number of variables which are potentially relevant. We therefore opted for a simple, two-way classification of (a) knowledge from identification and (b) knowledge from inference. Within each of these classifications we defined subgroups based on a survey of the text-question pairs in the corpus, as follows:

Type of Knowledge Required

1. Identification

    1.1 Answer stated directly in text; requires knowledge of English function-word vocabulary and English clause structure.

    1.2 Answer stated indirectly in text; requires knowledge of referential processes such as anaphora, paraphrase, appositionals (as well as clause structure).

2. Inference

    2.1 Answer stated indirectly in text; inferable from knowledge of discourse structure and/or world.

    2.2 Answer not stated in text, but inferable from world knowledge.

    2.3 Answer not stated in text, and not unequivocally inferable from world knowledge.

Each text-question pair in the reading comprehension test was classified according to the type of knowledge required to answer the question correctly, given the particular text being accessed. The classifications are listed in Appendix C.

The questions on the test addressed different linguistic/textual domains. Some were relatively straightforward questions with "who" or "what," for which the reader (or AI system) needed only to identify the relevant participant. Others asked the reader to assess the truth of a proposition, to identify the stated or implied discourse relation between two propositions, or to draw a meta-propositional inference. To address such differences, we decided to use a simple, four-way classification based on the question's linguistic/textual domain. These are broad groupings; we recognize that our classification lumps together distinctions which might be assigned to separate sub-categories given a larger sample size. The classifications are as follows:

Linguistic/Textual Domain

A. Intra-clausal. Identification of participant or constituent within the clause: WH - questions.

B. Clausal. Evaluation of proposition's truth: yes-no questions.

C. Inter-clausal. Identification of relation between two propositions, e.g., temporal sequence/concomitance, cause-effect, goal/reason, consequence, belief complements.

D. Discourse. Multi-clausal, e.g., identification of parallel, identification of topic of discourse segment.

Each question in the test corpus was assigned to one of these groups. The classifications are listed in Appendix C.

Methods to Determine Objective Difficulty of Texts

As mentioned previously, we explored two methods of determining objective difficulty of text passages. We looked at difficulty based on syntactic structures, and we considered established readability measures. Studies of assessment of difficulty in terms of syntactic structures are scarce in the literature. Most research into the ease of understanding of texts, or readability, focuses on indirect evaluation of syntactic difficulty, most often a simple account of the number of words per sentence or some variant of this variable (Klare, 1984). Klare (1984, p. 686) notes that most readability formulas are limited to two variables, one representing the semantic factor and one the syntactic factor. The semantic factor is typically a measure of word length in syllables and the syntactic factor a measure of sentence length in words. Very few of the studies that Klare references take specific syntactic structures into consideration. A formula developed by Williams, Siegel, Burkett, and Groff (1977) includes measures of transformational complexity, center embedding, and right-branching as variables, and a formula by Hull (1979) includes a measure of prenominal modifiers. But other than these two studies, examination of specific syntactic structures is rare.

We developed an experimental measure of syntactic difficulty which included as variables the number of passive constructions, subordinate clauses, ellipsis, order reversals, and anaphors. These structures correlate with those isolated by Oakhill and Garnham (1988) as structures which are acquired late in human acquisition of language. Oakhill and Garnham (1988) consider in detail the issue of developmental acquisition of syntactic structures. Their experimental studies indicate that many of the structures acquired after age five have an "irregular" relation between form and meaning. These structures are exceptions to the association of the most common form of English sentences (Noun Verb Noun) with the most common semantic role assignment (Subject Verb Object). The basis of our experimental measure of syntactic difficulty is that structures acquired iate in development are more complex and take longer to process than those acquired early. Although we ultimately decided that the development of a measure of syntactic difficulty was outside the scope of this study, the issue of syntactic difficulty vis-à-vis intelligent computer systems merits further investigation which could be attempted in an extension of this study.

In addition to measuring syntactic difficulty of the six texts on the reading comprehension test by means of our experimental measure, we also used the software program *Sensible Grammar* (Long, 1989) to calculate readability levels based on the Flesch Reading Ease formula (Flesch, 1949), an established readability formula with the variables of average sentence length and number of syllables per 100 words. The relative difficulty levels of the texts determined using our experimental measure of syntactic difficulty did not match the relative difficulty levels determined using the Flesch Reading Ease formula. Agreement between the two formulas is not to be expected however. The only variable in our syntactic structure measure that correlates with increasing sentence length and word length as measured by the Flesch formula is the number of subordinate clauses; all the other syntactic structure variables correlate with stable or decreasing sentence length and word length. This observation is only noted here; to pursue it and its implication for measurements of readability is beyond the scope of this paper.

We decided to use the Flesch Reading Ease formula, as implemented by the program *Sensible Grammar*, to determine objective difficulties of the text passages in this study. Readability measures have proved to be gross indicators of text difficulty although they do not directly measure causal factors. We recognize the deficiencies of and the problems associated with simple readability formulas (see, for example, Davison & Kantor, 1982; Manzo, 1970; Walmsley, Scott, & Lehrer, 1981) and have used the Flesch formula, keeping these problems in mind. The program *Sensible Grammar* determined the Flesch Reading Ease value and the Flesch-Kincaid Grade Level for each of the six texts on the reading comprehension test (see Appendix D). Ultimately the objective difficulty of text passages per se did not figure into our analyses. Still, we recognize its impact on reading comprehension and note in our discussion of the results of individual systems how text difficulty could be used in analysis.

Results: Reading Comprehension Test

The results reported below are initially presented in three sections. The first provides the descriptive statistics for subject performance by grade level in school. The second provides the descriptive statistics for subject performance by grade equivalence for reading.[5] The third provides the descriptive statistics

[5] Since subjects took four different levels of the CTBS, the subjects' raw scores were normed to scale scores which were converted to grade equivalency scores using the *CTBS Norms Book* (*Comprehensive Test of Basic Skills, Norms Book*, 1983). The purpose of using grade equivalents was to determine the reading levels of all students according to one scale independent of grade membership or level of the test taken. For example, a student in the 8th grade and a student in the 12th grade who took different levels of the CTBS reading comprehension section can both receive the same grade equivalent score. If they both receive a 10.9, for example, this indicates that the 8th grader scored as well as the average 10th grader in the ninth month of school would score on the test the 8th grader took and that the 12th grader scored as well as the average 10th grader in the ninth month of school would score on the test the 12th grader took. It does *not* mean that the 8th grader, for example, has mastered all the reading skills that the average 10th grader in the ninth month of school has.

for grade equivalency groupings, which become our unit of analysis for subsequent discussions.

Grade Level

Table 1 provides the descriptive statistics for the total reading comprehension test by the three grade levels in the study—6, 8, 11. It also includes the combined system responses for the six texts.

While our focus in the study is on benchmarking individual system performance, Table 1 provides an initial overall picture of subject performance compared to computer performance. It is important, however, to remember that the system response total provided in Table 1 is an aggregate of scores from the six text understanding systems in the study and should not be interpreted as a total score earned by a single system.

With the student responses, there is a clear increase in the mean score for each grade. Because of the grading system, 0-2 points possible per item, a perfect score on the test was 96. As a group, the 11th graders ($\overline{X}$=81.49) performed better than the 8th graders ($\overline{X}$=72.39) who in turn performed better than the 6th graders ($\overline{X}$=69.30). These differences as determined by a one way analysis of variance were significant. However, the standard deviations and the ranges show a considerable variation in subject performance especially in

Table 1

Descriptive Statistics for Total Score on the Reading Comprehension Test by Grade Level

	Reading Comprehension Test			
Grade Level	Mean	SD	Range (0-96)	N
6	69.30	10.09	36-85	61
8	72.39*	7.45	53-90	161
11	81.49**	5.97	67-90	43
System Response	81.00	–	–	–

* Significantly larger than the previous grade level at $p<.05$.

** Significantly larger than the previous grade level at $p<.01$.

grades 6 and 8. The 11th graders were a more homogeneous group in terms of reading ability as measured by the reading comprehension test. This was perhaps to be expected since the two 11th grades classes were honors classes from the same school. The total system response, 81.00, was slightly lower than the mean for the 11th grade. Also, it should be noted that at least some students at each grade level had a higher overall score than the aggregate computer score. These results are given only to provide an overall impression of the data.

Grade Equivalence

We were interested in benchmarking six text understanding systems to a performance-based measure of reading ability with students from 6th, 8th, and 11th grades. To this end, scores on the CTBS reading comprehension subtest were used to establish grade equivalencies in reading. The grade equivalencies for the students tested ranged from grade 4 to grade 13. Table 2 provides the descriptive statistics for grade equivalencies based on the total Reading Comprehension Score.

Table 2

Descriptive Statistics for Total Score on the Reading Comprehension Test by Grade Equivalencies

	Reading Comprehension Test			
CTBS Grade Equivalence	Mean	SD	Range (0-96)	n
4	61.75	17.86	36-76	4
5	64.92	8.81	54-82	13
6	63.92	7.81	51-76	12
7	67.00	4.00	63-71	3
8	70.29	5.06	63-82	14
9	69.51	8.45	37-84	59
10	73.15	6.03	61-85	34
11	73.54	7.21	54-85	41
12	74.75	6.14	65-87	12
13	80.03	6.24	63-90	73
System Response	81.00	–	–	1

The grade equivalencies in Table 2 must be viewed with extreme caution. The N sizes in general are too small to allow us to interpret the means with any confidence. Furthermore, the N sizes for 6 of the 10 grade levels are too small to allow for further statistical analysis. For this reason, we examined Table 2 with an eye towards grouping the grade equivalents in some meaningful way.

Grade Equivalency Groupings

There seemed to be natural breaks in the means, that is, distances of several points, which allowed us to produce the following 4 grade equivalent groups: Group 1 (grade equivalencies 4, 5, and 6), Group 2 (grade equivalencies 7, 8, and 9), Group 3 (grade equivalencies 10, 11, and 12, and Group 4 (grade equivalence 13). As it turned out, the groups correspond to school levels in the following way: Group 1, upper primary; Group 2, junior high school; Group 3, senior high school, and Group 4, college freshman and above. Table 3 provides the descriptive statistics for the four groups based on the total reading comprehension score.

Table 3

Descriptive Statistics for Total Score on the Reading Comprehension Test by Group

	Reading Comprehension Test			
Group	Mean	SD	Range (0-96)	N
Group 1 (4,5,6)[a]	64.07	9.62	36-82	29
Group 2 (7,8,9)	69.55*	7.77	37-84	76
Group 3 (10,11,12)	73.55*	6.57	54-87	87
Group 4 (13)	80.03*	6.24	63-90.	73
System Response	81.00	–	–	1

[a] Grade equivalencies per group.

* Significantly larger than the previous group at $p<.01$.

The results in Table 3 show that better readers, as determined by the CTBS, score higher in general on the reading comprehension test. See Appendix E for the correlation between the total score on the reading comprehension test and the CTBS grade equivalency groups and the correlations between scores on the individual system texts and the CTBS grade equivalency groups. Appendix E also provides the correlations for grade level rather than CTBS grade equivalency group. The correlations with the grade equivalency group are higher than the correlations with grade level. This validates the decision to use grade equivalency group rather than school grade level for benchmarking.

As mentioned above, we were interested in benchmarking the six text understanding systems whose texts comprised the reading comprehension test to the performance of humans. While our focus henceforth is on the individual systems, with Table 3 just as with Tables 1 and 2, it is possible to get an overall impression of subject performance compared to system performance. The means for the 4 grade equivalent groups show a marked increase from the lowest to the highest group. The mean increases are, in fact, significant. That is, the difference in the means between Groups 1 and 2 is significant, as is the difference between Groups 2 and 3, and so on. It should be noted, however, that the ranges indicate some overlap in performance among the groups on the reading comprehension test, most notably with Groups 1 and 2.

Results: Individual Systems

In the following discussions of the benchmarking of individual systems, benchmark levels are generally determined between grade equivalent groups. Only in certain instances of benchmark determination, where we considered information additional to group mean scores, was it possible to specify benchmark levels at particular grade equivalencies. Regardless of the system's level of performance, the range of student scores in each case included a score equal to or higher than the system score.

Discussion of Total System Scores by Group

Table 4 provides the descriptive statistics for the six computer systems by grade equivalency groups.[6] The data in this table were used for establishing system benchmarks.

With four of the six systems, PAM, JULIP, OpEd, and SAM, a comparison of group means with computer system responses indicated that the computer performed as well as or better than students in Group 4 (at or above grade equivalent 13).

PAM Group 4 mean = 6.76, system response = 7.00

JULIP Group 4 mean = 4.11, system response = 6.00

OpEd Group 4 mean = 3.89, system response = 4.00

SAM Group 4 mean = 5.73, system response = 6.00

In these four cases, the reading level benchmark can be placed at or above grade equivalent 13.

With the two remaining systems, BORIS and Kind Types, the benchmarks are lower. With BORIS the system response of 38.00 is slightly lower than the Group 4 mean of 38.44. The next highest mean is 36.86 for Group 3. Clearly BORIS falls between the two groups, closer to Group 4 than to Group 3. The 95% confidence intervals for Groups 3 and 4 indicate that BORIS should, in fact, be benchmarked at the low end of Group 4 (grade equivalent 13). The 95% confidence intervals for Groups 3 and 4 are [36.0, 37.8] and [37.5, 39.4] respectively.

With Kind Types, the system response of 20.00 falls between the means for Groups 2 ($\bar{X}$=19.26) and 3 ($\bar{X}$=20.53). The 95% confidence intervals for Groups 2 and 3, [18.6, 19.9] and [20.1, 20.9] respectively, place the benchmark for the Kind Types system between Groups 2 and 3, specifically between grade equivalents 9 and 10. With BORIS and Kind Types there were more questions asked about each text than in the other systems. It is unclear what impact, if any, increasing the number of questions associated with the texts in the other systems would have on subject performance.

[6] The N sizes in Table 4 and all other tables and appendices that distinguish different systems and/or question types may vary by grade, grade equivalency, or grade equivalency group because the subjects did not always answer all the questions.

Table 4

Descriptive Statistics for the Six Computer Systems by Grade Equivalency Groups

	AI Systems																	
	PAM Max Possible=8			JULIP Max Possible=6			BORIS Max Possible=48			OpEd Max possible=6			SAM Max Possible=6			Kind Types Max Possible=22		
Group	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range	Mean	SD	Range
Group 1 (4,5,6)[a] [n]	4.97	1.77	0-8 [35]	2.37	1.70	0-6 [35]	32.38	6.03	18-41 [34]	1.09	1.22	0-4 [34]	4.56	1.86	0-6 [34]	17.86	3.2	9-22 [29]
Group 2 (7,8,9) [n]	5.51	1.54	3-8 [80]	2.13	1.49	0-6 [79]	35.73*	4.54	23-43 [80]	1.61	1.46	0-6 [80]	5.21	1.1	2-6 [80]	19.26	2.7	6-22 [77]
Group 3 (10,11,12) [n]	5.90	1.5	2-8 [90]	2.57	1.66	0-6 [90]	36.86	4.30	23-45 [90]	2.32*	1.61	0-5 [90]	5.27	1.0	2-6 [90]	20.53	1.93	14-22 [87]
Group 4 (13) [n]	6.76*	1.3	2-8 [76]	4.11*	1.61	0-6 [76]	38.44**	4.03	25-45 [75]	3.89*	1.35	0-6 [74]	5.73	0.6	4-6 [73]	21.11	1.43	16-22 [73]
System Response [n=1]	7.00	–	–	6.00	–	–	38.00	–	–	4.00	–	–	6.00	–	–	20.00	–	–

[a] Grade equivalency per group.

* Significantly larger than the previous group at $p<.05$.

** Significantly larger than the previous group at $p<.01$.

It should be noted that in establishing system benchmarks we used descriptive data only. Due to the limitations of our sample, homogeneity problems emerged within certain groups, disallowing the use of an ANOVA to establish significant differences between means. In spite of this limitation, we feel that given an adequate sample size, the procedures described in this paper present a promising approach for establishing benchmarks for intelligent computer systems. Appendices F and G show total system scores on the reading comprehension test by school grade level and by grade equivalence respectively. While neither grade level nor individual grade equivalence is being used to determine benchmarks, the data are available for purposes of comparison.

Benchmarking by Question Types

In addition to benchmarking systems by total system scores, we were interested in comparing systems by (a) types of knowledge and (b) linguistic/ textual domains required to answer questions. Analyses of the data utilizing these two divisions from our scheme of question classification (see above) are presented in turn.

Before we begin our discussion, it should be noted that not all of the question types in our scheme of question classification are of equal difficulty. For example, type B questions, clausal yes/no questions, are generally easy to answer. This is reflected in the human and computer scores for this question type: Human scores were high and each computer system with type B questions scored the maximum possible number of points on them.

Type of Knowledge

The categories for type of knowledge required are the following:

1.1 Answer stated directly in text

1.2 Answer stated indirectly in text

2.1 Answer stated indirectly, inferable from discourse

2.2 Answer not stated, inferable from world knowledge

2.3 Answer not stated, not unequivocally inferable

Table 5 shows the results for all six computer systems for these five categories.· The unit of analysis remains the four groups specified by grade level equivalence for reading.

Not all systems included all question types. The two question types appearing in the largest number of systems, 2.1 (all systems) and 1.2 (five systems), are discussed first to provide perspective on how this type of information can be used as part of an evaluation of intelligent computer systems.

The system response for 2.1 questions for four of the six systems—PAM, JULIP, SAM, Kind Types—was slightly higher than the mean for Group 4, placing the benchmark for this question type at or above grade equivalent 13 for the four systems. PAM answered 7 out of a possible 8 questions correctly (88% correct) while the other three systems answered all of the 2.1 questions correctly. BORIS and OpEd, on the other hand, were lower in response accuracy. BORIS responded correctly to 75% of the 2.1 questions (15 out of 20) in the system while OpEd responded correctly to only 50% (2 out of 4). In both cases, the benchmark for the 2.1 question type falls between Groups 3 and 4.

As mentioned above, five of the six computer systems had questions classified as 1.2, answer stated indirectly in text. (PAM was the only system which did not have 1.2 type questions.) Of the five systems, four benchmarked at grade equivalent 13 or above. JULIP, OpEd, and SAM responded correctly to all 1.2 type questions; Kind Types responded correctly to 9 out of 10 (90% correct) but only benchmarked between Groups 2 and 3; and BORIS correctly answered 12 out of 14 (86% correct) which was still higher than the mean for Group 4 ($\overline{X}$=11.65).

Question type 2.2, answer not stated, inferable from world knowledge, is extremely interesting because the two systems that had this type of question benchmarked quite low. BORIS benchmarked between Groups 1 and 2 while Kind Types benchmarked below Group 1. It is probably not surprising that systems would have a difficult time with questions requiring world knowledge. Still, there were few questions per system for type 2.2 (BORIS had 4; Kind Types had 2), so the presumption of difficulty that emerges here is at best tentative.

Table 5

Descriptive Statistics for Question Type (Knowledge Required) by Grade Equivalency Groups on the Reading Comprehension Test

	1.1 Answer Stated Directly in Text				1.2 Answer Stated Indirectly in Text				2.1 Answer Stated Indirectly, Inferable From Discourse				2.2 Answer Not Stated, Inferable From World Knowledge				2.3 Answer Not Stated, Not Unequivocally Inferable			
PAM	Max Possible=0				Max Possible=0				Max Possible=8				Max Possible=0				Max Possible=0			
Groups	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
Group 1									4.97	1.77	0-8	35								
Group 2									5.51	1.54	3-8	80								
Group 3									5.90	1.54	2-8	90								
Group 4									6.76	1.38	2-8	76								
PAM									7.00			1								
JULIP	Max Possible=0				Max Possible=2				Max Possible=4				Max Possible=0				Max Possible=0			
Group 1					1.31	0.87	0-2	35	1.06	1.26	0-4	35								
Group 2					1.30	0.76	0-2	79	0.81	1.07	0-4	80								
Group 3					1.30	0.81	0-2	90	1.27	1.29	0-4	90								
Group 4					1.62	0.61	0-2	76	2.49	1.44	0-4	76								
JULIP					2.00			1	4.00			1								

Table 5 (continued)

	1.1				1.2				2.1				2.2				2.3			
	Answer Stated Directly in Text				Answer Stated Indirectly in Text				Answer Stated Indirectly, Inferable From Discourse				Answer Not Stated, Inferable From World Knowledge				Answer Not Stated, Not Unequivocally Inferable			
BORIS	Max Possible=0				Max Possible=14				Max Possible=20				Max Possible=4				Max Possible=10			
Groups	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
Group 1					10.59	1.84	4-14	34	12.41	3.59	6-19	34	2.94	1	1-4	35	5.91	1.62	2-9	34
Group 2					11.25	1.51	6-14	80	14.56	2.39	9-19	80	3.14	1.03	0-4	80	6.78	1.86	2-10	80
Group 3					11.5	1.54	7-14	90	14.80	2.55	6-19	90	3.33	0.96	0-4	90	7.22	1.53	2-10	90
Group 4					11.65	1.46	8-14	75	15.84	2.64	7-19	75	3.59	0.73	1-4	76	7.32	1.37	4-10	76
BORIS					12.00			1	15.00			1	3.00			1	8.00			1
OpEd	Max Possible=0				Max Possible=2				Max Possible=4				Max Possible=0				Max Possible=0			
Group 1					0.74	0.90	0-2	34	0.35	0.65	0-2	34								
Group 2					0.88	0.85	0-2	80	0.74	0.96	0-4	80								
Group 3					1.02	0.87	0-2	90	1.30	1.10	0-3	90								
Group 4					1.53	0.69	0-2	74	2.37	0.97	0-4	74								
OpEd					2.00			1	2.00			1								

Table 5 (continued)

	1.1				1.2				2.1				2.2				2.3			
	Answer Stated Directly in Text				Answer Stated Indirectly in Text				Answer Stated Indirectly, Inferable From Discourse				Answer Not Stated, Inferable From World Knowledge				Answer Not Stated, Not Unequivocally Inferable			
SAM	Max Possible=2				Max Possible=2				Max Possible=2				Max Possible=0				Max Possible=0			
Groups	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
Group 1	1.71	0.72	0-2	34	1.77	0.65	0-2	34	1.09	1.00	0-2	34								
Group 2	1.86	0.49	0-2	80	1.90	0.44	0-2	80	1.45	0.84	0-2	80								
Group 3	1.87	0.48	0-2	90	1.91	0.41	0-2	90	1.49	0.81	0-2	90								
Group 4	1.97	0.23	0-2	73	2.00	0.00	2-2	74	1.76	0.62	0-2	74								
SAM	2.00			1	2.00			1	2.00			1								
Kind Types	Max Possible=2				Max Possible=10				Max Possible=8				Max Possible=2				Max Possible=0			
Group 1	2.00	0.00	2-2	32	8.53	1.48	5-10	30	5.83	2.17	0.8	29	1.45	0.87	0-2	29				
Group 2	2.00	0.00	2-2	79	8.8	1.43	3-10	78	6.84	1.68	0-8	77	1.62	0.69	0-2	77				
Group 3	2.00	0.00	2-2	90	9.37	1.04	6-10	87	7.38	1.14	2-8	87	1.78	0.56	0-2	87				
Group 4	2.00	0.00	2-2	73	9.59	0.7	7-10	73	7.55	1.08	2-8	73	1.97	0.23	0-2	73				
Kind Types	2.00			1	9.00			1	8.00			1	1.00			1				

Linguistic/Textual Domains

We also investigated how the six systems performed on questions involving each of the four linguistic/textual domains identified in our scheme of question classification. The results are provided in Table 6. The four linguistic/textual domains are:

A: Intra-clausal, WH-questions

B: Clausal, yes-no questions

C: Inter-clausal

D: Discourse, multi-clausal.

As with the type of knowledge classification, not all systems included all linguistic/textual domain types. All systems that had an overall system benchmark above Group 4 (see Tabl 4) benchmarked above Group 4 for all of the linguistic/textual domain question types that occurred in that system. For example, JULIP, which had an overall system benchmark above Group 4, benchmarked above Group 4 for linguistic/textual domains A, C, and D. Questions with linguistic/textual domain B did not occur in JULIP.

The two systems that had an overall system benchmark at or below Group 4—BORIS and Kind Types—showed variability in their benchmark levels for the different linguistic/textual domains. BORIS benchmarked above Group 4, higher than its overall system benchmark, for linguistic/textual domains B, C, and D, but between Groups 1 and 2, well below its overall system benchmark, for domain A. Kind Types benchmarked at its overall system benchmark level, between Groups 2 and 3, for linguistic/textual domains A and C, but had an indeterminate benchmark for linguistic/textual domain B (scores for the four groups and the system are essentially the same). There were no questions with linguistic/textual domain D in the Kind Types text.

It is possible to use the system benchmarks for individual question types (type of knowledge or linguistic/textual domain) to describe system performance more specifically than is possible using only overall system benchmarks. For example, although BORIS has an overall system benchmark that is lower than JULIP's (between Groups 3 and 4 vs. above Group 4), BORIS's performance on type C questions benchmarks at the same level as JULIP's performance on type C questions (above Group 4). To take

Table 6

Descriptive Statistics for Question Type (Linguistic/Textual) by Grade Equivalency Groups on the Reading Comprehension Test

	A Intra Clausal WH-Question				B Clausal Yes-No Question				C Interclausal				D Discourse Multi-Clausal			
PAM	Max Possible=0				Max Possible=0				Max Possible=8				Max Possible=0			
Groups	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
Group 1									4.97	1.77	0-8	35				
Group 2									5.51	1.54	3-8	80				
Group 3									5.90	1.54	2-8	90				
Group 4									6.76	1.38	2-8	76				
PAM									7.00			1				
JULIP	Max Possible=2				Max Possible=0				Max Possible=2				Max Possible=2			
Group 1	1.31	0.87	0-2	35					0.26	0.66	0-2	35	0.80	0.99	0-2	35
Group 2	1.30	0.76	0-2	79					0.21	0.59	0-2	80	0.60	0.91	0-2	80
Group 3	1.30	0.81	0-2	90					0.29	0.67	0-2	90	0.98	0.99	0-2	90
Group 4	1.62	0.61	0-2	76					0.93	0.97	0-2	76	1.55	0.84	0-2	76
JULIP	2.00			1					2.00			1	2.00			1

Table 6 (continued)

	A Intra Clausal WH-Question				B Clausal Yes-No Question				C Interclausal				D Discourse Multi-Clausal			
BORIS	Max Possible=22				Max Possible=8				Max Possible=16				Max Possible=2			
Groups	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
Group 1	13.79	2.78	9-18	34	7.26	1.09	4-8	35	9.79	3.24	3-14	34	1.43	0.74	0-2	35
Group 2	15.50	2.52	9-20	80	7.44	1.07	4-8	80	11.20	2.34	4-15	80	1.59	0.57	0-2	80
Group 3	15.99	2.30	10-22	90	7.51	1.13	4-8	90	11.77	2.18	5-15	90	1.59	0.60	0-2	90
Group 4	16.67	2.31	11-21	75	7.62	0.80	4-8	76	12.45	2.13	7-16	76	1.62	0.59	0-2	76
BORIS	15.00			1	8.00			1	13.00			1	2.00			1
OpEd	Max Possible=0				Max Possible=0				Max Possible=6				Max Possible=0			
Group 1									1.09	1.22	0-4	34				
Group 2									1.61	1.46	0-6	80				
Group 3									2.32	1.61	0-5	9				
Group 4									3.89	1.35	0-6	74				
OpEd									4.00			1				

Table 6 (continued)

	A Intra Clausal WH-Question				B Clausal Yes-No Question				C Interclausal				D Discourse Multi-Clausal			
SAM	Max Possible=2				Max Possible=2				Max Possible=2				Max Possible=0			
Groups	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
Group 1	1.71	0.72	0-2	34	1.77	0.65	0-2	34	1.09	1.00	0-2	34				
Group 2	1.86	0.50	0-2	80	1.90	0.44	0-2	80	1.45	0.84	0-2	80				
Group 3	1.87	0.48	0-2	90	1.91	0.41	0-2	90	1.49	0.81	0-2	90				
Group 4	1.97	0.23	0-2	73	2.00	0.00	2-2	74	1.76	0.62	0-2	74				
SAM	2.00			1	2.00			1	2.00			1				
Kind Types	Max Possible=10				Max Possible=4				Max Possible=8				Max Possible=0			
Group 1	7.80	1.96	3-10	30	4.00	0.00	4-4	32	5.97	1.88	0-8	29				
Group 2	8.64	1.69	2-10	78	3.96	0.25	2-4	79	6.68	1.41	1-8	77				
Group 3	9.23	1.33	4-10	87	4.00	0.00	4-4	89	7.30	1.06	4-8	87				
Group 4	9.49	0.87	6-10	73	4.00	0.00	4-4	73	7.62	0.88	4-8	73				
Kind Types	9.00			1	4.00			1	7.00			1				

another example, although BORIS's overall system benchmark (between Groups 3 and 4) is higher than Kind Types's (between Groups 2 and 3), BORIS's performance on type A questions (between Groups 1 and 2) is lower than Kind Types's performance on type A questions (between Groups 2 and 3); it is in fact lower than Kind Types's performance on any type of question.

BORIS's performance on type A questions lowers its overall system benchmark below Group 4. Examples of type A questions from BORIS are the following:

8. Who is Paul?

23. How did Paul feel when Richard called?

It is clear that scores on these question reflect not only the linguistic/textual domain involved, but also the type of knowledge required to answer the question. For example, the knowledge required to answer question 23 correctly is not stated directly in the text; it is inferable from the discourse (classification 2.1).

BORIS's performance on type A questions is independent of the type of knowledge required to answer these questions. In general, the scores on these two scales (type of knowledge and linguistic/textual domain) are independent, but a specific system may manifest relations between certain types of knowledge and certain linguistic/textual domains, relations that are particular to that system. Such relations would give information about the system and the problems it has in text understanding. For example, if all the type A questions in a system's texts were 2.1 and 2.3, and the system benchmarked lower than its overall value for type A but not for types 2.1 and 2.3, this may indicate that the system has problems with question types 2.1 and 2.3 that are posed as intra-clausal questions (type A). Such relations might be observed in an expanded study which included texts with a good spread of question types and involved a large sample of human subjects. We did not observe any such relations in our preliminary study.[7]

Analysis of system benchmarks could be further clarified if the influence of text difficulty were factored out. For example, two systems may show different benchmarks for question type 2.1, but these benchmarks may reflect

[7] For an extended analysis focusing on a single text understanding system see Falk and Herl, 1990.

more than the fact that the questions under consideration are 2.1. The benchmarks may reflect overall difficulty of the questions and texts, among other things, as well. Suppose System A benchmarks at Group 2 for 2.1 questions and System B benchmarks at Group 3. Suppose that when the effects of text difficulty are normalized, both systems benchmark at Group 3. These latter benchmarks would be more accurate measures of the specific ability of the systems to answer 2.1 type questions. Although we did determine levels of difficulty for the various systems (see Appendix D), we were unable to factor text difficulty out of the results due to sample size.

Summary and Conclusions

The goal of this study has been to suggest a procedure for referencing the performance of intelligent computer systems, specifically text understanding systems, to the performance of humans on the same task. The guiding question has been "Can we compare machine output to human performance?" Or perhaps more accurately, "Can we compare machine output of a specific kind to human performance of the same kind, and if so, how?" Our underlying assumption has been that performance-based measures can be used to reference text understanding systems.

Our approach has involved a benchmarking methodology which allows us to evaluate the natural language processing abilities of intelligent computer systems. The scale to which we have benchmarked the systems' performances is a scale of human performance on reading comprehension. We have outlined methods for determining overall system benchmarks and also methods for investigating differential system abilities, that is, benchmark levels for questions of different types. Due to the relatively small and clustered subject sample we had access to, we were not able to benchmark on a continuous scale nor were we able to determine the statistical significance of many of our results. Nevertheless, the general descriptive results presented here indicate that a human benchmark methodology can distinguish certain kinds of natural language processing abilities of intelligent computer systems. Given a larger and more diverse subject sample, it would be possible to rigorously substantiate benchmark values.

If the benchmarking methodology described here were to be developed further, it would be useful, in addition to having a larger and more diverse subject sample serving as the benchmark population, to address some of the factors known to affect the reading comprehension of human subjects. Such a refinement of the benchmarking methodology would indicate whether these factors influence the ability of intelligent computer systems, as well as the ability of human subjects, to process texts.

Since the discourse knowledge and world knowledge required to answer a question differ from question to question, and since different cognitive processes are required for different questions, our analysis of the questions asked is one step towards determining and evaluating factors that influence reading comprehension for humans and computers.

Another factor influencing reading comprehension is text difficulty. There seemed to be a positive correlation between age and the ability to read one text passage (passage 4 in Appendix B) which contained the most difficult vocabulary and concepts. In our results section we considered how text difficulty values could be employed in analysis.

Vocabulary in the questions also affected performance. For example, two questions asked for the consequences of the characters' actions. Many students in the 6th and 8th grades did not understand the meaning of the word "consequences" and hence could not answer the question even though they may have comprehended the passage.

Although we did consider whether world knowledge was necessary to answer a question, it is clear that investigation of world knowledge accessed in answering a question is an area which could be explored further. Human subjects on occasion bring in world knowledge to supplement, alter, or override the literal text answer. For instance, one passage describes a man who almost hits someone while driving to a restaurant (see passage 3 in Appendix B). Shaken by the incident, the man has three drinks. One question asks "why did Richard get drunk?" The answer inferable from the text is that Richard got drunk because he was upset about almost hitting someone. Some human subjects gave this answer while others answered that he got drunk because he had three drinks, or because he was an alcoholic. Subjects who answered "because he had three drinks" may be drawing on experience which informs them that being upset is not a cause for getting drunk. Subjects who answered "because he is an alcoholic" are drawing on world knowledge of types of people who get drunk. The subjects' world knowledge about why people get drunk clearly affects the way they answer the question. Depending on the question and how it is scored, resorting to world knowledge can potentially raise or lower the score on a test item. Although a computer system could be devised to mimic this human behavior, it was not observed in the systems we used.

Attitudes and fears derived from world knowledge or experience can also affect how a person answers a question. For example, many subjects answered the question "What will happen as computers eliminate jobs?" in passage 2 (Appendix B) by saying that computers are going to take over everyone's jobs. The answer inferable from the text is that more jobs will be created. It would seem that an attitude or fear associated with computer aided manufacturing is influencing the subjects' ability to answer this question.

In addition to text difficulty and world knowledge, factors such as knowledge structures in a text affect the reading comprehension of human subjects. A knowledge structure is an abstract framework of conceptual relations. The use of analogy in passage 2 (see Appendix B) is an example of a knowledge structure which posed problems for many subjects in the 6th and 8th grades.

Unlike computers, humans make careless errors and exhibit test fatigue. In one student's answer to question 2 (see Appendix B), "Because John said he would break John's are," the two errors—"John" for "Bill" and "are" for "arm"—are probably due to careless lapses. Computer systems tend not to make errors of this sort. Humans also exhibit test fatigue: Some students gave humorous and silly answers towards the end of the test. Careless errors and test fatigue lower the average score for a group of students, even adults, but they do not affect computer scores.

We have discussed several factors which influence human text readers. What accounts for the computational systems' incorrect and partially-correct answers? As Table 4 shows, of the six computational systems, only two of them (JULIP and SAM) earned full credit on all questions when compared with the "correct" answers (determined here as the response given by a majority of the adult sample). One possible explanation for less-than-correct system answers is that the programmers knew that an answer was wrong, but were unable to program the system to produce the correct one. However, this explanation is unlikely for the texts used here, because the texts and questions were selected by the programmers to demonstrate the abilities of the systems. A more likely explanation here is that the programmers considered the system's answer to be correct. This study, then, suggests the incidental observation that, if human-like performance is the goal of a question-answering system, the programmers would do well to give attention to the answers humans actually give, rather than to rely wholly on introspection for identifying the "right" answer.

For benchmarking computational system capability to student capability, when a system provides a less-than-correct answer, it would be helpful to know whether this indicates a minor programming oversight or a major lack of system capability. Such questions were beyond the scope of this study; we considered system output only and did not try to assess system design or approach.

Another possible direction for future study would involve controlling the type of text and questions asked of each system. Our study was limited to the reported performance of each system. Each system read a different text and answered different questions. The texts differed in sentence length, word length, syntactic complexity, and discourse type, ranging from simple narrative to editorial argument. Relative benchmarking of individual systems based on number of correct answers tends to overlook the fact that, for a given system design, some types of questions will be easier than others to answer. A more revealing relative measure of system capability could be carried out if it were possible to use the same text for all systems and a full range of question types, comparing system performance to human performance.

To place our efforts in context, it is important to realize that the work is exploratory and was conceived as part of a larger plan to develop a methodology for assessing a range of intelligent computer systems. The results of this study complement the findings of an earlier study using a similar technique to benchmark a NL query system (Baker et al., 1990). Current research with an expert system has also involved the use of a benchmarking methodology as an evaluation procedure (O'Neil, Ni, & Jacoby, 1990; O'Neil, Ni, Jacoby, & Swigger, 1990). This growing body of work which focuses on a benchmark approach to the assessment of intelligent computer systems seems to suggest that we can compare system performance to human performance in a meaningful way using performance-based measures. Clearly the approach needs fine tuning, but this study as well as the others mentioned above provide direction for researchers who are interested in a methodology for assessing intelligent computer systems.

References

Alvarado, S.J. (1990). *Understanding editorial text: A computer model of argument comprehension.* Boston: Kluwer.

August, S., & Dyer, M. (1985). Understanding analogies in editorials. In *Proceedings of the 9th International Joint Conference on Artificial Intelligence* (pp. 845-847), Los Angeles, August 18-23.

August, S.E., & Dyer, M.G. (1986). *Analogy recognition and comprehension in editorials* (CSD 860073). Los Angeles: University of California, Computer Science Department.

Baker, E.L., & Lindheim, E.L. (1988, April). *A contrast between computer and human language understanding.* Paper presented at the annual meeting of the American Educational Research Association, New Orleans.

Baker, E.L., Turner, J.L., & Butler, F.A. (1990). *An initial inquiry into the use of human performance to evaluate artificial intelligence systems.* Los Angeles: University of California, Center for Technology Assessment/ Center for the Study of Evaluation.

Carbonell, J.G. (1979). *Subjective understanding: Computer models of belief systems* (Ph.D. dissertation; Research Report #150). New Haven, CT: Yale University, Computer Science Department.

Carroll, L. (1928). *Alice's adventures in wonderland; Through the looking glass; The hunting of the snark.* New York: Random House.

Comprehensive Test of Basic Skills, Form U. (1984). Del Monte Research Park, Monterey, CA: CTB/McGraw-Hill, Inc.

Comprehensive Test of Basic Skills, Norms Book, Forms U and V. (1983). Del Monte Research Park, Monterey, CA: CTB/McGraw-Hill, Inc.

Cullingford, R.E. (1978). *Script application: Computer understanding of newspaper stories* (Ph.D. dissertation, Research Report #116). New Haven, CT: Yale University, Computer Science Department.

Dahlgren, K. (1988). *Naive semantics for natural language understanding.* Boston: Kluwer.

Davison, A., & Kantor, R.N. (1982). On the failure of readability formulas to define readable texts: A case study from adaptations. *Reading Research Quarterly, 17,* 187-209.

DeJong, J. (1979). *Skimming stories in real time.* Unpublished doctoral dissertation, Yale University, New Haven.

Dyer, M. (1983). *In depth understanding: A computer model of integrated processing for narrative comprehension.* Cambridge, MA: MIT press.

Falk, T., & Herl, H. (1990, October). *Working paper: Benchmarking text understanding systems to human performance: An extended analysis using the BORIS system.* Los Angeles: University of California, Center for Technology Assessment/Center for the Study of Evaluation.

Flesch, R. (1949). *The art of readable writing.* New York: Harper & Bros. Pub.

Hieronymus, A.N., Hoover, H.D., Lindquist, E.F., & others. (1986). *Iowa Test of Basic Skills, Forms G/H, Teacher's Guide.* Chicago: The Riverside Publishing Co.

Hull, L.C. (1979, May). Measuring the readability of technical writing. In *Proceedings of the Twenty-sixth International Technical Communication Conference* (pp. E-73 – E-78), Los Angeles, CA.

Jacoby, A. (1989). *Natural language understanding systems. Working paper.* Los Angeles: University of California, Center for Technology Assessment/ Center for the Study of Evaluation.

Klare, G.R. (1984). Readability. In P.D. Pearson (Ed.), *Handbook of reading research.* New York: Longman, Inc.

Kolodner, J.L. (1980). *Retrieval and organizational strategies in conceptual memory: A computer model* (Tech. Rep. #187). New Haven, CT: Yale University, Computer Science Department.

Lebowitz, M. (1980). *Generation and memory in an integrated understanding system* (Research Rep. #186). New Haven, CT: Yale University, Computer Science Department.

Long, D. (1989). *Sensible grammar* (Version 1.5.7) [Computer program]. Sensible Software Inc.

Manzo, A. (1970). Readability: A postscript. *Elementary English, 47,* 962-965.

Oakhill, J., & Garnham, A. (1988). *Becoming a skilled reader.* Oxford, NY: Basil Blackwell.

O'Neil, H.F., Jr., Ni, Y., & Jacoby, A. (1990). *Literature review: Human benchmarking of expert systems.* Los Angeles: University of Southern California, Cognitive Science Laboratory, and University of California, Center for Technology Assessment/Center for the Study of Evaluation.

O'Neil, H.F., Jr., Ni, Y., Jacoby, A., & Swigger, K. (1990). *Human benchmarking methodology for expert systems.* Los Angeles: University of Southern California, Cognitive Science Laboratory; University of California, Center for Technology Assessment/Center for the Study of

Evaluation; and University of North Texas, Department of Computer Sciences.

Reeves, J.F., & Dyer, M.G. (1986). *Recognition and representing situational ironies* (CSD 860075). Los Angeles: University of California, Computer Science Department.

Walmsley, S.A., Scott, K.M., & Lehrer, R. (1981). Effects of document simplification on the reading comprehension of the elderly. *Journal of Reading Behavior, 13*, 237-248.

Wilensky, R. (1978). *Understanding goal-based stories* (Ph.D. dissertation, Research Rep. #140). New Haven, CT: Yale University, Computer Science Department.

Williams, A.R., Siegel, A.I., Burkett, J.R., & Groff, S.D. (1977). *Development and evaluation of an equation for predicting the comprehensibility of textual material* (AFHRL-TR-77-8). Brooks AFB, TX: Air Force Human Resources Laboratory, Air Force Systems Command.

Appendix A

Natural Language Processing—Text Understanding Systems

The eleven systems considered were the following:

SAM - Script Applier Mechanism, Yale University

PAM - Plan Applier Mechanism, Yale University

POLITICS - Yale University

FRUMP - Fast Reading Understanding and Memory System, Yale University

IPP - Integrated Partial Parser, Yale University

BORIS - Better Organized Reasoning and Inference System, Yale University

CYRUS - Computerized Yale Reasoning and Understanding System, Yale University

JULIP - AI Lab, UCLA

Reeves - AI Lab, UCLA

OpEd - Opinions to/from the Editor, AI Lab, UCLA

Kind Types System - NewSelector Project, IBM Los Angeles Scientific Center

The six selected systems were: SAM, PAM, BORIS, JULIP, OpEd, and the Kind Types System.

Appendix B

Reading Comprehension Test With Answer Key

Answers by thirteen adult subjects who took the test are provided for each question below. The number in the parentheses before each answer indicates the number of subjects who gave the same kind of answer (if the number is 0, this was an answer we discussed and added for guidance). The AI answers to the questions are provided first in italics. The **bold number** is the number of points on a 0-2 scale awarded to the answer. Answers receiving 1 point are followed by a brief explanation in **bold print**. Some questions include **Guidelines** for answers.

General Guidelines: If a two part answer is given with one part correct, the answer receives 2 points if the second part is a plausible inference, 1-0 points if not a plausible inference. Facetious answers that show understanding should receive credit.

(1)

John wanted Bill's bicycle. He went over to Bill and asked him if he would give it to him. Bill refused. John told Bill he would give him five dollars for it, but Bill would not agree. Then John told Bill he would break his arm if he didn't let him have it. Bill gave him the bicycle.

1. Why did John walk over to Bill?

Because he wanted to get his bicycle.

2 (6) To ask him if he would give him his bicycle.
2 (4) Because he wanted Bill's bicycle.
2 (3) To get Bill's bicycle.
1 (0) To borrow Bill's bicycle. **(not intended inference)**
1 (0) To buy Bill's bike. **(not intended inference)**
1 (0) To talk. **(not specific enough)**

2. Why did Bill give his bicycle to John?

Because he didn't want to get hurt.

2 (5) Because John threatened him.
2 (3) John told Bill he was going to break his arm if he didn't give him the bicycle.
2 (2) Bill was afraid of getting his arm broken if he didn't give it to him.
2 (2) Bill was afraid of getting his arm broken.
2 (1) Because John threatened that he would break his arm if he did not succumb.
2 (1) He was intimidated.

3. What were the consequences of John's walking over to Bill?

Guidelines: Answer needs to be an "action" and in the past tense with the notion of "transfer of bike" present for 2 pts. If answer is a conditional "would" or "will," receives 1 pt. If answer is a conditional "could" or "might," receives 0 pts. If answer conveys "purpose" as opposed to "consequences," receives 0 pts.

This enabled him to ask him to give him Bill's bicycle.
2 (9) John got Bill's bicycle.
2 (1) John threatened bill, Bill gave his bicycle to him.
2 (1) The bicycle was transfered from Bill to John.
1 (1) Bill refused to give up his bike to John. **(incomplete)**
1 (0) Asked if he would give him his bike.
1 (0) They argued.
1 (0) John would ask for the bike.
1 (0) Bill would lose the bike.
0 (1) ? No consequences were due to John's walking over to Bill. They were due to his question, offer, and threat.
0 (0) Bill might lose the bike.

4. What were the consequences of John's asking Bill to give him Bill's bicycle?

Guidelines: Bill's "refusal" must be present for 2 pts. If answer is a conditional "would" or "will" and conveys Bills "refusal," receives 1 pt. If answer is a conditional "could" or "might," receives 0 pts. If answer conveys "purpose" as opposed to "consequences," receives 0 pts.

Bill told him that Bill wouldn't give him Bill's bicycle.
2 (3) Bill refused.
2 (2) Bill said he wouldn't give John the bicycle.
2 (2) Bill refused, so he offered five dollars, Bill still refused, so he threatened Bill and he got the bike.
2 (1) Bill refused upon first request.
2 (1) John was initially refused the bicycle but later received it.
2 (1) An unpleasant exchange between John and Bill and the transfer of the bike.
2 (1) John's request was refused by Bill.
1 (0) He didn't get the bike. **(incomplete-need refusal first)**
1 (0) No. **(not specific enough)**
1 (1) He finally got his bike. **(not consequence of John's asking, but final outcome)**
0 (1) He didn't get the bicycle after the first try. He had to ask again.
0 (1) Bill gave John his bicycle.
1 (0) That Bill would say no.
1 (0) John and Bill were in an argument

(2)

Some people are against computer aided manufacturing (CAM) because CAM eliminates jobs. However the automobile industry did the same thing to people in the horse carriage industry. Yet consumer demand for autos was strong enough that more jobs were created in the automobile industry than jobs were lost in the horse carriage industry. In the end the economy benefitted by the introduction of the new technology.

5. To what is the computer industry being compared?

The computer industry is being compared to the automobile industry
2 (11) To the auto industry.
1 (0) Auto.
0 (1) The horse carriage industry.
0 (1) The computer industry is not mentioned.
0 (0) Auto and Horse Carriage Industry.

6. What did the auto industry do to people in the horse carriage industry?

Guidelines: Answer must refer to "people" in the horse carriage industry for 2 pts. Idea of "losing jobs" is the key point for 2 pts. Other plausible inferences receive 1 pt.

People in the horse carriage industry lost jobs.
2 (7) Put them out of jobs (or work).
2 (3) It made them lose their jobs.
2 (1) The auto industry reduced jobs in the horse carriage industry.
2 (0) Jobs were lost.
1 (1) Replaced their jobs with auto industry jobs. **(leap inference)**
1 (0) Put them out of business. **(broad inference)**
1 (0) More jobs were created. **(incomplete and missing loss of jobs)**
0 (1) The auto industry made the horse carriage industry obsolete.

7. What will happen as computers eliminate jobs?

Guidelines: For 2 pts, the key idea that as computers eliminate some jobs, they create others, must be present.

An even greater number of new jobs will be created.
2 (10) More jobs will be created.
2 (1) The economy will benefit by the introduction of the new technology.
2 (1) Who knows? If the economy parallels that of the auto industry, eventually the economy will be benefited.
0 (1) Not clear that computers do eliminates jobs.
0 (0) People will lose jobs.
0 (0) People won't have any jobs.

(3)

Richard hadn't heard from his college roommate Paul for years. Richard had borrowed money from Paul which was never paid back. But now he had no idea where to find his old friend. When a letter finally arrived from San Francisco, Richard was anxious to find out how Paul was.

8. What happened to Richard at home?

Guidelines: There is nothing in the passage to indicate Richard is at home. Further, a letter arriving is not necessarily something that happens to a person. Therefore, "received a letter" is the correct answer, but a qualified* "I don't know" is also correct.

Richard got a letter from Paul.
2 (4) He received a letter from Paul.
2 (1) I don't know. They don't say.
2 (1) I don't know. Is Richard at home?
2 (1) There is no mention of Richard at home. Only at college.
2 (1) It doesn't say anything to indicate when Richard was at home.
2 (1) I didn't understand this question. Maybe Richard received Paul's letter at his home.
2 (1) A letter from Paul arrived?
2 (1) Can't answer. I don't know if the letter arrived at his home or another place.
1 (2) I don't know. **(not specific enough)**
1 (0) Nothing happened. **(reasonable, but not correct)**
0 (1) He lost touch with his college roommate Paul.

9. Who is Paul?

Guidelines: Must refer to "roommate" as a past roommate situation.

Richard's friend.
2 (11) Richard's college roommate.
2 (2) Richard's friend.
1 (0) Richard's roommate. **(present tense roommate)**

10. Did Richard want to see Paul?

Guidelines: Richard may not necessarily want to "see" Paul. However, due to the content of the passage, an answer of "no" is incorrect. A qualified "I don't know" is correct.

* A "qualified I don't know" means the reason for not knowing is stated, and that reason has to do with the answer not stated in or not reasonably inferred from the text.

Yes, Richard wanted to know how Paul was.

2 (4) Yes.

2 (2) Maybe (maybe not). We don't know from the story.

2 (2) Not sure, but Richard seemed anxious to know how Paul was doing.

2 (1) Don't know, (but he did want to return the money and find out how his friend was.

2 (1) Richard mainly wanted to hear from Paul. Seeing him may or may not have been desired.

2 (1) I assume so since they are "old friends" and Richard might want to pay Paul back.

1 (1) I don't know. **(not specific enough)**

1 (1) No, he wanted to find out how Paul was. **("no" is incorrect)**

0 (0) No.

11. Had Paul helped Richard?

Yes, Paul lent money to Richard.

2 (7) Yes.

2 (2) Yes. He lent Richard money.

2 (1) In the sense that he loaned him money.

0 (2) No.

0 (1) I don't know. This information is not contained in the story.

12. Why didn't Richard pay Paul back?

Guidelines: Any inference other than "lost touch," e.g. "He was broke," is 0 pts.

Richard did not know where Paul was.

2 (4) I don't know. Not stated in passage.

2 (2) Insufficient information from the passage (or can't tell from the passage).

1 (3) I don't know. **(not specific enough)**

1 (3) They lost touch. **(leap inference)**

1 (1) He couldn't find him. **(leap inference)**

0 (0) He moved away.

13. How did Richard feel when the letter appeared?

Guidelines: Pulling "anxious" from the passage without "to find out how Paul was," is incorrect. Inferencing a more positive emotion is more correct.

Richard felt glad because Paul and he were friends.

2 (2) Anxious to find out how Paul was.

2 (0) He wanted to know (to find out) how Paul was.

1 (1) Anxious? Relieved?

1 (1) Curious and anxious.

1 (1) Happy.

1 (1) He was happy. He wanted to know how Paul was.

1 (1) Relieved that he could find out about Paul.

1 (1) Excited.

1 (0) Anxious to find Paul

0 (5) Anxious.

Unfortunately, the news was not good. Paul's wife Sarah wanted a divorce. She also wanted the car, the house, the children, and alimony. Paul wanted the divorce, but he didn't want to see Sarah walk off with everything he had. His salary from the state school system was very small. Not knowing who to turn to, he was hoping for a favor from the only lawyer he knew. Paul gave his home phone number in case Richard felt he could help.

14. What was the letter about?

Guidelines: Answer must refer to "divorce" or marital situation for 2. Must refer to Paul and/or Sarah as the people wanting a divorce.

Paul and Sarah were getting a divorce.
2 (3) Paul's wife wanted a divorce (or Paul's divorce).
2 (2) His troubles of life and marriage (or with his wife Sarah).
2 (1) Paul's divorce and problems.
2 (1) Paul and his wife getting a divorce and Paul needing legal help.
2 (0) Paul's problem with his wife.
1 (2) Paul was in trouble and wanted to know if Richard could help him. **(not specific enough)**
1 (1) The events that had happened to Paul in the recent past and an attempt to get help from Richard. **(not specific enough)**
1 (1) Paul's problems. **(not specific enough)**
1 (1) Paul's situation. **(not specific enough)**
1 (0) How everybody wanted a divorce. **(do not refer to Paul or Sarah)**
1 (0) Paul wanted Richard to be his lawyer. **(leap inference)**
0 (1) There is no mention of a letter.

15. Did Paul write Sarah?

Guidelines: This question is not directly stated in the passage, but the inferred answer is "no". "Not stated in text" answers receive 2 pts. "I don't know" without stating why, receives 1 pt.

No, it wasn't Sarah. It was Richard who got the letter.
2 (2) Not stated (or it doesn't say in the story).
2 (4) No.
2 (1) Insufficient information from the passage.
2 (1) Probably not.
2 (0) No, Paul wrote Richard.
1 (3) I don't know. **(not specific enough)**
1 (1) Maybe. **(not specific enough)**
1 (1) Paul wrote Richard. **(doesn't directly answer question)**

16. Why was Paul upset about the divorce?

Guidelines: Answers must refer to Sarah "wanting" everything, as opposed to "taking" or "getting" everything (since the divorce has not yet occurred). "Taking" answers get 1 pt.

Paul and Sarah were fighting over the family possessions.

2 (5) Because he was afraid Sarah was going to walk off with everything he had.

2 (2) Sarah wanted too much.

2 (1) He didn't want that bitch to get everything.

2 (1) Because of the things that Sarah wanted with it.

2 (1) His wife wanted all of their money, the house, and the children.

1 (1) For monetary reasons.

1 (1) Because he would have to give Sarah lots of money, and the car, house, and kids. **("would" should be substituted with "might")**

1 (1) He anticipated becoming impoverished. **(incomplete-also anticipates losing children)**

1 (0) Sarah was taking everything.

0 (0) Sarah got or took everything.

17. What did Paul do for a living?

Guidelines: It is not stated what Paul specifically does, so inferencing a "teacher" is not necessarily correct. However, it is more correct than "mechanic." Answering with a specific position such as teacher or janitor as probable is o.k. for 2 pts.

Paul was a teacher.

2 (10) State school system employee (He is probably a teacher).

2 (1) Probably a teacher or a principal or a secretary for a school. We don't know exactly.

1 (0) A teacher. **(insufficient)**

0 (2) Insufficient information from the passage (or can't tell from the passage).

18. What did Sarah want?

Guidelines: For 2 pts, "Divorce" must be included in answer along with most of the other things listed in the passage.

Sarah wanted to have the car and the house and the kids and the alimony.

2 (7) She wanted a divorce, the car, the house, the children, and alimony.

2 (1) A divorce as well as the house, the children and alimony.

1 (3) The house, car, kid and alimony. **(doesn't mention divorce)**

1 (1) Everything. **(not specific enough)**

1 (1) The car, the house, etc. **(doesn't mention divorce)**

19. Why did Paul write to Richard?

Guidelines: For 2 pts answer must include: writing for help specifically in "legal" advice, or for a "lawyer."

Paul wanted Richard to be his lawyer.

2 (4) He needed legal advice from Richard.

2 (2) He wanted to know if Richard could help him since Richard is a lawyer.

2 (1) He had no one to turn to and was hoping for a favor from his lawyer friend.

2 (1) Because Richard is a lawyer.

2 (1) He needed a lawyer.

1 (4) To ask if he would help him. **(not specific enough-need "lawyer," "legal")**

1 (0) To ask if Richard would be his lawyer. **(leap inference)**

Richard eagerly picked up the phone and dialed. After a brief conversation, Paul agreed to have lunch with him the next day. He sounded extremely relieved and grateful.

20. Why did Richard call Paul?

Guidelines: For 2 pts, the idea of Richard calling to <u>help</u> Paul must be present.

Richard wanted to arrange a meeting with Paul.

2 (4) He wanted to help (and maybe pay back the money).

2 (2) To talk to him? To offer to help him (with the divorce)?

2 (1) He was responding to Paul's situation.

2 (1) Probably to return the favor and get back in touch with his friend.

2 (0) He wanted Paul to talk about his problem.

2 (0) To return the favor.

2 (0) To talk to his friend about his problem.

1 (2) He was anxious to find out how Paul was. **(incomplete)**

1 (1) Perhaps he was worried and concerned by the bad news concerning his old pal. **(incomplete)**

1 (1) To meet with him. **(incomplete)**

1 (0) To talk.

1 (0) To pay Paul back.

0 (1) Insufficient information from the passage.

21. Did Richard contact Paul?

Yes. Richard called Paul.

2 (10) Yes.

2 (2) Yes, he called Paul.

1 (1) Yes, he tried, we don't know if it was busy and Paul called him but we presume. **(leap inference)**

22. How did Richard contact Paul?

By phone.
2 (8) By phone.
2 (4) He called him.
2 (1) He picked up the phone and dialed.

23. How did Paul feel when Richard called?

Guidelines: "Relieved and grateful" is most correct. Any plausible emotion (e.g., "happy") gets 1 pt.

Paul was happy because Richard agreed to be Paul's lawyer.
2 (11) (Extremely) relieved and grateful.
1 (1) Eager as well as relieved and grateful after they agreed to meet. **(Richard felt eager, not Paul)**
1 (0) Relieved.
1 (0) Grateful.
0 (1) Not clear from the text.

The next day, as Richard was driving into the restaurant, he barely avoided hitting an old man on the street. He felt extremely upset by the incident, and had three drinks at the restaurant.

When Paul arrived Richard was fairly drunk. After the food came, Richard spilled a cup of coffee on Paul. Paul seemed very annoyed by this so Richard offered to drive him home for a change of clothes.

24. Why did Richard eat out with Paul?

Guidelines: Answer must reflect idea of discussing, or helping with divorce case or problem for 2 pts.

Richard wanted to discuss the divorce case with Paul.
2 (4) So they could talk and see if he could help him.
2 (3) To discuss Paul's possible divorce (and legal problems).
1 (2) Paul agreed to have lunch with him. **(incomplete)**
1 (1) Richard asked him if he would eat with him. **(incomplete)**
1 (1) They had agreed to do so the day before. **(incomplete)**
1 (1) Because they had made plans to do this. **(incomplete)**
0 (1) Insufficient information from the passage.

25. What happened to Richard on the way to the restaurant?

Richard almost ran over an old man.

2 (10) He nearly (or almost) hit an old man on the street.
2 (2) He barely avoided an accident (or hitting an old man on the street).
2 (1) He narrowly missed plowing down an old man.

26. Why did Richard get drunk?

Guidelines: For 2 pts, the answer must explain "He was upset," with the incident regarding hitting the old man.

Richard was upset about almost running over the old man.

2 (9) He was upset about almost hitting an old man.
2 (1) Because he nearly hit the old man.
2 (1) He was shaken by the incident.
2 (0) He drank too much alcohol.
2 (0) He had 3 drinks.
2 (1) He was jittery over the near miss.
1 (1) Don't know - presume it was because he was upset. **(incomplete)**
1 (0) He was upset. **(incomplete) (not specific enough)**
1 (0) He drank alcohol. **(insufficient)**
0 (0) I don't know.

27. What happened at the restaurant?

Guidelines: For 2 pts must have both "Richard got drunk" and "Richard spilled coffee on Paul."

Richard spilled coffee on Paul.

2 (8) Richard got drunk and spilled coffee on Paul.
2 (3) Richard got drunk, Paul arrived, they ordered food, Richard spilled coffee on Paul, Richard offered to drive him home to change.
1 (2) Richard spilled coffee on Paul. **(incomplete)**

28. How did Richard feel when the coffee spilled?

Guidelines: Since answer is not stated, "not stated" is worth 2 pts. Any plausible emotion (e.g., bad, upset) gets 1 pt. Any plausible emotion preceded by "probably" or "maybe," etc., gets 2 pts.

Richard was unhappy.

2 (4) I don't know. Not stated.
2 (1) I don't know. Probably remorseful.
2 (1) I don't know. Drunk? Sorry?
1 (2) He was upset. **(leap inference)**
1 (1) He felt bad. **(leap inference)**
1 (1) Guilty. **(leap inference)**
1 (1) I don't know. **(not specific enough)**
0 (2) Annoyed. **(Paul felt annoyed)**

29. Why did Richard spill the coffee?

Guidelines: Although "drunk" is the most reasonable answer, it is not stated in the text, and other answers are plausible. So a qualified "I don't know" gets 1 pt.

Richard was drunk.
2 (10) He was drunk.
1 (3) I don't know. Not stated.
0 (0) I don't know.
0 (0) Accident or mistake.

30. When did Richard almost hit the old man?

While Richard was driving to the restaurant.
2 (6) While driving to the restaurant to meet Paul.
2 (5) On his way to the restaurant.
2 (1) On his way to meet his old roommate for lunch.
2 (1) As he was pulling into the restaurant.

31. Where did Richard have lunch with Paul?

At a restaurant.
2 (12) At a restaurant.
0 (1) Don't know.

(4)

The American machine-tool industry is seeking protection from foreign competition. The industry has been hurt by cheaper machine tools from Japan. The toolmakers argue that restrictions on imports must be imposed so that the industry can survive. It is a wrongheaded argument. Restrictions on imports would mean that American manufacturers would have to make do with more expensive American machine tools. Inevitably, those American manufacturers would produce more expensive products. They would lose sales. Then those manufacturers would demand protection against foreign competition.

32. What does the American machine-tool industry believe?

Guidelines: An answer receives 2 pts if it relates any part of the belief and includes the idea of "restrictions" or "protection against foreign competition." Other correct statements of belief will be awarded 1 pt.

The American machine tool industry believes that protectionist policy by the American government achieves the preservation of normal profits of the American machine tool industry.
2 (5) That there should be restriction on imports of machine tools.
2 (3) They need protection from foreign competition.
2 (1) They believe that import protection will save their industry.
2 (1) It should be protected from foreign competition by restricting imports.
2 (0) That the U.S. needs protection.

1 (2) That their industry has been hurt by cheaper machine tools from Japan. **(incomplete-restriction/protection not included)**
1 (1) They believe that their sales is declining due to cheaper foreign imports. **(incomplete-restriction/protection not included)**
1 (0) Japanese make cheaper machine tools.
1 (0) They are losing sales.

33. What does the author believe?

Guidelines: For 2 pts, the idea that (1) protectionist policies are the wrong approach must be conveyed **or** (2) the problems associated with protectionist policies must be conveyed. If answer in question 32 is correct and answer 33 states "opposite of" answer in 32, it is 2 pts.

The author believes that protectionist policy by the American government is bad because the author believes that protectionist policy by the American government motivates the preservationof normal profits of American industries. The author believes that the American Machine Tool Industry is wrong because the American Machine Tool Industry believes that protectionist policy by the American Government achieves the preservation of normal profits of the American Machine Tool Industry.

The author believes that protectionist policy by the American Government is bad because the author believes that protectionist policy by the American Government motivates the preservation of normal profits of American Industries; and the preservation of normal profits of American Industries intends persuasion plan by American Industries about protectionist policy by the American Government. The author believes that the American Machine Tool Industry is wrong because the American Machine Tool Industry believes that protectionist policy by the American Government achieves the preservation of normal profits of the American Machine Tool Industry.

2 (2) That if the process is begun then all manufacturers would begin to seek protection against foreign competition.
2 (1) That there shouldn't be restrictions on imports of machine tools.
2 (1) He believes that restrictions on foreign imports is a bad idea.
2 (1) That the restrictions would just continue down and be worse than no restrictions.
2 (1) Restricting imports of the less-expensive foreign-made machine tools would make industries that use such tools less competitive in the global market.
2 (1) The author believes trade restrictions would create more problems in the long run.
2 (1) That restrictions on Japanese tools will lead to restrictions on Japanese products.
2 (1) Restrictions would equal more expensive American products, loss of sales, and protection from foreign competition.
1 (2) The author believes that this is the wrong approach. **(incomplete-but # 32 is correct)**
1 (1) That in the long run trade protection harms those it was intended to help.
1 (1) The machine tool industry is wrong. **(incomplete)**
1 (0) The author believes import restrictions would mean fewer sales. **(copies question in # 34)**
0 (0) This is a wrongheaded argument. **(incomplete-and # 32 is incorrect)**

34. Why does the author believe that import restrictions would mean fewer sales?

Guidelines: The best answer includes: (1) restrictions mean American manufacturers have to use higher cost tools and (2) higher cost tools mean higher cost products and (3) higher cost products mean fewer sales. For **2** pts, the answer must include two of the above statements. Answers that include one statement receive 1 pt. (Numbers in **bold** in parentheses below indicate approximately which statements are present in answer)

The author believes that Protectionist policy by the American government motivates the preservation of normal profits of American industries because the author believes that as a
consequence of protectionist policy by the American government, American industry produces with high cost American machine tool; and if American industry produces with high cost American machine tool, then American industries produce high cost products; and if American
industries produce high cost products, then American industries sell fewer products; and if American industries sell fewer products, then there is a decrease in profits of American industries; and a decrease in profits of American industries motivates the preservation of normal profits of American industries.

2 (1) It would mean fewer sales by manufacturers who are forced to use the more expensive tools. Sales would be lost because their manufacturers would no longer have competitive prices as a result. **(1,3)**

2 (1) Import restrictions would result in company's having to buy more expensive American products. That increase would be included in product cost, and increased cost means fewer sales. **(1,2,3)**

2 (1) Because import restrictions would lead to more expensive products from the manufacturers which would lead to fewer sales. **(2,3)**

2 (1) Because restrictions would cause products to cost more and therefore not be as competitive. **(2,3)**

2 (1) Because manufacturers would have to buy more expensive machine tools thus raising the price of their production, making more expensive products. **(1,2)**

2 (1) The products produced would be more expensive than they would have been had they been produced with less expensive tools. **(2,1)**

2 (1) Because the price of the products would go up due to higher machine tool costs. **(2,1)**

1 (1) Because higher prices = fewer sales. **(3) (incomplete)**

1 (1) Because it would drive prices up of American products. **(2) (not specific enough)**

1 (1) Because American tools would become more expensive. **(1) (incomplete)**

1 (1) Because these products would also lose sales to cheaper foreign imports. **(3) (incomplete)**

1 (1 Because prices would increase. **(2) (not specific enough)**

1 (1) Higher cost tools would mean less could be bought. **(3) (incomplete)**

(5)

Friday evening a car swerved off Route 69. The vehicle struck a tree. The passenger, a New Jersey man, was killed. David Hall, 27, was pronounced dead at the scene by Dr. Dana Blanchard, medical examiner. Frank Miller, 32, of 593 Foxon Rd., the driver, was taken to Milford Hospital by Flanagan Ambulance. He was treated and released. No charges were made. Patrolman Robert Onofrio investigated the accident.

35. Was anyone killed?

Yes, Davis Hall died.

2 (9) Yes.

2 (2) Yes. The passenger was.

2 (2) Yes. David Hall died.

36. Why was Miller hurt?

Guidelines: For 2 pts answer must include "Miller's car hit a tree" or "Miller in auto accident," etc.

Because the automobile hit the tree.

2 (9) Because he hit (or drove into) a tree in the accident.

2 (2) He had gotten into an accident.

2 (1) He was in an accident.

1 (1) He was in the car when it swerved off the road. **(not specific enough)**

0 (0) He was the driver. **(not specific enough)**

37. How did Miller get to the hospital?

The Flanagan ambulance company took him to the Milford hospital.

2 (9) Flanagan ambulance (took him there).

2 (4) By an ambulance.

(6)

John is a programmer who works for a large corporation. John entered his manager's office. He was sitting at his desk. John asked for a raise. He wanted to buy a home computer for his child with the money. John and his manager reviewed his accomplishments.

38. Is John a programmer?

Yes.

2 (13) Yes.

39. Does John work for a corporation?

Yes.
2 (13) Yes.

40. Who was sitting at the desk?

The manager.
2 (10) His (or John's) manager.
2 (1) John's manager, I presume. Could have been anyone, we don't know for sure.
2 (1) His boss.
0 (1) John or his manager.

41. Who wanted to buy a computer?

John.
2 (13) John.

42. Where does John work?

Guidelines: Since question could require a specific name for place of work, and since passage does not state a name, "not stated" or "insufficient information" receives 1 pt.

In a large corporation..
2 (11) At a large corporation.
1 (1) Insufficient information from this passage.
0 (1) Don't know. **(not specific enough)**

43. How will John buy the computer?

With the money.
2 (13) With the money from the raise.
1 (0) With money. **(not specific enough)**

44. For whom does John want to buy the computer?

Guidelines: Son, daughter, kids (plural) or family receive 1 pt.

For his child.
2 (11) For his child.
1 (2) For his son.
1 (0) For his kids.

45. Why did John ask for a raise?

Because he wanted buy a home computer .

2 (11) He wanted to buy a computer for his child.

2 (2) So he could afford to buy a computer for his child.

0 (0) Because he didn't have enough money.

46. What was John's manager doing when John asked for the raise?

Sitting at his desk.

2 (12) Sitting at his desk.

2 (0) Sitting.

1 (1) Don't know. Perhaps sitting at his desk.

47. What was John's goal in entering his manager's office?

To ask for a raise.

2 (12) To get (or ask for) a raise.

0 (1) Not stated.

48. Why did John and his manager review his accomplishments?

Guidelines: For 2 pts answer must allude to the "possibility" of or to see if John "deserved" a raise.

Because John asked for a raise.

2 (10) To determine whether John should get a raise.

2 (1) Not stated. It is assumed that it was part of the discussion about John's raise.

2 (1) The manager's decision would be based on this information.

1 (0) To get a raise.

1 (0) To see how he was doing on the job.

0 (1) I don't know. (Because they didn't tell one).

Appendix C

Classification of Questions on Reading Comprehension Test by Linguistic/Textual Domain (A-D) and Type of Knowledge Required (1.1-1.2, 2.1-2.3)

(See pages 13 and 14 of this paper for description of classification scheme)

Question	Classification	Question	Classification
PAM		OpEd	
1	C 2.1	32	C 1.2
2	C 2.1	32	C 1.2
3	C 2.1	33	C 2.1
4	C 2.1	34	C 2.1
JULIP		SAM	
5	D 2.1	35	B 1.2
6	A 1.2	36	C 2.1
7	C 2.1	37	A 1.1
BORIS		Kind Types	
8	A 2.3	38	B 1.1
9	A 1.2	39	B 1.2
10	B 2.3	40	A 2.1
11	B 2.2	41	A 1.2
12	C 2.3	42	A 1.2
13	A 1.2	43	A 1.2
14	D 2.1	44	A 1.2
15	B 2.3	45	C 2.1
16	C 2.2	46	C 2.1
17	A 2.1	47	C 2.1
18	A 1.2	48	C 2.2
19	C 2.1		
20	C 2.1		
21	B 1.2		
22	A 1.2		
23	A 2.1		
24	C 2.1		
25	C 1.2		
26	C 2.1		
27	A 2.1		
28	A 2.3		
29	C 2.1		
30	A 1.2		
31	A 2.1		

Appendix D

Readability Analysis of AI Text Samples

AI System	Flesch Reading Ease	Flesch-Kincaid Grade Level
PAM	Very Easy	5th
JULIP	Very Difficult	15th
BORIS	Fairly Easy	7th
OpEd	Extremely Difficult	17th
SAM	Easy	6th
Kind Types	Fairly Easy	7th

Readability Formulas

Flesch Reading Ease

1.015 x (average sentence length)
= .846 x (number of syllabus per 100 words)
206.835 - Total = Flesch Reading Ease Score:

Score	Reading Difficulty	Approx. Grade Level
90-100	Very Easy	4th
80-90	Easy	5th
70-80	Fairly Easy	6th
60-70	Standard	7th - 8th
50-60	Fairly Difficult	Some High School
30-50	Difficult	High School-College
0-30	Very Difficult	College level & up

Flesch-Kincaid Grade Level

(0.39) x (average number of words per sentence)
= (11.8) x (average number of syllabus per word)
Total - 15.59 = Grade Level

Appendix E

Correlation Between Score on the Reading Comprehension Test and the CTBS Grade Equivalency Groups*

	Correlation Coeff.	Number of Cases
RCT	0.61	265

Correlations Between Scores on Individual System Texts and the CTBS Grade Equivalency Groups*

	Correlation Coeff.	Number of Cases
PAM	0.38	281
JULIP	0.40	280
BORIS	0.39	279
OpEd	0.57	278
SAM	0.30	277
Kind Types	0.44	266

*Note: All correlations are significant at the 0.01 level. Group is treated as a four category variable.

Correlation Between Score on Reading Comprehension Test and School Grade Level*

	Correlation Coeff.	Number of Cases
RCT	0.47	265

Correlations Between Scores on Individual System Texts and School Grade Level*

	Correlation Coeff.	Number of Cases
PAM	0.28	281
JULIP	0.43	280
BORIS	0.26	279
OpEd	0.48	278
SAM	0.17	277
Kind Types	0.28	266

*Note: All correlations are significant at the 0.01 level. Grade is trcated as a three category variable.

Appendix F

Descriptive Statistics for the Six Computer Systems by Grade Level

Grade	PAM				JULIP				BORIS				OPEd				SAM				Kind Types			
	Max Possible=8				Max Possible=6				Max Possible=48				Max Possible=6				Max Possible=6				Max Possible=22			
Level	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
6	5.30	1.63	1-8	64	2.27	1.63	0-6	64	34.91	6.05	18-43	64	1.73	1.51	0-5	64	5.22	1.24	0-6	64	19.34	3.19	9-22	61
8	5.97	1.59	0-8	174	2.58	1.65	0-6	173	36.41	4.24	22-44	172	2.13	1.64	0-6	171	5.16	1.26	0-6	170	19.94	2.31	9-22	162
11	6.58	1.52	2-8	43	4.70	1.30	2-6	43	38.65	4.48	25-45	43	4.35	1.09	1-6	43	5.86	0.41	4-6	43	21.35	1.07	17-22	43
System Response	7.00			1	6.00			1	38.00			1	4.00			1	6.00			1	20.00			1

Appendix G

Descriptive Statistics for the Six Computer Systems by Grade Equivalence

CTBS Grade Equiva-	PAM				JULIP				BORIS				OPEd				SAM				Kind Types			
	Max Possible=8				Max Possible=6				Max Possible=48				Max Possible=6				Max Possible=6				Max Possible=22			
lence	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n	Mean	SD	Range	n
4	5.17	1.83	2-7	6	2.5	1.64	1-5	6	28.33	9.33	18-38	6	0.83	0.98	0-2	6	4.00	2.53	0-6	6	16.00	4.97	9-20	4
5	4.93	1.79	0-8	15	1.67	1.63	0-4	15	32.50	5.13	25-41	14	1.07	0.92	0-2	14	5.29	1.27	2-6	14	18.69	2.56	14-22	13
6	4.93	1.86	1-8	14	3.07	1.59	1-6	14	34.00	4.71	22-41	14	1.21	1.58	0-4	14	4.07	1.94	0-6	14	17.58	3.18	12-22	12
7	4.75	1.50	4-7	4	1.25	0.50	1-2	4	36.00	2.71	32-38	4	0.50	0.58	0-1	4	6.00	0.00	6-6	4	19.67	1.15	19-21	3
8	5.43	1.60	4-8	14	2.71	1.54	0-5	14	35.00	3.66	29-42	14	2.00	1.30	0-4	14	4.86	1.29	2-6	14	20.29	1.54	17-22	14
9	5.58	1.54	3-8	62	2.05	1.49	0-6	61	35.87	4.83	23-43	62	1.60	1.51	0-6	62	5.24	1.18	2-6	62	19.00	2.99	6-22	60
10	6.03	1.32	4-8	35	2.17	1.72	0-6	35	36.85	3.32	28-42	35	2.71	1.71	0-5	35	5.14	1.06	2-6	35	20.26	2.27	14-22	34
11	5.81	1.61	2-8	42	2.81	1.73	0-6	42	36.71	5.12	23-45	42	2.05	1.43	0-5	42	5.38	0.99	2-6	42	20.68	1.68	15-22	41
12	5.85	1.91	4-8	13	2.85	1.07	2-5	13	37.13	3.97	31-43	13	2.15	1.82	0-5	13	5.23	1.30	2-6	13	20.75	1.76	17-22	12
13	6.76	1.37	2-8	76	4.11	1.61	0-6	76	38.44	4.03	25-45	75	3.89	1.35	0-6	74	5.73	0.65	4-6	73	21.11	1.43	16-22	73
System Response	7.00			1	6.00			1	38.00			1	4.00			1	6.00			1	19.00			1