Reading to learn and reading to integrate: new tasks for reading comprehension tests?

Latricia Trites *Murray State University* and
Mary McGroarty *Northern Arizona University*

To address the concern that most traditional reading comprehension tests only measure basic comprehension, this study designed measures to assess more complex reading tasks: Reading to Learn and Reading to Integrate. The new measures were taken by 251 participants: 105 undergraduate native speakers of English, 106 undergraduate nonnative speakers, and 40 graduate nonnative speakers. The research subproblems included determination of the influence of overall basic reading comprehension level, native language background, medium of presentation, level of education, and computer familiarity on Reading to Learn and Reading to Integrate measures; and the relationships among measures of Basic Comprehension, Reading to Learn, and Reading to Integrate. Results revealed that native language background and level of education had a significant effect on performance on both experimental measures, while other independent variables did not. While all reading measures showed some correlation, Reading to Learn and Reading to Integrate had lower correlations with Basic Comprehension, suggesting a possible distinction between Basic Comprehension and the new measures.

I Introduction

Each year, thousands of international students apply to American universities in the hope of obtaining a degree from an English-speaking university, and one of the hurdles they face is attaining a 'passing' score on the Test of English as a Foreign Language (TOEFL). Although not designed as a gatekeeper by Educational Testing Service (ETS), the TOEFL is often used as such by many institutions of higher education across the USA (Educational Testing Service, 1997). Prior to 2000, the test assessed basic reading and listening comprehension, as well as grammatical ability. While these skills are essential, they represent the minimum needed to succeed in

Address for correspondence: Latricia Trites, Assistant Professor, Murray State University, Department of English and Philosophy, 7C Faculty Hall, Murray, KY 42071, USA; email: latricia.trites@murraystate.edu

Language Testing 2005 22 (2) 174–210 10.1191/0265532205lt299oa

higher education. ETS, aware of this minimum standard and, as Bachman (2000) mentions, the need for task authenticity, embarked on a large-scale project to redesign TOEFL to better reflect the academic language skills required in higher education. Among other goals, the TOEFL 2000 project (Enright *et al.*, 1998) outlined plans to establish reading tasks for four distinct purposes:

- finding information;
- achieving basic comprehension;
- learning from texts; and
- integrating information.

The latter two purposes for reading represent a departure from traditional reading tests and constitute more complex tasks that require more cognitive processing. Tasks appropriate to measure these new purposes needed to be developed and validated.

The project reported here pursued the creation and evaluation of these new task types, the development of scoring rubrics, and the evaluation of native language effects on task and test performance (Educational Testing Service, 1998). In addition, because these were new reading tasks, some evidence for their validity was sought by establishing a baseline for native speakers and then comparing that baseline to performance of nonnative speakers. The TOEFL 2000 reading construct paper (Enright *et al.*, 1998) suggested that a Reading to Learn task would require students to recognize the larger rhetorical frame organizing the information in a given text and carry out a task demonstrating awareness of this larger organizing frame. Enright *et al.* (1998) hold that in reading to learn readers must integrate and connect information presented by the author with what they already know. Thus, readers must rely on background knowledge of text structures to form a Situation Model, a representation of the content, and a Text Model, a representation of the rhetorical structures of the text, as postulated by van Dijk and Kintsch (1983) and discussed by Perfetti (1997). Goldman (1997: 362) asserted that, to learn from texts, readers must have an awareness of text structure and know how to use it to aid comprehension. Reading to Learn can be assessed in a variety of ways. McNamara and Kintsch (1996) suggested that inferencing and sorting tasks requiring readers to process the text based on domain-specific knowledge of the text structures could yield a representation of the readers' ability to learn from the text. Hence, we postulated that one useful means of assessment would be to have participants recall information and reproduce information relationships reflecting their concept of text structure

(Enright *et al.*, 1998: 46–48). For the Reading to Learn task, we assessed readers' knowledge model through their ability to recall and categorize information from a single text (Enright *et al.*, 1998: 57).

Another goal of the project was to assess Reading to Integrate information, which requires readers to integrate information from multiple sources on the same topic. Reading to Integrate goes a step further than Reading to Learn because readers must integrate the rhetorical and contextual information found across the texts and generate their own representation of this interrelationship (Perfetti, 1997). Therefore, readers must assess the information presented in all sources read and accept or reject pieces of it as they create their own understanding. One means of assessing integration of information found in typical university assignments is the open-ended task of generating a synthesis based on one or more texts (Enright *et al.*, 1998: 48–49). We used a writing task, specifically a writing prompt, that elicited the reader's perception of the authors' communicative purposes (Enright *et al.*, 1998: 56) as well as amount of information retained from two texts to test Reading to Integrate.

II Related literature

Recent research has begun to explore the development of tasks that distinguish the constructs of Reading to Learn from basic comprehension. Researchers (van Dijk and Kintsch, 1983; McNamara and Kintsch, 1996; Goldman, 1997) have determined that reading to learn requires an interaction between the Text Model of a text as well as its Situation Model, thus resulting in a more difficult measure. These researchers further suggest that Reading to Learn can be assessed through measures that go beyond recall, summarization, and text-based multiple-choice questions.

The construct of Reading to Integrate requires that readers not only integrate the Text Model with the Situation Model, but also that they create what Perfetti (1997: 346) calls a Documents Model, consisting of two critical elements: 'An Intertext Model that links texts in terms of their rhetorical relations to each other and a Situations Model that represents situations described in one or more text with links to the texts.' He argues that the use of multiple texts as opposed to a single text brings into clearer focus the relationship between the Text Model and the Situation Model. This again suggests that Reading to Integrate should be more difficult than Reading to Learn.

Because these constructs go beyond basic comprehension, Reading to Learn and Reading to Integrate are hypothesized to be more difficult reading tasks than Reading to Find Information and Reading for Basic Comprehension. Perfetti (1997) further suggests that Reading to Integrate is a more difficult task than Reading to Learn because it not only requires an integration of a Text Model and a Situation Model but requires an integration of multiple Text Models and multiple Situation Models. Thus, current reading theory suggests a difficulty hierarchy of reading tasks based on the level of integration necessary to complete the tasks successfully. Several studies (Perfetti *et al.*, 1995; 1996; Britt *et al.*, 1996; Wiley and Voss, 1999) have attempted to move beyond basic comprehension and examine readers' ability to integrate the information from multiple texts into one cohesive knowledge base by having students make connections, compare, or contrast information across texts.

Additionally, recent research has addressed the effects of computers on reading and assessment; such research is relevant to the current project because the new TOEFL is administered via computers. Reading-medium studies have shown that the only effect that computers have on reading is related to task (Reinking and Schreiner, 1985; Reinking, 1988; van den Berg and Watt, 1991; Lehto *et al.*, 1995; Perfetti *et al.*, 1995; 1996; Britt *et al.*, 1996; Foltz, 1996; Wiley and Voss, 1999). Taylor *et al.* (1998) found that, after minimal computer training, familiarity with technology did not have a significant effect on examinees' performance on TOEFL-like questions. Because of the relevance of computer familiarity to TOEFL administration, a brief measure of computer familiarity was included in the research.

For this project, we asked three research questions:

1) Is performance on a measure of Reading to Learn affected by medium of presentation (paper versus computer), technology familiarity, native language (native versus nonnative speakers of English), or level of education (graduate versus undergraduate)?
2) Is performance on a measure of Reading to Integrate affected by medium of presentation (paper versus computer), technology familiarity, native language (native versus nonnative), or level of education (graduate versus undergraduate)?
3) To what extent are measures of finding information/basic reading comprehension, Reading to Learn, and Reading to Integrate related?

III Methods

1 Participants

Two hundred and fifty-one participants, the majority undergraduates, volunteered to take part in this study. The sample consisted of 105 undergraduate native speakers of English (NSUs), 106 undergraduate nonnative speakers (NNSUs), and 40 graduate nonnative speakers (NNSGs) of English at a midsized southwestern university. All data were collected between February and October 1999. All undergraduate participants were recruited through large undergraduate classes in the areas enrolling most NNSs (business administration, hotel management, engineering, social sciences, and humanities). We tested all NNSs accessible at the institution at the time of data collection; compared to a national sample of international students from the prior academic year, we had a relatively larger proportion of undergraduate relative to graduate students. Nearly all undergraduate participants were young adults with an average age of 21. Nonnative speakers were also recruited from students enrolled in the summer intensive English program, which is made up of students needing to increase TOEFL scores to at least 500 in order to enroll at a university. We included 46 participants (32% of NNS sample) with TOEFL scores below 500 in the nonnative sample. Graduate nonnative speakers ($n = 40$) were recruited from the entire university population and had an average age of 30.75. Nonnative speakers represented a range of language backgrounds: One third were Japanese, with other Asian, Germanic, and Romance languages also substantially represented. Both the relatively modest sample size and the all-volunteer nature of the participant sample preclude direct generalization to the worldwide TOEFL population, but participants were representative of the levels of international students at the institution where they were enrolled. Participants who completed all four data collection sessions received a payment of US$10 per hour (US$40 for the entire project).

2 Instruments

This project used three existing instruments, two to determine initial reading levels and one to assess levels of computer familiarity, and two new instruments, one for Reading to Learn and one for Reading to Integrate; these were developed especially for the project. Each of the new measures also served as the basis for an additional measure

of basic reading comprehension related directly to the text included in the new task. Thus, each participant completed a total of seven different instruments.

a Existing instruments: Initial levels of reading comprehension were determined based on the Nelson–Denny Reading Test (Nelson–Denny), Form G, used to identify the reading levels of the NSs, and three retired versions of the Institutional TOEFL Reading Comprehension Section (TOEFL Reading Comprehension), used to identify the reading levels of the NNSs. Although each of these tests was used to assess reading levels in the population for which it had been developed, all 251 participants took both tests in order to provide comparative data. All 251 participants also completed a brief computer familiarity questionnaire.

Participants' computer familiarity was determined through an 11-item questionnaire based on a longer, 23-item questionnaire previously developed by ETS (Eignor *et al.*, 1998). In the present study, we used only the 11 items that loaded the most heavily on the major factors resulting from administration to a large sample of TOEFL participants. For these 11 items, developers determined the reliability to be .93 using a split-half method (Eignor *et al.*, 1998: 22). This brief questionnaire took approximately 5 minutes to complete; reliability in our sample, using coefficient alpha was .87.

b Texts used for new measures: In developing the new tasks, we selected texts that would conform to the design specifications of TOEFL 2000. They were problem/solution texts recommended as one of the potentially relevant text types for TOEFL 2000 (Enright *et al.*, 1998). Longer texts were used because these represented more challenging and authentic academic tasks (Enright *et al.*, 1998). We used one 1200-word and two 600-word texts. The longer text (Tennesen, 1997) was used to assess Reading to Learn and the two 600-word texts (Monks, 1997; Zimmerman, 1997) were used to assess Reading to Integrate. We chose these text lengths based on work by Meyer (1985a) and further research by the first author indicating that natural science texts between 1200 and 1500 words included representation of all necessary macro-rhetorical structures of problem/solution texts with or without explicit signaling. While 1200–1500 word texts provide optimal representation of the macro-rhetorical structures, texts of 600-words provide all the basic macro-rhetorical structures present in problem/solution texts. Thus, these

lengths were long enough for adequate argumentation but not so long that they were excessively redundant (Enright *et al.*, 1998). Texts were also matched for readability according to standard readability scales such as the Flesch–Kincaid, Coleman–Liau, and Bormuth scales, and averaged a minimum of grade level 11.0 to 12.0 on these scales. Also, all texts pertained to natural and social sciences; each text covered environmental issues such as air and water pollution (Enright *et al.*, 1998). Thus, text topics were similar across tasks.

c New instruments used in the study: Three new reading measures were used in this study to assess Reading to Learn, Reading to Integrate, and Basic Comprehension. Trites (2000: Chapters 2 and 3) presents a more extensive review of literature and rationale for development of the new measures.

- Reading to Learn: The first new measure, completion of a chart, was used to determine participants' ability to read to learn. Spivey (1997: 69) suggests that readers' categorization of information in text offers insight into their cognitive processes and their making of meaning. We designed a measure to be used with a 1200-word text that students read on either paper or computer. Students were asked to recall, identify, and categorize information from the text on a chart reflecting macro-rhetorical structures, called macrostructures in this study (problems and solutions), and other types of information from problem/solution texts (causes, effects, and examples), categories based on the work of Meyer (1985a). The scoring rubric, based on work by Meyer (1985b) and later modified by Jamieson *et al.* (1993), awarded points only for the upper levels of textual structure represented on the chart (for task and scoring rubric, see Appendix 1). We weighted the information supplied on the chart as follows: 10 points for correct information in the problem and solution categories; five points for correct information supplied in the cause and effect categories; and one point for accurate examples. This weighting reflects Meyer's (1985b) hierarchical levels, which characterize problem and solution propositions as higher order structures, while the other categories represent lower order propositions.[1] The theoretical maximum score for this scale

[1]Students received no points for information improperly placed or for information not found in the text.

was 241, which would result from maximum points given in all categories. The first author and two research assistants spent 35–40 hours creating, revising, norming the scoring rubric, and developing the scoring guide (Trites, 2000: Chapter 3). To determine interrater reliability, we used coefficient alpha, rather than percentage of agreement because percentage of agreement inflates the likelihood of chance agreement (Hayes and Hatch, 1999). After norming, overall interrater reliability was .99 (coefficient alpha) with similarly high reliabilities assessed with similarly high alpha coefficients for all subcategories.[2]

- Reading to Integrate: The second new measure assessed Reading to Integrate. The task used to assess Reading to Integrate required participants to read two 600-word texts and compose a written synthesis. The prompt asked students to make connections across the range of ideas presented; thus, we asked readers to synthesize information rather than summarize or make comparisons (Wiley and Voss, 1999). This synthesis was scored based on an analytic scale ranging from 0 to 80, reflecting readers' ability to recognize and manipulate the structure of the texts, include specific information, and express connections across texts through the use of cohesive devices (for task and scoring rubric, see Appendix 2). The test was designed to measure the integration of content from both readings and did not assess other aspects of writing such as the creation of rhetorical style, grammaticality, or mechanics. The rubric was composed of three subcategories: integration ability, macrostructure recognition, and use of relevant details. The integration subscore was awarded the highest point values because this was the predominant skill being tested. It scored participants on their ability to make connections across texts based on the manipulation of the textual frames in both texts. The second subcategory awarded points for the ability to recognize and articulate the macrostructures (problem, cause, effect, or solution) present in each text. This subcategory was similar to the categorizing task used in the Reading to Learn measure with the additional constraint that participants had to express the connections overtly. The third subcategory in the scoring rubric analysed the ability to use

[2]We recognize that tasks requiring high inference measures plus extensive norming and revision of the scoring rubric pose feasibility issues in large-scale testing. Further research is needed to determine whether and how such scoring procedures could be adapted in standardized testing for numerous test-takers.

relevant details as support in the written synthesis. The first author and two research assistants spent 30 hours revising, norming the scoring rubric, and developing a decision guide, resulting in an overall interrater reliability of .99 (coefficient alpha) with similarly high alphas for all subcategories.

- Basic Comprehension: The third construct was measured by multiple-choice tests related specifically to the texts used in the new tasks. These tests were created by TOEFL Test Development staff and followed current TOEFL reading section specifications. We used two multiple choice tests, Basic Comprehension Test 1 (BC1) and Basic Comprehension Test 2 (BC2), 20 items each, one for the longer passage used to assess Reading to Learn and one for the two passages used to assess Reading to Integrate. Both were scored based on number of items answered correctly. Reliability on BC1, calculated based on 251 participants was .84 (coefficient alpha). Inadvertently, the order of the texts used in BC2 was different for the two different media; however, reliability on both versions of the test was high. For those who took BC2 based on paper texts ($n = 127$), reliability was .84 (coefficient alpha); for those who took BC2 based on computerized texts ($n = 124$), reliability was .86 (coefficient alpha).

3 Design for data collection

This study used a 2×2 repeated measures design to examine performance on the new reading tasks. Native speaker undergraduates and nonnative speaker undergraduates were divided into two groups each of equal ability as determined by performance on the baseline standardized measures of reading comprehension (Nelson–Denny or TOEFL). Half of each group read texts on paper; the other half read the same texts on a computer screen. A smaller group of nonnative speaker graduates, equally divided, were also included for a comparison between performance by graduate and undergraduate nonnative speakers. Additionally, the administration of the new measures was counterbalanced to control for any practice effect.

a Procedures: All participants met with the researchers in four sessions each lasting about an hour. The first two sessions were devoted to administering the existing instruments. During Session 1, participants received an introduction to the study and took one of the two

standardized basic reading comprehension measures (Nelson–Denny or TOEFL Reading Comprehension). Students completed the computer familiarity questionnaire and the Nelson–Denny Test at the same testing session because the Nelson–Denny was shorter than the TOEFL Reading Comprehension. During Session 2, participants took the other standardized basic reading comprehension measure.

Next, each participant group was subdivided into two subgroups for computer-based or paper reading of the texts for the new tasks. The subgroups were matched on their performance on initial reading measures; the Nelson–Denny was used for native speakers and the TOEFL Reading Comprehension was used for nonnative speakers. Independent *t*-tests run on these reading measures showed no significant difference in basic comprehension for the newly created subgroups assigned to each medium, ensuring that they were balanced for initial reading levels. Participants stayed in the same subgroups for the duration of the study. To ensure uniformity of response mode, all participants, whether they read the source texts on the computer or on paper, responded to the reading tasks using paper and pencil format.[3]

The last two sessions, each lasting approximately one hour, were dedicated to administration of the new measures. The Reading to Learn session took slightly longer to administer because administrative procedures were longer for this novel task. The new tasks were counterbalanced to control for practice effect; thus, half of the participants took the Reading to Learn measure first and half took the Reading to Integrate measure first. During Session 3, we administered the first new measure (for ease of discussion Reading to Learn is discussed first) and BC1. At this session, students were given 12 minutes to read a 1200-word passage either on computer or on paper. We limited the time allowed for reading based on 100 words per minute, thought to be ample (Grabe, personal communication, 1998). After examinees read the text, they were given 4 minutes to take notes on a half sheet of paper. Participants were instructed to take minimal notes due to the time constraints. Next, the text was removed and examinees were allowed 15 minutes to complete a chart based on the reading with the aid of their notes. After completing this Reading to Learn activity, participants were allowed to use the text and

[3]Although responses could have been entered and perhaps scored by computer, this would have introduced factors not directly related to our research questions and remains an area for further study.

were given 15 minutes to answer BC1. Following these new testing sessions, 49 participants were selected for a related interview concerning the cognitive processes used in task completion (for further details, see Trites, 2000: Chapter 6).

During Session 4, students were given 12 minutes to read two short texts (600 words each) either on computer or paper. After participants read the assigned texts, they were given 4 minutes to take one-half page of notes (Enright *et al.*, 1998). Next, the texts were removed and participants were asked to demonstrate Reading to Integrate by writing a synthesis of the texts with the aid of their notes (15 minutes allowed for this task). After completing the Reading to Integrate task, participants were allowed to see the texts again and answered BC2 (15 minutes allowed for this task). In one Reading to Integrate session, for unknown reasons, six of the seven participants read only one text. Because we cannot explain the cause of this anomalous session, we have eliminated scores from the session's seven participants from subsequent analyses, thus slightly reducing the *N* size for the Reading to Integrate measure.

b Variables used in study: The six independent variables included three nominal (Native Language Background, Medium of Text Presentation, and Level of Education) and three interval variables (Nelson–Denny, TOEFL Reading Comprehension, and Computer Familiarity). The four dependent variables were Reading to Learn, Reading to Integrate, BC1, and BC2.

IV Results

First we present the descriptive statistics for all reading measures followed by a systematic analysis of independent variables that might affect participant performance on the new measures. Scatterplots were checked for all reading measures to ensure normality of data. Kurtosis and skewness levels for all reading measures were found to be within normal limits, indicating a relatively normal distribution. Descriptive statistics for all existing measures are shown in Table 1. Means for these measures show a consistent pattern: the native speaker undergraduates had the highest mean followed by the nonnative speaker graduates followed by the nonnative speaker undergraduates. On the reading measures, Nelson–Denny and TOEFL Reading Comprehension, the nonnative speaker undergraduates

Table 1 Descriptive statistics for existing measures for three participant groups

Group	*n*	Mean	sd	k/Max
Nelson–Denny				
NSU	105	126.48	16.46	156
NNSU	106	67.24	31.91	156
NNSG	40	88.88	21.88	156
Total participants	251	95.47	36.93	156
TOEFL Reading comprehension				
NSU	105	61.30	4.24	67
NNSU	106	50.30	8.53	67
NNSG	40	57.15	4.55	67
Total participants	251	55.99	8.19	67
Computer familiarity				
NSU	104	38.08	3.60	44
NNSU	104	34.82	5.99	44
NNSG	40	35.63	6.02	44
Total participants	248	36.31	5.33	44

Note: k/Max: number of items or maximum possible score

showed the largest variance in performance, while on the computer familiarity measure, the variance of both nonnative speaker groups was substantially larger than that of the native speakers.

The same pattern emerged for the means on the new measures (see Table 2) as for the existing measures. The native speaker undergraduate group performed better on all new measures than both of the nonnative speaker groups. The nonnative speaker graduate group performed better than the nonnative speaker undergraduate group on all measures as well. This robust pattern of performance was also found in the variance of three of the four new measures. On BC1 and BC2 the performance of the native speaker undergraduates showed the least amount of variance, followed by the nonnative speaker graduates, followed by the nonnative speaker undergraduates. On Reading to Integrate, the native speaker undergraduate group showed substantially less variance than the nonnative speaker groups; however, the variance of the two nonnative speaker groups was almost identical. On Reading to Learn, all three groups showed considerable variance.

Table 3 reveals the range of awarded points achieved by all participant groups. The nature of the Reading to Learn point system created a maximum possible point value (241) that no participant achieved. We speculate that there are at least three possible causes of the discrepancy between the theoretical maximum and the range of observed scores:

Table 2 Descriptive statistics for new measures for three participant groups

Group	*n*	Mean	sd	k/Max
Reading to Learn (chart)				
NSU	105	51.85	19.86	241
NNSU	106	31.73	19.50	241
NNSG	40	44.68	19.27	241
Total participants	251	42.21	21.64	241
Basic Comprehension Test 1				
NSU	105	16.98	2.47	20
NNSU	106	11.73	4.25	20
NNSG	40	14.98	3.50	20
Total participants	251	14.44	4.23	20
Reading to Integrate (synthesis)				
NSU	101	63.65	11.05	80
NNSU	103	37.24	21.76	80
NNSG	40	53.60	21.03	80
Total participants	244*	50.86	21.63	80
Basic Comprehension Test 2				
NSU	105	15.91	2.78	20
NNSU	106	9.75	4.54	20
NNSG	40	12.85	3.61	20
Total participants	251	12.82	4.72	20

Notes: k/Max: number of items or maximum possible score; *n size reduced for reading to integrate because of anomalous testing session

- task novelty: no participant reported ever doing such a task previously;
- time allowed for task completion; and
- space on the response sheet: space constraints may have limited the amount of information that participants could include.

Future research would need to address these issues. However, for the Reading to Integrate measure, the full range of possible point totals was achieved by at least one participant in each group.

1 Computer familiarity

The overall plan for the analyses was to check the influence of the independent variables on the dependent measures, with computer familiarity being addressed first. Initially, we had proposed that if computer familiarity was significantly different across groups, it would be entered into all calculations as a covariate. To determine this, it was necessary to conduct an Analysis of Variance (ANOVA) for computer familiarity across the six participant/medium subgroups.

Table 3 Range of scores for new measures for three participant groups

Group	*n*	Minimum	Maximum	k/Max
Reading to Learn (chart)				
NSU	105	14	120	241
NNSU	106	0	86	241
NNSG	40	3	94	241
Total participants	251	0	120	241
Reading to Integrate (synthesis)				
NSU	101	38	80	80
NNSU	103	0	80	80
NNSG	40	5	80	80
Total participants	244*	0	80	80

Notes: k/Max: number of items or maximum possible score; *n size reduced for reading to integrate because of anomalous testing session

The resulting ANOVA ($F = 4.70$; $p > .05$) showed a significant difference between subgroups on the computer familiarity questionnaire; therefore, a *post hoc* Scheffé test was done to locate significant contrasts. After analysis of all possible subgroup contrasts, the *post hoc* Scheffé revealed that the only significant difference in subgroups appeared between the native speaker undergraduates and nonnative speaker undergraduates who read texts on paper. Hence, although there was one significant contrast, it occurred in two subgroups reading on paper, not in any of the subgroups who read on computer. All groups generally scored high on computer familiarity although, as noted, variance of the nonnative groups was greater. It was thus established that computer familiarity had no significant effect on participants who read texts on computer, so we did not use computer familiarity as a covariate in further analyses and proceeded to the three research questions of central interest to this study.

Because both Research Questions 1 and 2 are similar – except that they address the two different new reading measures, Reading to Learn and Reading to Integrate – we approached them in the same manner through ANOVA to identify the independent variables that could have significantly affected the results on the new measures.

2 *Research Question 1*

The first research question asked if performance on a measure of Reading to Learn was affected by medium of presentation, computer familiarity, native language, or level of education. We calculated a univariate ANOVA with Type III sums of squares on Reading to Learn with

Table 4 Performance on Reading to Learn measure by groups, medium, and test order ($n = 251$) (univariate analysis of variance)

Source	Type III sum of squares	df	Mean square	F
Group	21869.29	2	10934.64	28.10*
Medium	1215.86	1	1215.86	3.13
Test order	294.98	1	294.98	0.76
Group × medium	334.81	2	167.40	0.43
Group × test order	4.37	2	2.19	0.01
Medium × test order	391.73	1	391.73	1.01
Group × medium × test order	575.29	2	287.65	0.74
Error	92997.45	239	389.11	

Note: $*p < .05$

group status, medium of text presentation, and test order as possible contributing factors.[4] Table 4 shows that there were no significant interactions for any of the group, medium or test order combinations. The only significant main effect was group membership.

Because group membership was a combined measure that included both native language background as well as level of education, *post hoc* analysis was needed to identify the significant contrasts. Table 5 shows that there was a significant difference in performance on the Reading to Learn measure between the native speaker undergraduate and the nonnative speaker undergraduate groups, as well as a significant difference between the nonnative speaker undergraduate and nonnative speaker graduate groups. There was no significant difference in performance between the native speaker undergraduate and the nonnative speaker graduate groups. Therefore, the answer to Research Question 1 is that native language background and level of education did have a significant effect on performance on the Reading to Learn measure, but that medium of text presentation did not. Further, order of testing, whether participants took Reading to Learn or Reading to Integrate first, had no significant effect.

3 Research Question 2

The second research question, related to the first, asked if performance on Reading to Integrate was affected by medium of presentation,

[4]Test order was added as an additional variable to double check that our counterbalancing had been effective in controlling for any practice effect.

Table 5 *Post hoc* Scheffé for Reading to Learn measure ($n = 251$)

Group	n	Group	n	Mean difference	Standard error
NSU	105	NNSU	106	20.12*	2.72
		NNSG	40	7.17	3.67
NNSU	106	NNSG	40	−12.95*	3.66

Note: *$p < .05$

computer familiarity, native language, or level of education. Again, to ensure that counterbalancing of tests controlled for any practice effect, test order was added as an additional variable.

To answer this question, we proceeded to calculate a univariate ANOVA on the Reading to Integrate measure with group status, medium of text presentation, and test order entered as possible contributing factors. The results (Table 6) show, as for Research Question 1, that there were no significant interactions for any of the group, medium, or test order combinations; the only significant main effect was group membership. The answer for Research Question 2 is that native language background and educational level had a significant effect on Reading to Integrate, but medium of text presentation did not. *Post hoc* analysis of group contrasts showed that all three groups were distinct in their performance on Reading to Integrate (see Table 7).

4 Research Question 3

The third research question asked to what extent measures of basic comprehension, Reading to Learn, and Reading to Integrate were

Table 6 Performance on Reading to Integrate measure by groups, medium, and test order ($n = 244$[a]) (univariate analysis of variance)

Source	Type III sum of squares	df	Mean square	F
Group	36294.33	2	18147.17	55.82[b]
Medium	192.95	1	192.95	0.59
Test order	97.83	1	97.83	0.30
Group × medium	11.82	2	5.91	0.02
Group × test order	30.14	2	15.07	0.05
Medium × test order	1098.72	1	1098.72	3.38
Group × medium × test order	1488.58	2	744.29	2.29
Error	75430.37	232	325.13	

Notes: [a]*n* size reduced for Reading to Integrate because of anomalous testing session; [b]$p < .05$

Table 7 *Post hoc* Scheffé for Reading to Integrate measure ($n = 244$[a])

Group	*n*	Group	*n*	Mean difference	Standard error
NS	101	NNSU	103	26.41[b]	2.53
		NNSG	40	10.05[b]	3.37
NNSU	103	NNSG	40	−16.36[b]	3.36

Notes: [a]*n* size reduced for Reading to Integrate because of anomalous testing session; [b]$p < .05$

related. We used correlational analysis as the first step in answering this question. Results for the total participant population (see Table 8) showed moderate to high correlations across all reading measures. However, the analyses done for Research Questions 1 and 2 revealed that group status had a significant effect on performance on Reading to Learn and Reading to Integrate. Further, we realize that correlations are sensitive to variance, so the high correlations seen in the total population could have been an artifact of combining the three groups. Therefore, we examined the correlations among all reading measures for each group (available in Trites, 2000: Appendix 1, pp. 230–33). While the reading measures were still correlated often moderately, sometimes highly, magnitudes differed and sometimes dropped substantially. The text-specific multiple-choice measures, BC1 and BC2, consistently correlated more highly with the Nelson–Denny and TOEFL Reading Comprehension tests than with Reading to Learn and Reading to Integrate based on the same texts, suggesting a test method or construct effect. Because comparisons between different measures of basic comprehension were not a goal of the project, BC1 and BC2 were not used in further analyses. We conclude that, as expected, all reading measures were related, but the lower correlations between Reading to Learn and Reading to Integrate and the traditional basic comprehension measures led us to consider further types of analysis to identify the possible distinctiveness of the new measures.

5 Discriminant analysis

Because we were interested in determining how constructs differed, we sought additional analyses to help us better characterize the new constructs. Of the several possible statistical methods that could have been employed, two are most plausible: multivariate analysis of variance, usually associated with experimental research,

Table 8 Correlations for all reading measures for all participants ($n = 251$)

	TOEFL Reading Comprehension	Basic Comprehension Test 1	Basic Comprehension Test 2	Reading to Learn	Reading to Integrate[a]
Nelson–Denny	.90[b]	.85[b]	.84[b]	.66[b]	.69[b]
TOEFL Reading comprehension	1.00	.85[b]	.84[b]	.64[b]	.69[b]
Basic comprehension 1		1.00	.84[b]	.68[b]	.68[b]
Basic comprehension 2			1.00	.68[b]	.70[b]
Reading to Learn				1.00	.59[b]

Notes: [a]*n* size reduced for Reading to Integrate because of anomalous testing session; [b]$p < .05$

and discriminant analysis, usually associated with descriptive research (Tabachnick and Fidell, 1996). The present research was conducted with samples of naturally occurring student groups and was not experimental. Moreover, we were interested in finding ways to compare participant performance on the measures of the new constructs Reading to Learn and Reading to Integrate with performance on more traditional measures of basic comprehension. Thus we opted to use discriminant analysis because of its parsimony of description and clarity of interpretation (Stevens, 1996). Discriminant analysis, a technique recommended to describe group differences or predict group membership based on a comparison of multiple predictors (Huberty, 1994), has been used in other areas of applied linguistic research to investigate creation of a student profile of success or failure on Computer Assisted Language Learning (CALL) lessons (Jamieson *et al*., 1993) and accurate classification of text types into registers (Biber, 1993) among other purposes.

To further distinguish basic comprehension from the new constructs, we conducted discriminant analysis on each of the two language groups (native and nonnative) to determine whether Reading to Learn and Reading to Integrate would classify participants in the same way that Basic Comprehension would. We divided the native speaker and nonnative speaker groups into three levels: high, middle (mid), and low reading ability scorers, based on the basic comprehension measure chosen for that group (Nelson–Denny for native speakers; TOEFL Reading Comprehension for nonnative speakers). Research methodologists (Tabachnick and Fidell, 1996: 513) note that robustness is expected when the smallest group has is least 20; our smallest group was 25. To check the assumption of homogeneity of the variance/covariance matrices, we examined the outcomes of Box's M Test and found them all nonsignificant (Klecka, 1980). Each group was checked for outliers using Mahalanobis distance treated as Chi-Square, and no outliers were found (Tabachnick and Fidell, 1996). Thus the data met all assumptions required for use of discriminant analysis.

a Discriminant analysis for nonnative speakers: To organize the discriminant analysis in order to see if the Reading to Learn/Reading to Integrate Composite classified participants similarly to the measure of Basic Comprehension (for nonnative speakers, TOEFL Reading Comprehension), we divided the entire nonnative speaker group ($n = 146$) into three levels of basic comprehension: high, mid,

Table 9 Descriptive statistics for nonnative speakers by TOEFL reading comprehension reading ability groups (n = 143*)

Reading ability group	*n*	Mean on TOEFL reading comprehension	sd
High (≥56)	57	59.68	3.03
Mid (50–55)	42	52.81	1.67
Low (≤49)	44	42.11	5.47
Total	143*	52.26	8.22

Note: *Total n = 143 due to loss of three cases in anomalous Reading to Integrate session

and low. These reading ability groups of high, mid, and low were based on typical TOEFL Reading Comprehension score levels required for program entry. Participants were classified as high if their scores were greater than or equal to 550, or 56 and above on the scaled score on the TOEFL Reading Comprehension (550 is the cut score often used for graduate entry). Participants were classified as mid if scores ranged between 500 and 549, or 50 to 55 on the TOEFL Reading Comprehension. Participants were classified as low if their scores fell below 500, or 49 and below on TOEFL Reading Comprehension (500 is a minimum TOEFL score sometimes used for undergraduate admission often with the proviso that students enroll in ESL classes either prior to official enrollment or concurrently). For our entire nonnative speaker group, descriptive statistics on basic comprehension reading ability group membership levels appear in Table 9.

We ran SPSS Discriminant Analysis with initial grouping variables of high, mid, and low reading ability. We compared initial reading ability levels with high, mid, and low categories on the Reading to Learn/Reading to Integrate Composite, a new variable reflecting level of performance on the Reading to Learn and Reading to Integrate measures combined.[5] The discriminant analysis yielded one discriminant function with an eigenvalue of .92, responsible for 99.9% of the variance in outcomes. Wilk's Lambda for this function was .52, significant at ≤ .001; the associated Chi-Square value was extremely large (90.93) and highly significant ($p \leq .001$), indicating the group centroids on the composite Reading to Learn/Reading to Integrate function for the three nonnative speaker reading ability groups were significantly different. Both Reading to Learn and

[5]We first calculated two separate discriminant analyses, one for Reading to Learn and one for Reading to Integrate, but we found that both loaded on a single function, so we used the composite in subsequent analyses.

Reading to Integrate loaded significantly on the discriminant function at .86 for Reading to Learn and .78 for Reading to Integrate ($p \leq .05$).

Over half (64.9%) of the high reading ability group remained high on the new measure; less than half (42.9%) of the mid reading ability group remained classified as mid. Hence, the Reading to Learn/Reading to Integrate Composite was particularly influential in reclassifying the mid reading ability group and, to a lesser extent, the high group. However, most (81.8%) of the low reading ability group remained low on the Reading to Learn/Reading to Integrate Composite (see Table 10). Of the 143 nonnative speaker participants, 92 (64%) remained in the initial basic comprehension category on the composite; the rest moved, but in different directions. Twenty-one participants (14.7%) were classified into a higher category on the Reading to Learn/Reading to Integrate Composite than their initial basic comprehension level would have suggested, while 31 (21.7%) were reclassified into a lower category. Thus, 51 participants (36.4%), just over one third of the sample, were classified differently based on their Reading to Learn/Reading to Integrate Composite performance.

b Discriminant analysis for native speakers: Because one of our goals in this project was to probe the possible validity of these new measures by assessing performance of two groups, native as well as

Table 10 Discriminant analysis comparison of nonnative speaker reading ability groups with reading to learn/reading to integrate composite (n = 143*)

Reading ability group	Predicted group membership for Reading to Learn/Reading to integrate composite			Initial classification total
	High	Mid	Low	
Count				
High (≥56)	37	17	3	57
Mid (50–55)	13	18	11	42
Low (≤49)	2	6	36	44
Reclassification total	52	41	50	143
Percentage				
High (≥56)	64.9	29.8	5.3	100.0
Mid (50–55)	31.0	42.9	26.2	100.0
Low (≤49)	4.5	13.6	81.8	100.0

Note:*Total n = 143 due to loss of three cases in anomalous Reading to Integrate session

nonnative speakers, we conducted a parallel discriminant analysis for native speakers. Thus, for the native speakers we followed the same procedure, dividing the entire native speaker group ($n = 105$) into three levels of basic comprehension, high, mid, and low, based on score distances of $\pm .5$ standard deviations from the sample mean of the Nelson–Denny test for these participants. Native speakers need not take reading comprehension tests when entering the university, so the three-way split was based entirely on our sample data. Participants were classified as high if their scores on the Nelson-Denny were greater than or equal to 135. Participants were classified as mid if their scores ranged between 119–134 on the Nelson-Denny. Participants were classified as low if their scores fell at or below 118 on the Nelson-Denny. Descriptive statistics for the entire native speaker sample on basic comprehension group membership appear in Table 11.

The discriminant analysis yielded one discriminant function with an eigenvalue of .31, responsible for 98.7% of the variance in outcomes. Wilk's Lambda for this function was .76, significant at $\leq .001$; the associated Chi-Square value was large (26.39) and highly significant ($p \leq .001$), indicating the group centroids on the discriminant function for the three reading ability groups on the Reading to Learn/Reading to Integrate Composite were significantly different. Pooled within groups correlations between discriminating variables showed that Reading to Learn correlated with the first discriminant function at a level of .81; Reading to Integrate correlated with the second discriminant function at .71. This contrasts with findings for the nonnative speakers, where scores on the combined new measures loaded significantly on only one discriminant function. For native speakers, then, there is evidence for two significant discriminant functions, although the first accounts for almost all of the variance. Although these two measures (Reading to Learn and Reading to Integrate) loaded on two separate discriminant functions, they still

Table 11 Descriptive statistics for native speakers by Nelson–Denny reading ability groups ($n = 101^{*}$)

Reading ability group	*n*	Mean on Nelson–Denny	sd
High (≥135)	39	141.26	5.19
Mid (119–134)	37	125.65	5.09
Low (≤118)	25	102.88	10.98
Total	101*	126.04	16.52

Note: *Total $n = 101$ for discriminant analysis due to loss of four cases in anomalous reading to integrate session.

Table 12 Discriminant analysis comparison of native speaker reading ability groups with Reading to Learn/Reading to Integrate Composite (n = 101*)

Reading ability group	Predicted group membership for Reading to Learn/Reading to Integrate Composite			Initial classification total
	High	Mid	Low	
Count				
High (≥56)	37	17	3	
Count				
High (≥135)	28	3	8	39
Mid (119–134)	10	7	20	37
Low (≤118)	5	7	13	25
Reclassification total	43	17	41	101
Percentage				
High (≥135)	71.8	7.7	20.5	100.0
Mid (119–134)	27.0	18.9	54.1	100.0
Low (≤118)	20.0	28.0	52.0	100.0

Note: *Total n = 101 for discriminant analysis due to loss of four cases in anomalous reading to integrate session.

showed moderate correlations with the alternate function,[6] justifying the composite calculations.

Results of discriminant analysis for native speakers, seen in Table 12, show a different pattern than that observed for nonnative speakers. Nearly three-fourths (71.8%) of the native speakers classified as high in the basic comprehension reading ability group remained high on the Reading to Learn/Reading to Integrate Composite. For the mid group, however, only 18.9% remained classified as mid. Just over half (52%) of the low reading ability group members remained low on the Reading to Learn/Reading to Integrate Composite. As with the nonnative speakers, participants in the mid category on basic comprehension showed the most frequent reclassification. Forty-eight of the 101 (47.5%) native speaker participants remained in the initial classification categories. Twenty-two (21.8%) were reclassified into a higher category and 31 (30.7%) were reclassified into a lower category. Thus 53 participants (52.5%), over half of the sample, were classified differently based on their performance on the Reading to Learn/Reading to Integrate Composite.

[6]Reading to Learn correlated with function 2 at -.58; Reading to Integrate with function 1 at .71

V Interpretation

Analyses done to answer Research Questions 1 and 2 showed that performance on Reading to Learn and Reading to Integrate measures was significantly influenced by language background and level of education (graduate vs. undergraduate, for nonnative speakers only). Moreover, level of computer familiarity had no significant effect on Reading to Learn and Reading to Integrate performance.

Correlations showed that, as expected, all reading measures correlated to some degree, generally answering Research Question 3. The most interesting results came from the discriminant analyses because they showed a pattern of differential classification on the new tasks sensitive to initial reading ability. This pattern differed for nonnative speakers and native speakers. Examination of the reclassifications based on the new tasks revealed some of the problems of using a basic comprehension-only test to predict performance on more challenging literacy tasks. These results also imply the need for further work on new measures of advanced literacy skills such as Reading to Learn and Reading to Integrate to reflect trends in construct-driven assessment (Pellegrino *et al.*, 1999).

For the nonnative speakers, most (81.8%) of the participants with TOEFL Reading Comprehension below 50 remained classified as low on the Reading to Learn/Reading to Integrate Composite, suggesting the existence of a lower threshold of academic English proficiency. Participants below this threshold were unlikely to perform well on tasks assessing Reading to Learn and Reading to Integrate. Even tasks involving only selection (such as BC1 and BC2) rather than production of responses were difficult for this group. However, for the nonnative speaker readers in the mid and high reading ability groups, basic comprehension level was not nearly as consistent a predictor of performance on the Reading to Learn/Reading to Integrate Composite. This was especially striking for those in the mid reading ability group. Approximately one fourth of the mid group dropped into the low category on the Reading to Learn/Reading to Integrate Composite, indicating that they experienced more difficulty in completing the new measures, but approximately one fourth could indeed categorize and synthesize information relatively well despite middling performance on basic comprehension. This finding suggests that for nonnative speaker readers in the mid reading ability category, basic comprehension measures were insufficient to predict their performance on the new measures.

Almost two-thirds of the high basic comprehension nonnative speakers remained high on the Reading to Learn/Reading to Integrate Composite, but one third dropped, suggesting that they could not categorize and synthesize information from texts as well as they could recognize information. For this third, the recognition-only task, multiple-choice basic comprehension, overestimated ability to manipulate and synthesize information. For the high and mid reading ability nonnative speaker groups, basic comprehension-only tests may, then, overestimate academic English proficiency.

Results for the native speakers showed a more definitive pattern of reclassification. For the low reading ability group, approximately half were misclassified, suggesting that the basic comprehension measure underrepresented their ability to categorize and synthesize information in academic texts. Most of the mid reading ability group were reclassified as lower on the Reading to Learn/Reading to Integrate Composite, showing that basic comprehension results overestimated ability to succeed on more challenging tasks. On the other hand, almost one third of the mid reading ability group did better on the Reading to Learn/Reading to Integrate Composite; for them basic comprehension underrepresented their ability to complete more challenging reading tasks. For the high ability readers, nearly one third dropped on the Reading to Learn/Reading to Integrate Composite. For them, the basic comprehension measure overestimated their level of academic English proficiency. The other two thirds of the high reading ability group remained high, suggesting a higher threshold of academic English proficiency.

Considering results from both the nonnative speakers and native speakers, we conclude that the new tasks did assess something different from basic comprehension, once a lower level threshold of basic academic English proficiency had been achieved. Examination of scatterplots based on discriminant functions indicated an obvious separation between participants who could and could not perform on the Reading to Learn/Reading to Integrate Composite, thus providing some tentative evidence for concurrent validity (Messick, 1989; Chapelle, 1999). We had hoped to find clear evidence of a hierarchy suggesting that Reading to Learn was demonstrably more difficult than basic comprehension and Reading to Integrate demonstrably more difficult than Reading to Learn, but results did not yield an obvious hierarchy. Results did, however, suggest an even simpler pattern, a dichotomy. For those nonnative speakers above the lower threshold of English language proficiency, which in our data would be approximately 500 or 49 on the scaled TOEFL Reading

Comprehension scores, some could perform well on the new measures, some could not, revealing two rather than three groups on the Reading to Learn/Reading to Integrate Composite. These results lead us to speculate that the new measures tap additional skills such as 'sophisticated discourse processes and critical thinking skills' (Enright and Schedl, 1999: 24) in addition to language proficiency. For native speakers, the discriminant loading suggested a possible hierarchy of difficulty, but the scatterplots and classification tables revealed a dichotomy. Reclassification based on the Reading to Learn/Reading to Integrate Composite nearly eliminated the mid basic comprehension reading ability category, with only 16.8% of the native speakers qualifying as mid on the Reading to Learn/Reading to Integrate Composite.

VI Conclusions

While we acknowledge that there are some limitations to this project, such as the time required to administer and score new measures, the results are illuminating, useful, and suggest some considerations for future research. The first consideration relates to appropriate classification of students based on academic reading abilities. This corresponds to the role of TOEFL as a gatekeeper by many institutions. For TOEFL test-takers whose current TOEFL scores would be in an intermediate to high intermediate range, the new tasks could assess additional abilities relevant to academic performance. For such participants, additional test development efforts are warranted. The majority of high reading ability nonnative speakers (64.9%) could perform the new measures; a third could not. Our results suggest that some unknown number of mid reading ability nonnative speakers are more capable of succeeding at more challenging academic reading tasks than their current level of basic comprehension assessment would indicate. At the same time, there are some high and mid reading ability participants who could not perform the more demanding tasks; admitting such students directly into university programs could result in failure. These results suggest that for most students at lower levels of basic comprehension (in our study, those with TOEFL scores below 490 to 500), development of new tasks is unnecessary. Nevertheless, our results indicate that a certain small percentage (in our sample, 8 students, 18.2%) of nonnative speakers classified as low reading ability on basic comprehension did in fact

perform adequately on more challenging tasks like Reading to Learn and Reading to Integrate. Given the very large number of TOEFL test-takers around the world, this finding merits further research. It would be potentially unfair to exclude such students from university study based on basic comprehension-only measures such as the current TOEFL. We would urge institutions that use TOEFL to consider scores on these new, more demanding tasks if doing so would address institutional needs. According to interview data collected in conjunction with this project (Trites, 2000: Chapter 6), participants perceived that successful completion of the new tasks required a thorough understanding of the texts, whereas the multiple-choice tests were less demanding, requiring only superficial grasp of content. This finding corroborates the observation of Freedle and Kostin (1999: 3) who note that often examinees do not need to comprehend the accompanying text to answer the test item. Therefore, for all TOEFL test-takers, incorporating more challenging tasks such as the Reading to Learn and Reading to Integrate measures into typical English language instruction could have positive washback effects.

The second area of consideration is the focus on additional relevant research. For predictive validity, it is important to determine the correlation between the Reading to Learn/Reading to Integrate Composite measures and actual academic performance in university classes. Correlations might differ depending on the degree of categorization and synthesis required by different major fields of study. Another area for future research is related to the novelty of the Reading to Learn task as an assessment technique. Current pedagogical trends in literacy instruction emphasize the use of graphic organizers as a means to understand and manipulate information in texts. To our knowledge, graphic organizers are typically used as classroom activities, rather than assessment tools. They have good potential for use in assessment if students are familiar with them and if appropriate scoring systems can be developed. This project demonstrates that it is possible, though labor intensive, to develop reliable scoring systems. Further research is needed to explore refinement of this task type and related scoring systems for use in testing programs with large numbers of test-takers and scorers. Although the Reading to Integrate task (generating a synthesis) was more familiar, the scoring system was innovative because it reflected a reader's ability to recognize textual frames as well as integrate information. If test developers are interested in the abilities of nonnative speaker readers to perform such tasks, further development of similar tasks and scoring systems is warranted. While a substantial investment of time

would be required to refine the administration and scoring systems needed for more complex tasks such as these, it should be weighed against the possible danger of under-representing the likelihood of academic success based on results of the basic comprehension-only measures still most often used to assess academic proficiency.

Acknowledgements

This project was funded by a grant from Educational Testing Service as part of the TOEFL 2000 effort. We appreciate the cooperation of the ETS staff members who assisted in the development of passage specific tasks and other aspects of the research; however, no official endorsement of Educational Testing Service should be inferred.

VII References

Bachman, L. 2000: Modern language testing at the turn of the century: assuring that what we count counts. *Language Testing* 17, 1–42.

Biber, D. 1993: Using register-diversified corpora for general language studies. *Computational Linguistics* 19, 219–41.

Britt, M., Rouet, J. and **Perfetti, C.** 1996: Using hypertext to study and reason about historical evidence. In Rouet, J., Levonen, J., Dillon, A. and Spiro, R., editors, *Hypertext and cognition*. Mahwah, NJ: Lawrence Erlbaum, 43–72.

Chapelle, C. 1999: Validity in language assessment. *Annual Review of Applied Linguistics* 19, 1–19.

Educational Testing Service 1997: *TOEFL test and score manual*. Princeton, NJ: Educational Testing Service.

—— 1998: *Draft TOEFL 2000 research agenda framework: areas of research*. Research agenda three. TOEFL 2000 internal document. Princeton, NJ: Educational Testing Service.

Eignor, D., Taylor, C., Kirsch, I. and **Jamieson, J.** 1998: *Development of a scale for assessing the level of computer familiarity of TOEFL examinees*. TOEFL Research Report No. 60. Princeton, NJ: Educational Testing Service.

Enright, M. and **Schedl, M.** 1999: *Reading for a reason: using reader purpose to guide test design*. TOEFL 2000 Internal Report. Princeton, NJ: Educational Testing Service.

Enright, M., Grabe, W., Mosenthal, P., Mulcahy-Ernt, P. and **Schedl, M.** 1998: *A TOEFL 2000 framework for testing reading comprehension: a working paper*. Princeton, NJ: Educational Testing Service.

Foltz, P. 1996: Comprehension, coherence, and strategies in hypertext. In Rouet, J., Levonen, J., Dillon, A. and Spiro R., editors, *Hypertext and cognition*. Mahwah, NJ: Lawrence Erlbaum, 109–36.

Freedle, R. and **Kostin, I.** 1999: Does the text matter in a multiple choice test of comprehension? The case for the construct validity of TOEFL's minitalks. *Language Testing* 16, 2–32.

Goldman, S. 1997: Learning from text: reflections on the past and suggestions for the future. *Discourse Processes* 23, 357–98.

Hayes, J. and **Hatch, J.** 1999: Issues in measuring reliability: correlation versus percentage of agreement. *Written Communication* 16, 354–67.

Huberty, C. 1994: *Applied discriminant analysis*. New York: John Wiley.

Jamieson, J., Campbell, J., Norfleet, L. and **Berbisada, N.** 1993: Reliability of a computerized scoring routine for an open-ended task. *System* 21, 305–22.

Jamieson, J., Norfleet, L. and **Berbisada, N.** 1993: Successes, failures, and dropouts in computer-assisted language lessons. *Computer Assisted English Language Learning Journal* 4, 12–20.

Klecka, W. 1980: Discriminant analysis. In Lewis-Beck, M., editor, *Quantitative applications in the social sciences*. Volume 19. Newbury Park, CA: Sage.

Lehto, M., Zhu, W. and **Carpenter, B.** 1995: The relative effectiveness of hypertext and text. *International Journal of Human-Computer Interaction* 7, 293–313.

McNamara, D. and **Kintsch, W.** 1996: Learning from texts: effects of prior knowledge and text coherence. *Discourse Processes* 22, 247–88.

Messick, S. 1989: Validity. In Linn, R.L., editor, *Educational measurement*. 3rd edition. New York: American Council on Education, Macmillan, 13–103.

Meyer, B. 1985a: Prose analyses: purposes, procedures, and problems. In Britton, B. and Black, J., editors, *Understanding expository text*. Hillside, NJ: Lawrence Erlbaum, 11–64.

—— 1985b: Prose analysis: purposes, procedures, and problems. Part 2. In Britton, B. and Black, J., editors, *Understanding expository text*. Hillside, NJ: Lawrence Erlbaum, 269–97.

Monks, V. 1997, August/September: Two views, same waterway. *National Wildlife* 35, 36–37.

Pellegrino, J., Baxter, G. and **Glaser, R.** 1999: Addressing the 'two disciplines' problem: linking theories of cognition and learning with assessment and instructional practice. *Review of Research in Education* 24, 307–53.

Perfetti, C. 1997: Sentences, individual differences, and multiple texts: three issues in text comprehension. *Discourse Processes* 23, 337–55.

Perfetti, C., Britt, M.A. and **Georgi, M.** 1995: *Text-based learning and reasoning: studies in history*. Hillside, NJ: Lawrence Erlbaum.

Perfetti, C., Marron, M., and **Foltz, P.** 1996: Sources of comprehension failure: theoretical perspectives and case studies. In Cornoldi, C. and Oakhill, J., editors, *Reading comprehension difficulties: processes and intervention*. Mahwah, NJ: Lawrence Erlbaum, 137–65.

Reinking, D. 1988: Computer-mediated text and comprehension differences: the role of reading time, reader preference, and estimation of learning. *Reading Research Quarterly* 23, 484–500.

Reinking, D. and **Schreiner, R.** 1985: The effects of computer-mediated text on measures of reading comprehension and reading behavior. *Reading Research Quarterly* 20, 536–53.

Spivey, N. 1997: *The constructivist metaphor: reading, writing, and the making of meaning*. San Diego, CA: Academic Press.

Stevens, J. 1996: *Applied multivariate statistics for the social sciences*. 3rd edition. Mahwah, NJ: Lawrence Erlbaum.

Tabachnick, B. and **Fidell, L.** 1996: *Using multivariate statistics*. 3rd edition. New York: Harper Collins.

Taylor, C., Jamieson, J., Eignor, D. and **Kirsch, I.** 1998: *The relationship between computer familiarity and performance on computer-based TOEFL test tasks*. TOEFL Research Report No. 61. Princeton, NJ: Educational Testing Service.

Tennesen, M. 1997, November/December: On a clear day. *National Parks* 71, 26–9.

Trites, L. 2000: Beyond basic comprehension: reading to learn and reading to integrate for native and non-native speakers. Unpublished doctoral dissertation, Northern Arizona University, Flagstaff, AZ.

Van den Berg, S. and **Watt, J.** 1991: Effects of educational setting on student responses to structured hypertext. *Journal of Computer-Based Instruction* 18, 118–24.

Van Dijk, T.A. and **Kintsch, W.** 1983: *Strategies of discourse comprehension*. New York: Academic Press.

Wiley, J. and **Voss, J.** 1999: Constructing arguments from multiple sources: tasks that promote understanding and not just memory for text. *Journal of Educational Psychology* 91, 310–11.

Zimmerman, T. 1997, December 29: Filter it with billions and billions of oysters: how to revive the Chesapeake Bay. *US News and World Report* 123, 63.

Appendix 1a Chart completion task

Directions: Complete the following chart. Fill in as much detail as possible from the text read by categorizing the information into the different areas on the chart. Do not use single words for your responses; form your responses in phrases or complete sentences. Include examples from the text.

Make no judgments about the accuracy of causes or effects or the effectiveness of the solutions mentioned in the text. Solutions are seen

as any action taken in response to the problem(s). Solutions can take many forms such as proposed solutions, attempted solutions, or failed solutions. Also, space is provided under each category for examples. Examples are specific examples found in the text that are used by the author(s) to exemplify the problems, causes, effects, or solutions in the text. However, there may not be examples for every category.

Points will be awarded for correct responses only. There is no penalty for incorrect responses. Points will be awarded in the following manner.

Problems and Solutions	10 points each
Causes and Effects	5 points each
Examples	1 point each

Problems	Causes	Effects	Solutions
Examples	Examples	Examples	Examples

Appendix 1b Reading to Learn scoring rubric: 0–241 total possible

Problems (10 points) 1 word (6 points): 40 maximum	Causes (5points) 1 word (3 points): 55 maximum	Effects (5 points) 1 word (3 points): 30 maximum	Solutions (10 points end) 1 word (6 points): 90 maximum
• **Air pollution/Smog in the National Parks** • **Opposition**/or **Ignoring** (or lack of cooperation) to environmental standards (regulations) • No true 'point sources' (**myriad of smaller pollution sources** rather than one large source) • **National Parks limited jurisdiction** of pollution sources outside of the parks	• Sulfur/nitrogen emissions • Acid rain • Ground-level ozone ***Urban/Industrial emissions:*** • **Automobile** emissions • **Power plants**/factories/plants emissions • Emissions from **Smokestacks (chimneys)** • Emissions from **Kilns** • Unregulated pollution/smog (from other countries/Mexico) • Smoke from **Controlled burns** • Emissions from large cities	• **Trees and plants affected** (injured) growth hindered • **Visibility decreased** • **Metals loosened** into waters (surface/groundwater water polluted) • **Nutrients removed** (leached) from soil and/or plants • **Public outcry** over stance of big business/government • **Aquatic life** injured (damaged)	***Stricter Environmental Laws (resolutions, acts):*** • 1977 Clean Air Act **Amendments labeling** NPs as **Class I areas** • Regional haze **regulations** proposed by the **EPA** • **Reduction** of allowable **pollution standards** from industry • Clean Air Act set **visibility goals** • **Objecting** to **construction** permits • **Identification** of pollution **sources** • **Industry modifications/ Installation of devices** (scrubbers) to reduce pollution • **Public pressure** to protect parks

Appendix 1b (continued)

Examples (1 point) 5 maximum	**Examples (1 point or 5 points with overarching cause listed above)**	**Examples (1 point) 6 maximum**	**Examples (1 point or 10 points with overarching solution listed above)**
• Great **Smoky Mountain** National Park • **Grand Canyon** National Park • Smoky Mountain: Sources from Ohio, New York, Atlanta, etc. • Grand Canyon: Sources from CA, NV, UT, AZ, NM, and Mexico • TN Luttrell corp. building permit situation	• **Automobile** emissions • **Power plants**/factories/ plants emissions • Emissions from **Smokestacks (chimneys)** • Emissions from **Kilns** • Unregulated pollution/smog (from other countries/Mexico) • Smoke from **Controlled burns** • Emissions from large cities	• Leaves of plants turning purple and brown (stippling) • Hindering **photosynthesis** of plants and trees • **Reduction of visibility** at **Grand Canyon** • *Visibility reduced 90% of the days* • *See vague blue masses* • *See half as far as in 1919*	• 1977 Clean Air Act **Amendments labeling** NPs as **Class I areas** • Regional haze **regulations** proposed by the **EPA** • **Reduction** of allowable **pollution standards** from industry
	Examples (1 point): 10 maximum • Navajo Generating Station, Page, AZ (Power Plant, AZ) • Power plants in TN and OH river valley • Southern California Edison Plant, Laughlin, NV • Tennessee Luttrell Kilns, TN • Industry in AZ • Cars in CA • Smokestacks in NM, NV, UT • Los Angeles • Atlanta • New York		**Examples (1 point): 5 maximum** • Objecting to kilns in TN • Researchers using scientific technology to ID sources (radioactive isotopes) • Southern California Edison Plant in Laughlin, NV identified as point source • **Scrubbers** installed at Navajo Generating Plant • 10% reduction of pollution in 10–15 years (goal of no man-made pollution)

Appendix 2a Reading to Integrate task integration activity

Directions: For the next 15 minutes, reflect on what you read and compose a short essay that combines the information in the material read and makes connections across the range of ideas and arguments presented

Appendix 2b Reading to Integrate scoring rubric

Integration:

50	Excellent	Integrates texts accurately and successfully on multiple levels and creates a true Documents Model. Generalizes at least two macrostructure concepts common across texts (this may be simply identifying the existence of a macrostructure followed by the supporting macrostructures from each text). ***Effectively integrates relevant support*** (i.e., details or macrostructures) ***to support generalizations*** and may still discuss each article separately to some degree. Integrates macrostructures present in both texts.
40	Good	Integrates through the creation of a well developed and accurate introduction and/or a conclusion yet summarizes articles separately; AND/OR generalizes one major macrostructure that is fully developed ***with support***. Uses some relevant support.
30	Fair	Attempts to integrate through the creation of a partially developed, possibly inaccurate introductory AND/OR concluding statement usually related to the main topic of the articles, but has no substantive development. Creates a Text and Situation Model of each text separately. Does not make generalizations for integration beyond the one statement. May make evaluative or editorial statements; however, these may be inaccurate.
20	Poor	Creates a well developed and accurate Text Model and Situation Model for each of the texts, yet attempts no integrative connection across the texts.
10	Very Poor	Ineffectively or inaccurately attempts to summarize each article in a very reporting style, revealing little or no use of background knowledge, contextualization, or evaluation. Response is simply a recall of information from the texts. No introduction or conclusion. OR participant confuses texts and sees separate texts as one issue (problem).

0	No Response	Participant does not attempt response or only addresses one article.

Macrostructures:
Participants will receive one point for each macrostructure accurately identified in each text with a minimum score of 4 awarded for the inability to accurately identify any macrostructures.

25 (Total = 8)	Excellent	Accurately identifies all 4 macrostructures present in both texts
20 (Total = 6/7)	Good	Accurately identifies all 4 macrostructures in one text and 3 in the other OR accurately identifies 3 of the four macrostructures in both texts. (Or 4 in one and 2 in the other.)
15 (Total = 4/5)	Fair	Accurately identifies 3 of the four macrostructures in one text and 2 of the four in the other. (Or 4 in one and 1 in the other.) OR accurately identifies 2 of the four macrostructures in both texts. (Or 4 in one and 0 in the other, or 3 in one and 1 in the other.)
10 (Total = 2/3)	Poor	Accurately identifies 2 of the macrostructures in one text and 1 in the other. (Or 3 in one and 0 in the other.) Accurately identifies 1 of the four macrostructures in both texts. (Or 2 in one and 0 in the other.)
5 (Total = 0/1)	Very Poor	Accurately identifies 1 of the four macrostructures in one text and 0 in the other OR unable to accurately identify any macrostructures in the texts.
0	No Response	Participant does not attempt response.

Use of relevant details:

5	Excellent	Effectively uses multiple relevant details as support, no irrelevant or erroneous details.
4	Good	Effectively uses relevant details as support, may include some inaccurate or irrelevant details. (More than 50% of details are relevant)
3	Fair	Possibly uses one or two relevant details, yet several irrelevant details or inaccurate details erroneous or fabricated) appear in the synthesis. (50% or less of details are relevant)
2	Poor	Erroneous or fabricated details used in attempt to support arguments presented; does not include relevant details.
1	Very Poor	No details used or listed.
0	No Response	Participant does not attempt response