北市大語文學報 TMUE Journal of Language and Literature
1.6 (June 2011): 83-117

Evaluating the English Reading Comprehension Items of the SAET and the DRET

Chin-Ni Liu
Taipei Municipal University of Education
Wen-Ying Lin
Taipei Municipal University of Education

Abstract

This study aimed to evaluate the reading comprehension items of the Scholastic Ability English Test (SAET) and the Department Required English Test (DRET) from 2004 to 2008. Specifically, the study intended to answer the following three research questions: (1) What reading skills were measured on the SAET and the DRET reading comprehension sections and what was the percentage of the items for each of these skills identified? (2) How did the examinees in general perform on reading comprehension items measuring each of the reading skills on the SAET and the DRET? (3) For both tests across the five years, which reading skill identified could consistently best discriminate between the high scorers and the low scorers? For the purpose of answering the research questions, Nuttall's (2000) categorizations of reading skills and question types were mainly used as the coding scheme. Two experts in the field of English were invited as raters to classify each of the 134 reading comprehension items into one of the 11 reading skills. The results showed that six reading skills were identified on the SAET from 2004 to 2008, including (1) *Interpreting* (39.24%), (2) *Comprehending literal meaning* (25.32%), (3) *Reorganizing* (18.99%), (4) *Recognizing implications and inferences* (7.59%), (5) *Recognizing functional value* (6.33%), and (6) *Recognizing and interpreting cohesive devices* (2.53%). As for the DRET, the same six reading skills were also identified along with one more sub-skill, *Recognizing style and tone*. The respective percentages of the seven reading sub-skills identified on the DRET were: *Interpreting* (40%), *Recognizing implications and inferences* (18.18%), *Reorganizing* (16.36%), *Comprehending literal meaning* (12.73%), *Recognizing and interpreting cohesive devices* (5.45%), *Recognizing functional value* (3.64%), and *Recognizing style and tone* (3.64%). The SAET takers performed best on the *Comprehending literal meaning* items, but worst on the *Recognizing functional value* items, whereas the DRET takers performed best on the *Recognizing functional value* items, but worst on the *Recognizing style and tone* items. Furthermore, the examinees generally performed better on the SAET than those on the DRET, in terms of the mean passing rate for each of the reading skills identified. Finally, none of the reading skills could consistently best discriminate the high scorers from the low scorers for both tests across the five years.

Key words: reading comprehension, Bloom's taxonomy, Nuttall's categorizations of reading skills, interactiveness, construct validity

INTRODUCTION

The College Entrance Examination in Taiwan, a two-stage testing system, has been implemented every year by the College Entrance Examination Center (CEEC) to serve as a nationwide college placement test. The major purpose of this examination is to determine which university each third-year senior high school student will be admitted to. In the first stage, students are required to take the Scholastic Ability Test (SAT) in late January or early February, which aims to evaluate whether students have acquired the basic scholastic knowledge and abilities for college education. The second-stage test, the Department Required Test (DRT), is intended to identify those students who perform well in certain subject areas required by university departments. With such a purpose, the DRT tends to focus on the assessment of students' higher-order cognitive abilities, such as judgment, inference, and analysis (Yin, 2005).

Both the SAT and the DRT include the assessment of students' English achievement since English is commonly taught as an academic subject in Taiwan's senior high schools. One similarity shared by both the English Achievement Test of the SAT (abbreviated as SAET) and that of the DRT (abbreviated as DRET) is that students' reading comprehension is one of the main components in assessing their English reading ability. In both tests, a reading comprehension section consists of a series of passages, each followed by three to five multiple-choice questions. These questions are intended to assess an array of different reading sub-skills.

With regard to the reading sub-skills, several studies (Fan, 2008; Hsu, 2005; Lan, 2007; Lu, 2002) have been conducted to categorize the reading comprehension questions on both the SAET and the DRET into different reading sub-skills or question types. For example, using Mo's (1987) classification of question types, Lu (2002) attempted to categorize each test item of the SAET from 1995 to 2002 into various question types designed to measure different reading sub-skills. Likewise, Lan (2007) analyzed and categorized the reading comprehension questions on the SAET and the DRET into various reading sub-skills based on the revised Bloom's (Anderson & Krathwohl, 2001) Taxonomy. However, Mo's classification contains only six categories of question types, while the revised Bloom's Taxonomy is not specifically designed for language learning. As such, further research is warranted to evaluate the test items of the SAET and the DRET using categorizations that not only are specifically intended for language learning but also include a fairly extensive list of reading sub-skills.

Given that the test scores of the two high-stake tests have a tremendous impact on students' future study in university, results obtained from research along this line can be of great value to English teachers in Taiwan's senior high schools. Specifically, with the results of this study, they can better understand what specific reading sub-skills are most needed for students to achieve high test scores on the SAET and

the DRET. That said, they can then design or modify their reading instructions or teaching materials accordingly. In addition, the results of this study can also provide SAET and DRET constructors with information on whether or not certain reading sub-skills are over-represented or under-represented in the tests.

LITERATURE REVIEW

Defined by Almasi (2003) as "the ability to understand and construct meaning from what one reads" (p. 74), the construct of reading comprehension is generally viewed as a group of receptive skills. With its unobservable nature, one cannot see the process of reading, nor can one observe a specific product of reading. Therefore, the challenge for language test writers has always been to construct test tasks which will not only cause test takers to exercise reading, but also result in behavior that will demonstrate successful reading. To deal with the challenge, language test writers often believe in the multi-dimensional nature of the reading comprehension construct and translate it into various reading sub-skills, which are usually based on taxonomies or categorizations proposed by researchers in the related fields. The following describes taxonomies or categorizations used to construct or evaluate reading comprehension tests.

Taxonomies / Categorizations

Bloom's Taxonomy of Educational Objectives in the Cognitive Domain, being widely influential in the classroom instruction and language tests, encompasses the following six major categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation (Bloom, Engelhart, Frost, Hill, & Krathwohl, 1956). Except for the Knowledge category, the remaining types are labeled as abilities or skills. Five of the six categories comprise sub-categories (see Table 1). These categories are hierarchically arranged from simple and concrete entities to complex and abstract constructs. The mastery of simple categories is a prerequisite for the advancement into the complex constructs (Krathwohl, 2002; Kreitzer and Madaus, 1994). However, this taxonomy is questioned with regard to this hierarchical structure. Some demands for the sub-categories under Knowledge level appear more complex than certain demands for those under the Analysis or Evaluation levels. Similarly, some demands for the sub-categories under the Evaluation level seem less complex than those under the Synthesis level, because several researchers, such as Kreitzer and Madaus (1994), believe that the Synthesis level in fact also involves evaluation.

Table 1
Structure of the Original Bloom's Taxonomy

1.0 Knowledge
- 1.10 Knowledge of specifics
  - 1.11 Knowledge of terminology
  - 1.12 Knowledge of specific facts.
- 1.20 Knowledge of ways and means of dealing with specifics
  - 1.21 Knowledge of convention
  - 1.22 Knowledge of trends and sequences
  - 1.23 Knowledge of classifications and categories
  - 1.24 Knowledge of criteria
  - 1.25 Knowledge of method0logy
- 1.30 Knowledge of universals and abstractions in a field
  - 1.31 Knowledge of principles and generalization
  - 1.32 Knowledge of theories and structures

2.0 Comprehension
- 2.1 Translation
- 2.2 Interpretation
- 2.3 Extrapolation

3.0 Apply

4.0 Analyze
- 4.1 Analysis of elements
- 4.2 Analysis of relationships
- 4.3 Analysis of organizational principles

5.0 Synthesis
- 5.1 Production of a unique communication
- 5.2 Production of a plan, or proposed set of operations
- 5.3 Derivation of a set of abstract relations

6.0 Evaluation
- 6.1 Evaluation in terms of internal evidence
- 6.2 Judgments in terms of external criteria

Note. Adopted from "A Revision of Bloom's Taxonomy: An Overview," by D.R. Krathwohl, 2002, *Theory into Practice*, *41*(4), p.213.

Nowadays, meaningful learning is deemed as one of very important educational goals from the constructivist perspective. That is, with meaningful learning, students tend to engage themselves in active knowledge processing and meaning construction of their selective information through integration with their existing knowledge (Mayer, 2002). What learners know (knowledge) and how they think (cognitive processing) are thus highly emphasized in constructivist learning (Anderson and

Krathwohl, 2001). Learners' acquired knowledge enables teachers to know what to teach, whereas their cognitive processing provides teachers with information on ways to help them retain and transfer their acquired knowledge. Based on the above constructivist position, Anderson and Krathwohl (2001) revised Bloom's Taxonomy and divided the framework into two dimensions: the knowledge dimension and the cognitive process dimension. The knowledge dimension entails four types of knowledge—factual knowledge, conceptual knowledge, procedural knowledge, and metacognitive knowledge. Under each type of knowledge, a number of subtypes are also listed. For the cognitive process dimension, the revised taxonomy encompasses six categories or levels: Remember, Understand, Apply, Analyze, Evaluate, and Create. Each of the six levels also includes its sub-categories (see Table 2).

One thing to note is that both the original and the revised Bloom's Taxonomies are not specifically developed for language learning, though they have been used by several language researchers (e.g., Chern, 2006; Lan, 2007; You, 2004) in Taiwan to evaluate reading comprehension items of some nationwide English entrance examinations. There are several other categorizations that particularly aim at language learning, such as the categorizations by Mo (1987) and by Nuttall (2000). Each of the two classifications is described in the following.

The taxonomy proposed by Mo (1987) focuses on what cognitive strategies or abilities are involved in language test tasks. Specifically, he claimed that a reading test should include questions that assess textual comprehension and questions that require test takers to clarify the organization of the text. Accordingly, he came up with a classification of reading sub-skills, which includes the following six categories: (1) identifying the main idea, (2) comprehending literal meaning, (3) finding implications

Table 2

Structure of the Cognitive Process Dimension of the Revised Bloom's Taxonomy

1.0 Remember—Retrieving relevant knowledge from long-term memory
1.1Recognizing
1.2 Recalling
2.0 Understand—Determining the meaning of instructional messages, including oral, written and graphic communication
2.1 Interpreting
2.2 Exemplifying
2.3 Classifying
2.4 Summarizing
2.5 Inferring
2.6 Comparing
2.7 Explaining
3.0 Apply—Carrying out or using a procedure in a given situation

3.1 Executing
3.2 Implementing
4.0 Analyze—Breaking material into its constituent parts and detecting how the parts relates to one another and to an overall structure or purpose
4.1 Differentiating
4.2 Organizing
4.3 Attributing
5.0 Evaluate—Making judgments based on criteria and standards
5.1 Checking
5.2 Critiquing
6.0 Create—Putting elements together to form a novel, coherent whole or make an original product
6.1 Generating
6.2 Planning
6.3 Producing

Note. From "A Revision of Bloom's Taxonomy: An Overview," by D.R. Krathwohl, 2002, *Theory Into Practice*, *41*(4), p.215.

and drawing inferences and conclusions from the text, (4) recognizing style and tone, (5) clarifying text organization and cohesive devices, and (6) determining the meaning of words or phrases in the text.

More recently, Nuttall (2000) suggested an extensive list of reading sub-skills. In her list, she further classified them into two kinds of text-attack skills: skills necessary to read for plain sense and skills necessary to read beyond plain sense. According to Nuttall, the skills necessary to read for plain sense belong to bottom-up strategies which encompass (1) understanding the syntax, (2) recognizing and interpreting cohesive devices, and (3) interpreting discourse markers. The skills necessary to read beyond plain sense pertain to top-down strategies. They include (1) recognizing functional value, (2) recognizing text organization, (3) recognizing the presuppositions underlying the text, (4) recognizing implications and making inferences, and (5) predicting. In addition to these two kinds of text-attack skills, Nuttall classified most reading comprehension questions into the following six types: (1) questions of literal comprehension, (2) questions involving reorganization or interpretation, (3) questions of inference, (4) questions of evaluation, (5) questions of personal response, and (6) questions concerned with how writers say what they mean.

Taken together, of the four taxonomies or categorizations mentioned above, the last two categorizations proposed by Mo (1987) and by Nuttall (2000) are specifically developed for language learning. Furthermore, several sub-skills, such as the skill of recognizing implications and making inferences, are included in both categorizations. Finally, when language test writers construct reading comprehension questions or

when language test evaluators analyze, evaluate, or validate reading comprehension questions of nationwide entrance examinations, the four categorizations have been used as a sort of framework, together with some of the criteria that will be described in the next section.

Criteria Used to Evaluate Language Tests

For decades, the four criteria commonly used to evaluate a language test are reliability, practicality, washback, and validity (Hughes, 2003). Each of the four criteria has played an important role in both developing a language test and evaluating an existing assessment procedure. However, in 1996, Bachman and Palmer added two more criteria and proposed a model of test usefulness for designing and evaluating language tests. They believed that the most important quality of a test is its usefulness, which can be described as a function of six different qualities, including reliability, practicality, impact (washback), construct validity, authenticity, and interactiveness.

The last two qualities are the added criteria. The first added criterion, authenticity, according to Bachman and Palmer (1996), means "the degree of correspondence of the characteristics of a given test to the features of a target language use (TLU) task: task that the test taker is likely to encounter outside the testing situation, and to which we want our inferences about language ability to generalize" (p.23). If a test task is authentic, then this task is likely to be acted out in the real world. Take a reading comprehension test for example. In general, the text is a crucial part in any reading comprehension test. Therefore, if the topical content of each text in the reading test matches the kinds of topics that the test taker may read outside the testing situation, then we can assume that this reading test/task is authentic. The other added criterion, interactiveness, as defined by Bachman and Palmer, is "the extent and type of involvement of the test taker's individual characteristics in accomplishing a test task" (p.25). According to Bachman and Palmer, there are three aspects of the test taker's individual characteristics that are most relevant for language testing: language ability (including language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata. Therefore, if a language test has a high degree of interactiveness, then the test taker's areas of language ability, topical knowledge, and/or affective schemata are engaged when he/she takes the test. On the contrary, if a language test lacks interactiveness, then the test taker may get points simply by using his/her common sense rather than his/her language ability, topical knowledge, and/or affective schemata.

Furthermore, Bachman and Palmer (1996) confined the criterion of validity to construct validity, which refers to the degree to which scores on an assessment instrument permit inferences about its underlying trait(s). Specifically, construct validity consists of two aspects. First, it pertains to "the meaningfulness and appropriateness to which we can interpret a given test score as an indicator of the

ability or construct that we intend to measure" (p.21). The second aspect of construct validity deals with the generalization of the test score to the TLU domain that the test tasks correspond to. That is to say, we want our interpretations of test score about language ability (construct) to generalize beyond the testing situation itself to a particular TLU domain.

Following Bachman and Palmer's definition, one can easily see that interactiveness is closely related to the first aspect of construct validity, which concerns the meaningfulness and appropriateness of a given test score. If a language test has a high degree of interactiveness, then the test taker should be required to use his/her language ability, topical knowledge, and/or affective schemata when s/he takes the test. If any of these three aspects of interactiveness engaged by the test taker is what the language test intends to measure, then one can make inferences, based on his/her performance (i.e., the test score), about the targeted language ability, topical knowledge, and/or affective schemata of the test taker. In other words, if a test taker employs the targeted or intended language ability, topical knowledge, and/or affective schemata while taking the language test, the test then is said to have a certain degree of interactiveness, which would in turn lend some evidence to the construct validity of the test. It is in this sense that interactiveness is linked with construct validity.

Studies on Evaluation of Reading Comprehension Items

Among the three aspects (i.e., language ability, topical knowledge, and affective schemata) of interactiveness, language ability has been used implicitly as a criterion to analyze, evaluate, or validate the reading comprehension questions of some nationwide English entrance examinations in Taiwan. Take, for example, the reading comprehension items of the Basic English Competence Test (abbreviated as BECT) at junior high school level, which is held in late May and mid July each year. The test objectives of the BECT are based on the Core Competence Indicators of the Grades 1-9 Curriculum Guidelines. On the BECT, the test items range from 40 to 45 and the test format consists of multiple-choice items only. Among the test items, about 15 to 20 items measure test takers' knowledge of vocabulary, phrases, and grammar, and around 20 to 25 items measure their reading comprehension. In a study that analyzed all the reading comprehension items of the BECT from 2001 to 2003, You (2004) concluded that the BECT items had high validity because each item of the BECT can be categorized into Bloom's taxonomy. However, it seems that You's conclusion may not be appropriate because Bloom's taxonomy is not specifically aimed at language learning. Other categorizations of reading skills that are particularly for language learning, such as Nuttall's categorizations of reading skills, should have been used instead.

Therefore, a more resent study was conducted by Chern (2006) to find out what cognitive or reading skills were involved in the reading comprehension items of the

BECT from 2001 to 2006 by using not only Bloom's learning taxonomy but also Nuttall's categorization of text-attack skills. The results of her study showed that the majority of the test items fell into the first two levels in Blooms Taxonomy, i.e., Knowledge and Comprehension, and that the sixth level (i.e., Evaluation) in the taxonomy was not tested. As to Nuttall's categorization, she found that more items of the BECT measured top-down than bottom-up reading skills.

Likewise, the reading comprehension items of the SAET and the DRET at the senior high school level have also been evaluated and validated implicitly by the criteria of the language ability aspect of interactiveness. Both the SAET and the DRET contain two major parts: one is multiple-choice items and the other is non-multiple-choice items. Multiple-choice items intend to measure test takers' vocabulary, grammar, and reading comprehension, whereas non-multiple-choice items aim to measure test takers' writing ability. In general, non-multiple-choice items include English translation and essay writing.

To date, several studies (Fan, 2008; Hsu, 2005; Lan 2007; Lu, 2002) have been conducted to examine the test by categorizing each of the reading comprehension items into various different question types or skills. For instance, Lu (2002) ran a study in 2002 to classify each reading comprehension item of the SAET from1995 to 2002 into various question types using Mo's (1987) classification of question types. Her study showed that the items on details (56%) were the major question type category whereas the organization items (1%) were the minor category. She also computed the mean passing rate for each question type for the eight years. Her results showed that the examinees generally performed best on the word meaning items (mean passing rate = 55.04%) and performed worst on the style/tone items (mean passing rate = 27%).

Unlike the study of Lu (2002), a more recent study by Lan (2007) analyzed the items of the SAET and the DRET from 2002 to 2006 by applying the revised Bloom's Taxonomy. Her results indicated that the reading comprehension items on both tests only measure the following four lowest cognitive levels of the revised Bloom's Taxonomy: Remember (41%), Understand (46%), Apply (4%), and Analyze (9%). She also found that the question type preferred was different between the SAET and the DRET. Specifically, Executing questions (i.e., questions requiring test takers to use a procedure to carry out a familiar task) were more common on the SAET, whereas Inferring questions (i.e., questions requiring test takers to draw a logical conclusion from presented information) were favored on the DRET. Moreover, Lan also compared the performance between the high scorers and the low scorers on each question type. She concluded that it was hard to determine which type of question could best discriminate the high scorers from the low scorers on both the SAET and the DRET because no significant effect was found in the major question types on the discrimination index.

To sum up, the BECT, the SAET, and the DRET have been examined with the implicit use of the criterion of interactiveness. With regard to the BECT, the original Bloom's Taxonomy of educational objectives in the cognitive domain has been used by both You (2004) and Chern (2006). However, this taxonomy is not specifically developed for language use. As such, in addition to the original Bloom's Taxonomy, Chern (2006) also applied Nuttall's categorization of text-attack skills. As for the SAET and the DRET, Lu (2002) used Mo's classification, which contains only six categories of reading sub-skills, and Lan (2007) employed the revised Bloom's Taxonomy, which is not specifically aimed at language learning. Hence, further research is warranted to evaluate the items of the SAET and the DRET by using categorizations that not only are specifically intended for language learning but also include a fairly extensive list of reading sub-skills. For the purpose of knowing more clearly about to what extent the reading comprehension questions of the SAET and the DRET are interactive, a taxonomy that specifically aims at language learning and includes a relatively extensive list of sub-skills, such as Nuttall's taxonomy, should be used instead. Moreover, in Lan's study, she invited two graduate students rather than domain-specific experts as raters to categorize each of the test items into various reading skills. The credibility of Lan's study may therefore tend to be low. Hence, the need is warranted to conduct a study that evaluates and validates the reading comprehension questions of the SAET and the DRET by adopting Nuttall's taxonomy of reading skills and including experts' judgments.

RESEARCH QUESTIONS

The purpose of the study was mainly to evaluate and validate the reading comprehension questions of the SAET and the DRET from 2004 to 2008. Specifically, the present study intended to answer the following research questions: (1) To what extent were the reading comprehension questions of the SAET and the DRET interactive, from the aspect of language ability involved? That is, what reading skills were involved and measured on the SAET and the DRET reading comprehension items and what was the percentage of the items for each of the skills identified? (2) How did the examinees in general perform on the reading comprehension questions measuring each of the reading skills identified on the SAET and the DRET? (3) For both tests across the five years, which reading skills identified could consistently best discriminate between the high scorers and the low scorers?

METHOD

For the purpose of answering the three research questions, two experts in the field of English were invited as raters to analyze and evaluate the reading comprehension items on the SAET and the DRET by classifying each of the test items into 11 reading skills, which were mainly based on Nuttall's (2000) categorizations of reading skills and question types. The test items evaluated, the domain-specific raters involved, the instrument used, the data collected, and the coding procedures followed are briefly described in the following.

The Test Items

This study included a total of 35 reading passages and 134 reading comprehension items of the SAET (20 passages with 79 items) and the DRET (15 passages with 55 items) from 2004 to 2008. Table 3 presents, for each of the five years, the topic categorizations and the number of reading passages, and the number of the test items that were evaluated.

Table 3

The Topic Categorizations and the Number of Reading Passages and the Number of Test Items

Year	SAET			DRET		
	No. of reading passages	No. of test items	Topic	No. of reading passages	No. of test items	Topic
2004	4	15	Medicine Culture Animals Education	3	11	Sports Science Art
2005	4	16	Education Nature Health Animals	3	11	Art Communication Business
2006	4	16	Health Culture Technology Language	3	11	Environment Animals Ethics
2007	4	16	Health Business Animals Education	3	11	Literature Medicine Ethics

2008	4	16	Animal Nature Professions Medicine	3	11	Business Medicine History
total	20	79		15	55	

Two Domain-specific Raters

Two female professors from the Department of English Instruction at one public university in northern Taiwan were recruited to serve as the domain-specific raters of the study. Both professors had more than eight years of English teaching experience in senior high schools. The first rater (Rater A) graduated from the National Taiwan Normal University with a PhD in English Literature; the other rater (Rater B) graduated from Fu Jen Catholic University with a PhD in Comparative Literature.

The Instrument

The instrument used in this study was the coding scheme sheets administered to the raters to classify the 134 reading comprehension items into 11 categories of reading skills. The coding scheme sheets included the definition and the example item for each skill. Table 4 presents the 11 reading skills used on the coding scheme sheets. As shown in Table 4, Nuttall's (2000) categorization of text-attack skills was used in the present study as the major framework for raters because her categorization not only is specific for language learning but also is quite extensive. As stated earlier, Nuttall's extensive list of the text-attack skills includes three bottom-up skills (i.e., *Understanding syntax*, *Recognizing and interpreting cohesive devices*, and *Interpreting discourse markers*) and five top-down skills (i.e., *Recognizing functional value*, *Recognizing text organization*, *Recognizing presuppositions*, *Recognizing implications and inferences*, and *Predicting*). However, a close examination of her descriptions about the three bottom-up skills indicated that her first bottom-up skill, *Understanding syntax*, tends to be very broad, and overlaps her remaining two bottom-up skills (*Recognizing and interpreting cohesive devices* and *Interpreting discourse markers*). Therefore, the skill *Understanding syntax* was dropped from the framework of the present study. Furthermore, after a preliminary check of the 134 items on the SAET and the DRET, it was found that Nuttall's categorization of text-attack skills still was not comprehensive enough. As such, two (out of six) categories from Nuttall's classification of question types were also included in the framework of the present study: *Questions of literal comprehension* and *Questions involving reorganization or reinterpretation*. As each category of Nuttall's text-attack skills was named with a gerund in the beginning, the two question types included in the present study were therefore renamed. As such, "*Questions of literal comprehension*" was renamed as "*Comprehending literal meaning"*, and "*Questions*

involving reorganization or reinterpretation" was renamed as "*Reorganizing*".

Table 4
The 11 Reading Skills Used on the Coding Sheet

Reading skill	Code		Derived from
1. Comprehending literal meaning	LM	Bottom-up	Nuttall's question type
2. Recognizing and interpreting cohesive devices	CD	Bottom-up	Nuttall's text-attack skill
3. Interpreting discourse markers	DM	Bottom-up	Nuttall's text-attack skill
4. Recognizing functional value	FV	Top-down	Nuttall's text-attack skill
5. Recognizing text organization	TO	Top-down	Nuttall's text-attack skill
6. Recognizing presuppositions	PS	Top-down	Nuttall's text-attack skill
7. Recognizing implications and inferences	IF	Top-down	Nuttall's text-attack skill
8. Predicting	PD	Top-down	Nuttall's text-attack skill
9. Reorganizing	RO	Top-down	Nuttall's question type
10. Interpreting	IT	Top-down	revised Bloom's taxonomy
11. Recognizing style and tone	ST	Top-down	Mo's taxonomy

Taken together, a total of nine categories of reading skills from Nuttall's (2000) classification of text-attack skills and her categorization of question types were used as the main framework in the present study for the raters to classify the 134 reading comprehension questions of the SAET and the DRET. The nine categories of reading skills included Nuttall's two bottom-up text-attack skills, five top-down text-attack skills, and two question types.

However, as shown in Table 4, in addition to the nine categories, two more reading skills from other taxonomies were also included in the framework of the coding scheme sheets because, after a further check of the 134 items on the SAET and the DRET, some items were found to measure the skills that were not yet included in the categorizations by Nuttall (2000). The two added skills were the skill "*Interpreting*," which is one of the sub-categories from the second level (Understand) of the revised Bloom's taxonomy (Anderson & Krathwohl, 2001), and the skill "*Recognizing style and tone*," which is derived from Mo's (1987) taxonomy. The former skill "*Interpreting*" refers to the test-taker's ability to identify a restatement of a sentence or a passage. The latter skill "*Recognizing style and tone*" pertains to the test-taker's ability to recognize the author's tone, mood, voice, attitude, or the text style. The two skills were added in the framework of this study.

Consequently, as listed in Table 4, at the final stage a total of 11 reading skills were employed in the coding scheme in the present study. They were *Comprehending literal meaning* (LM), *Recognizing and interpreting cohesive devices* (CD), *Interpreting discourse markers* (DM), *Recognizing functional value* (FV),

Recognizing text organization (TO), *Recognizing presuppositions* (PS), *Recognizing implications and inferences* (IF), *Predicting* (PD), *Reorganizing* (RO), *Interpreting* (IT), and *Recognizing style and tone* (ST). The definition for each skill is presented in the Appendix.

The Data Collected

The data collected in this study included three parts. The first part was the passing rates of all examinees for each reading comprehension question on the SAET and the DRET from 2004 to 2008. The second part of the data contained 10 sets. The first five sets of data were responses from a group of 5000 randomly-selected examinees to each item on the SAET each year from 2004 to 2008. Similarly, the other five sets of data were responses from a group of 5000 randomly-selected examinees to each item on the DRET each year from 2004 to 2008. Both parts of the data were provided by the CEEC. The last part of the data was the raters' coding. The 11 skills were numbered from 1 to 11. For each item the raters would assign a number from 1 to 11 after they had decided the category of skill that each item attempted to measure.

The Coding Procedures

In this study, every reading comprehension question was coded by the two domain-specific experts. Prior to the formal coding, a rater training was arranged. During the training phase, the two domain-specific raters first read over the coding scheme sheets (see Appendix). The raters then practiced coding together. The reading comprehension questions of the SAET and the DRET for 2002 were provided as practice items. Both tests included 15 items. Based on the coding scheme sheets, the two raters categorized each item into one of the 11 reading skills. Then the raters practiced coding independently the reading comprehension questions of the SAET and the DRET for 2003. Each of the two tests for every year contained 15 items. Similarly, they classified each item into one of the 11 reading skills. If there was a disagreement between the two raters, a consensus-building discussion was then followed. Finally, Cohen's Kappa, which is one type of inter-rater reliability index, was computed. Specifically, the inter-rater reliability between the two raters was 100 percent for the SAET and 90.73 percent for the DRET.

After the inter-rater reliability was calculated, the formal coding was then implemented. All the reading comprehension questions from 2004 to 2008 were coded by the two raters. The correct answers to the reading comprehension questions of the SAET and the DRET from 2004 to 2008 were provided for raters as reference. Similar to the training phase, if there was a disagreement between the raters, a consensus-building discussion was followed. The final results were obtained after 100% agreement had been reached through the discussion.

Calculation of the Mean Passing Rate for Each Reading Sub-skill

To examine how the examinees in general performed on the reading comprehension questions that measured each of the reading skills on the SAET and the DRET, the mean passing rate for each item measuring each reading skill was calculated. The passing rate refers to the proportion of the test takers who answered an item correctly. For example, an item with 50 percent passing rate means that 50 percent of the test takers answered the item correctly. The mean passing rate for each reading skill was obtained by summing up all passing rates and then dividing that number by the number of items identified for that reading skill. Take, for example, *Comprehending literal meaning* items. Six items were identified to measure this skill in the 2004 SAET. The passing rates for the six items measuring the skill were 48, 50, 52, 83, 73, and 77. To obtain the mean passing rate for the skill *Comprehending literal meaning* items, the six passing rates for the six items were summed up and then the number (383) was divided by six. The obtained value, 63.83, was the mean passing rate for *Comprehending literal meaning* items in the 2004 SAET. If there was only one item identified to measure a certain skill, the mean passing rate for that skill would be the passing rate for that single item. Similarly, to compare the mean passing rates for each reading skill across the five years, the average of the mean passing rates over the five years was calculated. The average was obtained by summing up the mean passing rates of a particular skill from 2004 to 2008 and then dividing that number by the number of years (i.e., five). Again take, for example, *Comprehending literal meaning*. The mean passing rates for this skill from 2004 to 2008 were 63.83, 67.33, 60.75, 58.33, and 68.75 respectively. To obtain the average of the mean passing rates for the skill, the mean passing rates for this skill over the five years were summed up and then the number (318.99) was divided by five. The obtained value, 63.80, was the average of the mean passing rates for *Comprehending literal meaning* from 2004 to 2008 on the SAET.

Calculation of the Mean Discrimination Index for Each Reading Sub-skill

To understand how well the items identified for each skill can discriminate between high and low scorers, the mean discrimination indices were also computed and examined. The discrimination index D_i (where i = 1, 2, 3, 4, 5, 6) for item i was obtained by subtracting the passing rate (*PH*) for the low scorers from that (*PI*) for the high scorers. The mean D for items measuring one particular reading skill was obtained by summing up the $D_i's$ and dividing the number by the number of items identified for that reading skill. Take, for example, the skill *Comprehending literal meaning*. Six items were identified to measure this skill in the 2004 SAET. Suppose the D_i for each of the six items was 55, 63, 31, 42, 53, and 47. To obtain the mean D

for the skill *Comprehending literal meaning*, the D_i's for the six items were summed up and then the number (291) was divided by six. The obtained value, 48.50, was the mean for *Comprehending literal meaning* items in the 2004 SAET. Furthermore, the average of the means for each reading skill over the five years was also calculated. The average was obtained by summing up the D's for a particular skill from 2004 to 2008 and dividing the number by the number of years. Again take *Comprehending literal meaning* for example. The means for this skill from 2004 to 2008 were 48.50, 59.67, 62.75, 56.67, and 61.00 respectively. To obtain the average of the means for the skill *Comprehending literal meaning*, the D's for this skill over the five years were summed up and then the number (288.59) was divided by five. Hence, the obtained value, 57.72, was the average of the means for *Comprehending literal meaning* over the five years from 2004 to 2008 on the SAET.

RESULTS AND DISCUSSION

Reading Skills Measured on Both Tests

In this study, a total of 134 reading comprehension items for the SAET (79 items) and the DRET (55 items) over the five years from 2004 to 2008 were evaluated by the two raters. The percentage of the items identified for each of the 11 reading skills is presented in Table 5. As seen from Table 5, the 134 items were classified into seven (out of the 11) reading skills. The seven skills were *Comprehending literal meaning*, *Recognizing and interpreting cohesive devices*, *Recognizing functional value*, *Recognizing implications and inferences*, *Interpreting*, *Reorganizing*, and *Recognizing style and tone*. Of the seven reading skills identified, *Interpreting* (53 items or 39.55%) was the most frequently measured skill on the SAET and the DRET. The second most frequently measured skill was *Comprehending literal meaning* (27 items or 20.15%), followed by *Reorganizing* (24 items or 17.91%) and *Recognizing implications and inferences* (16 items or 11.94%). The three least frequently measured skills were *Recognizing functional value* (7 items or 5.22%), *Recognizing and interpreting cohesive device* (5 items or 3.73%), and *Recognizing style and tone* (2 items or 1.5%).

The remaining four skills, *Interpreting discourse markers*, *Recognizing text organization*, *Recognizing presuppositions*, and *Predicting*, were not identified on either tests. Several plausible reasons may explain why these four skills were not identified. First of all, the skill *Interpreting discourse markers* is usually tested on the cloze section of the two tests. Similarly, the skill *Recognizing text organization* is normally assessed in the discourse structure section of the DRET. The items measuring the skill *Recognizing presuppositions*, they tend to measure the background knowledge or experience that the test writer expects examinees to have in order to answer them correctly. However, it is hard to be certain that all examinees possess the

particular background knowledge. Hence, this type of question is normally not included on the two tests for the sake of fairness. As to the failure of locating items measuring the skill *Predicting* (i.e., the ability to predict what is likely to come next and what is not), the reason is not quite clear, given the ease and importance of constructing this question type. This finding may serve as a reminder for future SAET and DRET item constructors to include items measuring the sub-skill *Predicting*.

Table 5
Percentage of the Items Identified to Measure Each of the 11 Reading Skills

Reading skill	No. of items	% of total
Interpreting	53	39.55
Comprehending literal meaning	27	20.15
Reorganizing	24	17.91
Recognizing implications and inferences	16	11.94
Recognizing functional value	7	5.22
Recognizing and interpreting cohesive devices	5	3.73
Recognizing style and tone	2	1.50
Interpreting discourse markers	0	0
Recognizing text organization	0	0
Recognizing presuppositions	0	0
Predicting	0	0
Total	134	100.00

Table 6 shows the respective percentage of the items on the SAET and the DRET categorized into each reading skill. Of the 79 items on the SAET, 31items (39.24%) were identified as measuring *Interpreting*, which was the most frequently measured sub-skill, and *Recognizing and interpreting cohesive devices* was the least frequently measured sub-skill with only 2 items (2.53%) identified. Of the 55 items on the DRET, 22 items (40%) were identified as measuring *Interpreting*, which was the most frequently measured skill. The two least frequently measured skills were *Recognizing functional values* (2 items or 3.64%) and *Recognizing style and tone* (2 items or 3.64%).

Table 6

Percentage of the Items Identified for Each Reading Sub-skill on the SAET and the DRET across the Five Years

Reading skill	SAET		DRET	
	No. of items	% of total	No. of items	% of total
Comprehending literal meaning	20	25.32	7	12.73
Recognizing and interpreting cohesive devices	2	2.53	3	5.45
Interpreting discourse makers	0	0	0	0
Recognizing functional value	5	6.33	2	3.64
Recognizing text organization	0	0	0	0
Recognizing presuppositions	0	0	0	0
Recognizing implications and inferences	6	7.59	10	18.18
Predicting	0	0	0	0
Reorganizing	15	18.99	9	16.36
Interpreting	31	39.24	22	40.00
Recognizing style and tone	0	0	2	3.64
Total	79	100.00	55	100.00

A close look at Table 6 reveals some similarities between the SAET and the DRET over the 2004-2008 period. One similarity worthy of mentioning is that, for both tests, *Interpreting* was the most frequently measured skill. Specifically, for both tests the largest proportion of items measured the skill *Interpreting*, accounting for 39.24% and 40%, respectively. This finding appeares to suggest that both tests emphasized the importance of measuring examinees' ability to identify a restatement of a sentence or a passage. Another similarity between the two tests is that *Reorganizing* was the third most frequently measured sub-skill for both tests, with 18.99% for the SAET and 16.36% for the DRET.

In terms of the percentage of items measuring each reading sub-skill, several interesting differences between the SAET and the DRET can also be observed from Table 6. For example, one striking difference between the two tests is that the skill *Recognizing style and tone* occurred only on the DRET (in 2006 and 2007). In other words, *Recognizing style and tone* was never measured on the SAET over the 2004-2008 period. A possible reason for this finding may be the decision on the part of test constructors to differentiate the purposes between the two tests. As noted by Yin (2005), the SAET is designed to measure examinees' general scholastic ability whereas the DRET is intended to assess their relatively advanced scholastic ability.

Hence, it is reasonable to observe the sub-skill *Recognizing style and tone* only on the DRET, as the skill is usually perceived as a relatively advanced reading sub-skill. This perception is also evidenced in Lan's (2007) classification of the sub-skill into the subcategory "Attributing" under "Analyze", a higher (the fourth) level of cognitive processing in the revised Bloom' s Taxonomy.

Table 6 also reveals a few more differences in the skill type preferred between the two tests. For instance, the SAET had more *Comprehending literal meaning* items than the DRET. Specifically, the percentage of *Comprehending literal meaning* items was 25.32% for the SAET, considered as the second most frequently measured skill; on the other hand, the percentage of *Comprehending literal meaning* items for the DRET was only 12.72%. As for the sub-skill *Recognizing implications and inferences*, more items were identified for the DRET than for the SAET. In particular, the percentage of *Recognizing implications and inferences* items for the DRET was 18.18% (the second most frequently measured skill), whereas for the SAET it was only 7.59% (the fourth largest proportion). These findings, roughly in line with those of Lan's (2007) study, appeared to suggest that the SAET puts more emphasis on the skill *Comprehending literal meaning* while the DRET underscores the skill *Recognizing implications and inferences*. A possible reason might be that the *Comprehending literal meaning* skill, which is normally viewed as a reading skill that is relatively important but basic, better fits one of the main purposes of the SAET (i.e., to measure examinees' basic scholastic or reading ability). As for *Recognizing implications and inferences*, it is similar to *Recognizing style and tone* in the sense that the two skills are usually considered to be associated with high-level reading skills, which fit the general aim (i.e., to measure examinees' advanced scholastic ability) of the DRET. Hence, it is quite reasonable to find that the DRET includes more *Recognizing implications and inferences* items than the SAET.

Another interesting finding was that, of the seven reading skills identified in the two tests, only two (i.e., *Comprehending literal meaning*, and *Recognizing and interpreting cohesive devices*) are considererd the bottom-up processing skills, according to Nuttall (2000). The other five identified reading skills are categorized as the top-down processing skills. In fact, the ratio of bottom-up processing skills to top-down processing skills was found to be 27.85: 72.15 (1: 2.59) for the SAET and 18.18: 81.82 (1: 4.5) for the DRET. Similar to the results of Chern (2006) about the BECT, our findings indicate that, like the BECT, both the SAET and the DRET (especially the DRET) tended to measure more top-down reading skills than bottom-up skills. However, no matter what ratio was found between the top-down and bottom-up reading skills, the six reading skills identified on the SAET and the seven on the DRET appeared to suggest that the test takers were assumed to have employed these reading skills while completing the items. Hence, the findings seemed to, according to the definition of Bachman and Palmer (1996), provide evidence for the

interactiveness (and thus the construct validity) of the reading comprehension items in both tests.

Similarities and Differences between the Two Tests in the Percentage of the Reading Sub-skills by Year

Table 7 presents the percentage of the reading sub-skills identified on the SAET and the DRET for each year over the 2004-2008 period. One similarity between the two tests was that the number of reading sub-skills measured in both tests each year ranged from four to six. Specifically, both tests measured five reading sub-skills in 2004 and four sub-skills in 2008. In 2008, the two tests had not only the same number of sub-skills but also measured exactly the same sub-skills — *Comprehending literal meaning, Recognizing implications and inferences, Interpreting, and Reorganizing*. Another similarity between the two tests was that the sub-skill *Interpreting* had the highest percentage in 2008 — 50% for the SAET and 54.5% for the DRET. Further, for both tests, the sub-skill *Recognizing and interpreting cohesive devices* was measured in two years only and the respective percentages of the items tended to be low.

One major difference between the two tests, also shown in Table 7, was the variation in the percentage of items measuring certain sub-skills over the 2004-2008 period. In particular, the percentage of *Reorganizing* items on the SAET was quite stable over the period, ranging from 13.3% to 25%, whereas the percentage of *Reorganizing* items on the DRET varied widely from 0% to 45.5% over the same period. Similarly, for the sub-skill *Recognizing Implications and Inferences,* the percentage of items on the SAET was fairly stable over the five years, ranging from 6.2% to 12.5%; on the other hand, the percentage of *Recognizing Implications and Inferences* items on the DRET changed considerably from 9.1% to 36.4% over the five years. This greater variation with respect to the DRET calls for close attention on the part of DRET item constructors in order to maintain year-to-year stability in the percentage of items measuring certain sub-skills.

Table 7

Percentage of Each Reading Sub-skill Identified on the SAET and the DRET by Year

			LM	CD	DM	FV	TO	PS	IF	PD	IT	RO	ST
SAET	2004	No. of items	6	1	0	0	0	0	1	0	5	2	0
		% of total	40.0%	6.7%	.0%	.0%	.0%	.0%	6.7%	.0%	33.3%	13.3%	.0%
	2005	No. of items	3	1	0	3	0	0	2	0	3	4	0
		% of total	18.8%	6.2%	.0%	18.8%	.0%	.0%	12.5%	.0%	18.8%	25.0%	.0%
	2006	No. of items	4	0	0	1	0	0	1	0	7	3	0
		% of total	25.0%	.0%	.0%	6.2%	.0%	.0%	6.2%	.0%	43.8%	18.8%	.0%
	2007	No. of items	3	0	0	1	0	0	1	0	8	3	0
		% of total	18.8%	.0%	.0%	6.2%	.0%	.0%	6.2%	.0%	50.0%	18.8%	.0%
	2008	No. of items	4	0	0	0	0	0	1	0	8	3	0
		% of total	25.0%	.0%	.0%	.0%	.0%	.0%	6.2%	.0%	50.0%	18.8%	.0%
DRET	2004	No. of items	0	2	0	1	0	0	1	0	5	2	0
		% of total	.0%	18.2%	.0%	9.1%	.0%	.0%	9.1%	.0%	45.5%	18.2%	.0%
	2005	No. of items	1	0	0	0	0	0	1	0	4	5	0
		% of total	9.1%	.0%	.0%	.0%	.0%	.0%	9.1%	.0%	36.4%	45.5%	.0%
	2006	No. of items	2	1	0	1	0	0	3	0	3	0	1
		% of total	18.2%	9.1%	.0%	9.1%	.0%	.0%	27.3%	.0%	27.3%	.0%	9.1%
	2007	No. of items	2	0	0	0	0	0	4	0	4	0	1
		% of total	18.2%	.0%	.0%	.0%	.0%	.0%	36.4%	.0%	36.4%	.0%	9.1%
	2008	No. of items	2	0	0	0	0	0	1	0	6	2	0
		% of total	18.2%	.0%	.0%	.0%	.0%	.0%	9.1%	.0%	54.5%	18.2%	.0%

Note. LM refers to comprehending literal meaning. CD refers to recognizing and interpreting cohesive devices. DM refers to interpreting discourse markers. FV refers to recognizing functional value. TO refers to recognizing text organization. PS refers to recognizing presuppositions. IF refers to recognizing implications and inferences. PD refers to predicting. IT refers to interpreting. RO refers to reorganizing. ST refers to recognizing style and tone.

Examinees' Performances on the Reading Comprehension Questions of Both Tests

The averages of the mean passing rates for each reading sub-skill over the 2004-2008 period are shown in Table 8. With respect to the SAET, the average of the mean passing rates for each sub-skill was greater than 50%, which indicates that the items were easy to medium-difficult. Of the six reading sub-skills identified on the SAET, *Comprehending literal meaning* items (63.80%) earned the highest average of the mean passing rates and *Recognizing functional value* items (50.67%) obtained the lowest average of the mean passing rates. The finding about the test-takers' best performance on the *Comprehending literal meaning* items appeared to be consistent with that of Lu's (2002). As to the DERT, only two reading sub-skills (i.e., *Comprehending literal meaning* and *Recognizing functional value*) obtained an average of the mean passing rates that was higher than 50%. This finding suggests that, in general, the items for the DRET over the 2004-2008 period were medium-difficult to difficult. Of the seven reading sub-skills identified on the DRET, *Recognizing functional value* items had the highest average of the mean passing rates (65.50%), and *Recognizing style and tone* items (28.50%) had the lowest average of the mean passing rates. The fact that test takers performed the worst on *Recognizing style and tone* items was also found in Lu's (2002) study, where an average of the mean passing rate as low as 27% was found.

A close examination of Table 8 reveals one similarity between the SAET and the DRET in the average of the mean passing rates for each reading sub-skill. That is, the averages of the mean passing rates for the *Comprehending literal meaning items* were found to be very high for both tests. Specifically, for this sub-skill, the items for the SAET had the highest average (63.80%) of the mean passing rates and those for the DRET had the second highest average (51.13%) of the mean passing rates. This finding seems to suggest that, for both tests, examinees tended to perform quite well on items measuring the skill *Comprehending literal meaning,* which is categorized as a bottom-up skill and is considered to be relatively essential and basic. This similarity was also found for items on *Recognizing functional value.* That is, the averages of the mean passing rates (50.60% for the SAET and 65.50% for the DRET) for this reading sub-skill were higher than 50%.

Table 8
Averages of the Mean Passing Rates for Each Reading Sub-skill Identified on the SAET and the DRET over the 2004-2008 Period

Reading skill	SAET		DRET	
	N	*M*	N	*M*
Comprehending literal meaning	20	63.80	7	51.13
Recognizing and Interpreting cohesive devices	2	57.00	3	37.75

Recognizing functional value	5	50.67	2	65.50
Recognizing implications and inferences	6	55.90	10	44.17
Interpreting	31	58.74	22	44.75
Reorganizing	15	54.25	9	48.97
Recognizing style and tone	0	.	2	28.50
Total number	79		55	
Average		56.73		45.82
SD		4.42		11.51

Note. N refers to the number of items measuring each particular reading skill from 2004 to 2008

On the other hand, a further examination of Table 8 also brought out several differences in examinees' performance between the two tests. The first difference was that the averages of the mean passing rates for the six reading sub-skills on the SAET fell into a narrow range, between 50.67% and 63.80% (with a standard deviation of 4.42%), whereas those on the DRET ranged widely from 28.50% to 65.50% (with a standard deviation of 11.51%). That is, the level of item difficulty for the DRET tended to vary more drastically for different reading sub-skills than that for the SAET.

The second difference, again revealed in Table 8, was that, except for the sub-skill *Recognizing function value*, the average of the mean passing rates for each sub-skill identified on the SAET was higher than that on the DRET. This finding suggests that the reading comprehension items on the DRET are in general more difficult than those on the SAET. Several reasons can explain this phenomenon. First of all, based on the results of a related study of Yin (2005), the length of the sentence in the reading passages on the DRET was found to be generally longer than on the SAET. Similarly, the length of the passages (around 200 to 300 words) on the DRET on average was found to be longer than the length of the passages (around 150 to 250 words) on the SAET. In addition, compared with that on the SAET, the sentence structure of the reading passages on the DRET was found to be more complex. A similar comment was also made by Yu (2006) who recommended that the use of words and the readability level of the text are more difficult on the DRET than on the SAET. Hence, one would expect the text difficulty of the passages and the difficulty level of the reading comprehension items on the DRET to be higher than the corresponding items on the SAET. Another possible reason resulting in the lower passing rates for DRET examinees than for SAET examinees is the difference in the length of time allowed for the two tests. Although both tests have an equal number of test items, the length of time allotted for the SAET is 100 minutes while for the DRET it is 80 minutes. Finally, a difference in the grading scheme may also account for the difference in the passing rates for the two tests. Unlike that for the SAET, the grading scheme for the DRET penalizes examinees for answering test items incorrectly. Hence,

DRET examinees tend not to answer items they only have partial knowledge about. With all the above reasons possibly at work, it is not surprising to find that DRET examinees did not perform as well as SAET examinees.

Another difference, which can also be found from Table 8, was that DRET takers tended to perform best on items measuring *Recognizing functional value*, while SAET takers tended to perform worst on items measuring this sub-skill. Specifically, the average of the mean passing rates for the *Recognizing functional value* items was 65.50% for the DRET and 50.67% for the SAET. In other words, *Recognizing functional value* items seemed to be the easiest items on the DRET but the most difficult items on the SAET. One possible reason for this unexpected finding is difference in item stems used between the two tests. On the DRET, the item stem used to measure this skill was "This passage is most likely taken from a _____." DRET takers can determine the answer simply by using their topical knowledge and the key words appearing in the text. On the other hand, the item stem used on the SAET to measure the skill was "The main purpose of the passage is to _____." To answer this type of item correctly, SAET takers need to read through the entire passage before obtaining a thorough understanding of the text. Obviously, items with this type of stem on the SAET tended to be more difficult than those on the DRET. This may explain why SAET takers performed much worse on the *Recognizing functional value* items, while DRET takers performed much better on items measuring this skill.

Another point to note is that DRET takers performed considerably worse on *Recognizing style and tone* items, with an average of the mean passing rates of 28.50%, which is below the minimum standard rate of 33% set by Jeng et al. (1999). The finding may be due to two possible reasons. First, this type of item may be too difficult for most DRET takers. Second, DRET takers may not have received enough formal instruction on this skill during their study in senior high school.

One final finding worthy of mention is that the top-down sub-skills (e.g., *Recognizing functional value*, *Reorganizing*, and *Recognizing style and tone*) did not necessarily result in averages of the mean passing rates lower than those of the bottom-up sub-skills (e.g., *Comprehending literal meaning* and *Recognizing and interpreting cohesive devices*). Take the DRET for example. The sub-skill *Recognizing functional value*, a top-down sub-skill, had an average of the mean passing rates of 65.50%, much higher than that of 37.75% for *Recognizing and interpreting cohesive device,* a bottom-up sub-skill. This was also the case for the SAET, where the top-down sub-skill *Interpreting* was found to produce an average of the mean passing rates of 58.74% slightly higher than that of 57.00% for the bottom-up sub-skill *Recognizing and interpreting cohesive devices*.

Table 9 presents the mean passing rate for each reading sub-skill measured on the SAET and the DRET each year. For the SAET, the mean passing rate ranged from 31.00% to 76.00%. Three (out of six) reading sub-skills had quite stable mean passing

rate over the 2004-2008 period, including *Interpreting cohesive devices* (*SD* = 1.41%), *Comprehending literal meaning* (*SD* = 4.36%), and *Interpreting* (*SD* = 5.34%). In addition, the mean passing rate for each of the three sub-skills in each year was all above 50%. The other three sub-skills, on the other hand, varied widely over the five years. As for the DRET, the mean passing rate for the seven sub-skills identified ranged from 27.00% to 71.00%. The mean passing rates for two reading sub-skills (out of seven) were relatively quite stable over the five years. The two sub-skills were *Recognizing style and tone* and *Recognizing functional value,* with an *SD* of 0.71% and 3.54%, respectively. On the other hand, the mean passing rates for the other five sub-skills varied relatively widely over the five years.

Table 9
Mean Passing Rate for Each Reading Sub-skill by Year

Reading skill								
		LM	CD	FV	IF	RO	IT	ST
SAET	2004	63.83	56.00	–	66.00	54.00	52.40	–
	2005	67.33	58.00	54.00	55.50	38.25	65.33	–
	2006	60.75	–	31.00	76.00	51.67	55.86	–
	2007	58.33	–	67.00	41.00	61.67	57.00	–
	2008	68.75	–	–	41.00	65.67	63.13	–
M		*63.80*	*57.00*	*50.67*	*55.90*	*54.25*	*58.74*	–
SD		*4.36*	*1.41*	*18.23*	*15.41*	*10.59*	*5.34*	–
DRET	2004	–	44.50	63.00	27.00	37.50	37.00	–
	2005	57.00	–	–	38.00	55.40	47.00	–
	2006	50.50	31.00	68.00	38.33	–	36.33	28.00
	2007	42.00	–	–	46.50	–	51.25	29.00
	2008	55.00	–	–	71.00	54.00	52.33	–
M		*51.13*	*37.75*	*65.50*	*44.17*	*48.97*	*44.78*	*28.50*
SD		*6.66*	*9.55*	*3.54*	*16.52*	*9.96*	*7.68*	*0.71*

Note. LM refers to comprehending literal meaning. CD refers to recognizing and interpreting cohesive devices. FV refers to recognizing functional value. IF refers to recognizing implications and inferences. RO refers to reorganizing. IT refers to interpreting. ST refers to recognizing style and tone.

The Mean Discrimination Index (D) for Each Reading Skill

Table 10 provides the average of the mean *D*'s for each reading sub-skill on the SAET and the DRET over the five years. For the SAET, the averages of the mean *D*'s

over the five years fell in a narrow range of between 46.11% and 58.50%. In fact, the average of the mean *D*'s for each reading sub-skill was higher than 50%, except for *Recognizing functional value* (46.11%). The highest average of the mean *D*'s was found for *Recognizing implications and inferences* items. Unlike those on the SAET, the averages of the mean *D*'s on the DRET varied widely from 20.00% to 56.80%. Of the seven reading sub-skills, the *Comprehending literal meaning* items had the highest average of the mean *D*'s (56.38%) while the *Recognizing style and tone* items had the lowest average of the mean *D*'s (20.00%).

Table 10
Averages of the Mean Discrimination Indices (D) for Each Reading Sub-skill on the SAET and the DRET

Reading skill	SAET		DRET	
	D	SD	D	SD
Comprehending literal meaning	57.72	5.61	56.38	12.84
Recognizing and interpreting cohesive devices	58.50	0.71	45.25	4.60
Recognizing functional value	46.11	7.70	44.50	9.19
Recognizing implications and inferences	55.80	10.57	39.25	13.83
Interpreting	50.46	3.96	50.94	6.46
Reorganizing	52.37	11.57	52.93	8.67
Recognizing style and tone	—	—	20.00	11.31

A close look at Table 10 indicates some similarities and differences between the SAET and the DRET. One similarity between the two tests was that three skills (*Comprehending literal meaning*, *Interpreting*, and *Reorganizing*) had their averages of the mean *D*'s higher than 50%. Furthermore, on both tests, the averages of the mean *D*'s for the skill *Recognizing functional value* were lower than 50%, being 46.11% and 44.50% respectively. As to the differences between the two tests, it is clear that, except for *Interpreting* and *Reorganizing*, the SAET items tended to outperform the DRET items in terms of the discriminating power of the items measuring each reading skill. In fact, all averages of the mean *D*'s for the six reading skill identified on the SAET over the five years were higher than 50%, except for the skill *Recognizing functional value*. On the DRET, only three skills (*Comprehending literal meaning*, *Interpreting*, and *Reorganizing*) had averages of the mean *D*'s higher than 50%.

Two possible reasons may account for the finding that the discriminating power obtained for the DRET was lower than that obtained for the SAET. First, many

examinees did not take the DRET because they had been admitted to their desired universities based on their acceptable SAET scores. As a result, the range as well as the variance for DRET scores would diminish, which in turn would decrease the discriminating power of the items on DRET test. Second, one may speculate that the more complicated sentence structure of the texts and thus higher difficulty level for the texts on the DRET may have led to a poorer performance for both high and low scorers, which in turn would lead to a smaller variance for DRET scores. However, the second reason is merely a speculation and awaits future investigation.

Another finding to note was that the overall pattern of differences in the discriminating power between the items on the bottom-up skills and those on the top-down skills was not the same between the SAET and the DRET. For the SAET, items classified as the bottom-up skills (e.g., *Comprehending literal meaning*, *Recognizing and interpreting cohesive devices*) had discriminating power of items higher than those classified as the top-down skills (e.g., *Reorganizing* and *Interpreting*). This was not necessarily the case for the DRET. For example, the items on *Comprehending literal meaning* (a bottom-up skill) for the DRET did produce an average of the mean *D*'s (56.38%) higher than that of 52.93% for the items on *Reorganizing* (a top-down skill). However, the average of the mean *D*'s (45.25%) for the items on *Recognizing and interpreting cohesive devices* (a bottom-up skill) was much lower than that of 52.93% for those on *Reorganizing* (a top-down skill).

One final point worthy of mentioning is that, as seen from both Tables 8 and 10, the averages of the mean passing rates and those of the mean *D*'s for the SAET seemed to be less varied across the types of reading skills than those for the DRET. Specifically, for the SAET, the averages of the mean passing rates for the six reading skills fell in a relatively narrow range from 50.67% to 63.80%, whereas the averages of the mean passing rates for the seven reading skills for the DRET ranged widely from 28.50% to 65.50%. Similarly, the averages of the mean *D*'s for the SAET fell in a narrow range between 46.11% and 58.50%, whereas the averages of the mean *D*'s ranged widely from 20.00% to 56.38%. That is, over the five-year period, the level of item difficulty and the discriminating power for the DRET tended to vary more widely across different reading skills than those for the SAET.

Table 11 presents the mean *D* for items measuring each of the reading sub-skills on the SAET and the DRET each year over the 2004-2008 period. As shown in Table 11, the mean *D* for each reading skill on the SAET ranged from 37.50% to 69.00%. In addition, the *Comprehending literal meaning items* had the highest mean *D* in 2005 (59.67%) and in 2006 (62.75%); the *Reorganizing* items had the highest mean D in 2008 (66%); and the *Recognizing implications and inferences* item had the highest mean *D* in 2004 (69%) and in 2007 (64%). As for the DRET, the mean *D* for each reading skill ranged between 12.00% and 68.50%.

Similar to the SAET, the *Comprehending literal meaning* items had the highest

mean *D* in 2005 (63%) and in 2006 (68.50%), and the *Reorganizing* items had the highest mean *D* in 2008 (59%). Furthermore, the *Recognizing functional value* items had the highest mean *D* in 2004 (51%) while the *Interpreting* items had the highest mean *D* in 2007 (53.75%). These findings, which were consistent with those of Lan's (2007) study, appear to suggest that, for both tests over the five years, none of the reading skills identified could be found to consistently best discriminate between the high-scoring group and the low-scoring group.

Table 11
Mean Discrimination Index (D) for Each Reading Sub-skill on the SAET and the DRET by Year

Reading skill								
		LM	CD	FV	IF	RO	IT	ST
SAET	2004	48.50	59.00	–	69.00	37.50	51.60	–
	2005	59.67	58.00	53.33	50.00	44.00	46.67	–
	2006	62.75	–	47.00	53.00	59.67	46.29	–
	2007	56.67	–	38.00	64.00	54.67	52.13	–
	2008	61.00	–	–	43.00	66.00	55.63	–
M		*57.72*	*58.50*	*46.11*	*55.80*	*52.37*	*50.46*	–
SD		*5.61*	*0.71*	*7.70*	*10.57*	*11.57*	*3.96*	–
DRET	2004	–	48.50	51.00	22.00	43.00	40.60	–
	2005	63.00	–	–	43.00	56.80	53.50	–
	2006	68.50	42.00	38.00	30.00	–	49.33	12.00
	2007	39.00	–	–	43.25	–	53.75	28.00
	2008	55.00	–	–	58.00	59.00	57.50	–
M		*56.38*	*45.25*	*44.50*	*39.25*	*52.93*	*50.94*	*20.00*
SD		*12.84*	*4.60*	*9.19*	*13.83*	*8.67*	*6.46*	*11.31*

CONCLUSIONS

Based on the above results, the following conclusions can be made in the order of the three research questions stated earlier: (1) Six out of the 11 reading skills were identified on the SAET over the 2004-2008 period, including *Comprehending literal meaning*, *Recognizing and interpreting cohesive devices*, *Recognizing functional value*, *Recognizing implications and inferences*, *Interpreting*, and *Reorganizing*. For the DRET, *Recognizing style and tone* was also identified in addition to the above six

reading skills. Of the seven identified skills, *Interpreting* was the most frequently measured skill for both tests. Given the reading skills identified in this study, it appears that the two tests could be considered interactive. (2) SAET takers performed best on the *Comprehending literal meaning* items but worst on the *Recognizing functional value* items, whereas DRET takers performed best on the *Recognizing functional value* items but worst on the *Recognizing style and tone* items. In addition, DRET takers generally did not perform as well as SAET takers on the reading skills identified. (3) For both tests, the reading skill found to best discriminate high scorers from low scorers tended to change over the five-year period. Hence, none of the reading skills identified could be claimed to consistently best discriminate high scorers from low scorers for the two tests over the period.

Given the results of this study, several pedagogical implications can be made for classroom practice. First, knowing that a total of six to seven reading skills were identified from the two tests, English teachers should help their students develop and master these reading skills. In particular, given that a high percentage of items were identified on both tests to measure the *Interpreting* skill, teachers need to make certain that this basic but essential skill is taught or assessed in the classroom. Second, teachers are strongly recommended to strengthen their students' ability to recognize style and tone, and their ability to recognize and interpret cohesive devices, based on the results that the DRET takers showed relatively low passing rates on items assessing these two sub-skills. Third, most items on the DRET had lower passing rates than those on the SAET. As stated earlier, one of the reasons for this finding, also found by some previous studies (Yin, 2005; Yu, 2006), was that the reading passages on the DRET generally contain more difficult words, longer sentences, and more complex sentence structure than those on the SAET. Hence, teachers should not only assist students in building up their reading skills identified on the DRET, but also make efforts to help students improve their lexical and syntactical competence.

Apart from the implications for classroom practice, the differences in the results of this study between the SAET and the DRET can also have some implications for the construction of reading comprehension items. To begin with, most of the SAET items had higher mean passing rates (and thus higher mean item difficulty indices) than the DRET items. Simply put, the SAET items in general were easier than the DRET items. Similarly, the SAET items in general tended to outperform the DRET items in terms of the discriminating power of the items measuring the reading sub-skills. Hence, given the fact that items with medium item difficulty indices tended to have high discriminating power, if a goal of the DRET is to discriminate among examinees, these findings may serve as a reminder for the DRET constructors when constructing items with a medium (rather than higher) level of item difficulty. In the meantime, considering the reasons mentioned earlier for the lower passing rates and the discriminating power of the DRET items, DRET constructors may need to think

about not only whether to remove the penalty for answering DRET items incorrectly, but also whether to lengthen the time for the DRET test. In addition, over the 2004-2008 period, the level of item difficulty and the discriminating power of the DRET tended to vary more drastically across different reading skills than for the SAET. This finding may also point to a need for DRET constructors to reflect on the causes for such a big variation across various reading skills. Finally, the current study found that *Recognizing functional value* items seemed to be the easiest items on the DRET, while they appeared to be the most difficult items on the SAET. One possible reason, as mentioned earlier, is the difference in the item stems used between the two items measuring the same type of reading skill. This finding, suggesting that items assessing the same reading skill but being phrased differently in the item stems could result in a huge difference in their item difficulty, may help to remind test constructors to exercise extra caution when phrasing the item stems during their item construction.

The results of this study were subject to some limitations. First, this study employed only two raters to categorize the test items into one of the 11 reading skills. Hence, in some cases, the results may not be consistent with those of other studies employing different raters. Second, the two raters' item categorization could only represent their predictions about the cognitive or language processing that each item attempted or intended to assess. The question about whether test takers really applied the cognitive or language processing skills identified while answering the items requires further investigation. As Alderson, Clapham, & Wall (1995) pointed out, information on how test takers actually respond to test items -- the process they undergo and the reasoning they engage in when responding -- can be crucial indications of what the test is gauging. This kind of introspective data can be gathered concurrently or retrospectively in the form of "think aloud" or "in-depth interviews." Third, due to the limit of its scope, the present study did not take topic variation into consideration when discussing the differences in the passing rate and discriminating power among the items measuring different reading sub-skills. Future studies incorporating topic variation could be useful in providing DRET and SAET constructors with more insight into various potential reasons affecting examinees' performance on these reading comprehension tests.

REFERENCES

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Almasi, J. F. (2003). *Teaching strategic processes in reading*. New York, NY: Guilford.

Anderson, L. W., & Krathwohl, D. R., (Eds.) (2001). *A taxonomy for learning,*

teaching, and assessing: A revision of Bloom's educational objectives. New York: Longman.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bloom, B. S., Engelhart, M. D., Frost, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives: Handbook I, Cognitive Domain.* New York: David Mckay.

Fan, Y. S. (2008). A Comparison of Scholastic Aptitude English Test and Department Required English Test. Unpublished master's thesis, National Chung Cheng University, Chiayi.

Hsu, W. L. (2005). An Analysis of the reading comprehension questions in the JCEE English test. Unpublished master's thesis, National Kaohsiung Normal University, Kaohsiung.

Hughes, A. (2003). *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Jeng, H. Hs., Yang, I. L., Chen, Hs. Ch., Chen, L. Hs., Chen, K. T., & Wu, H. Ch. (1999). *An experiment in designing English proficiency tests of two difficulty levels for the college entrance examination in Taiwan.* Taipei: CEEC.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into Practice*, *41*, 212-218.

Kreitzer, A. E., & Madaus, G. F. (1994). Empirical investigation of the hierarchical structure of the taxonomy. In Anderson, L. W., & Sosniak, L. A. (Eds.) *Bloom's taxonomy: A forty-year retrospective* (pp.64-81). Chicago, IL: The National Society for the Study of Education.

Lan, W. H. (2007). *An analysis of reading comprehension questions on the SAET and the DRET using revised Bloom's taxonomy.* Unpublished master's thesis, National Taiwan Normal University, Taipei.

Lu, J. J. (2002). *An analysis of the reading comprehension test given in the English subject ability test in Taiwan and its pedagogical implications.* Unpublished master's thesis, National Chengchi University, Taipei.

Mayer, R. E. (2002). Rote versus meaningful learning. *Theory Into Practice, 41*, 226-232.

Mo, C. C. (1987). A study of English reading comprehension and general guidelines for testing reading. *Journal of National Chengchi University*, *55*, 173-206. Taipei: National Chengchi University.

Nuttall, C. (2000). *Teaching reading skills in a foreign language*. London: Heinemann.

You, H. L. (2004). *Analysis of the basic competence English test for junior high school.* Unpublished master's thesis, National Yunlin University of Science & Technology, Yunlin.

Yu, H. Y. (2006). The development of English testing and teaching in Taiwan: A survey of the college entrance English exam and high school English teaching. *English Teaching and Learning*, *Special Issue 2*, 133-151.

陳秋蘭 (Chern, C. L.) (民85)。**基本學力測驗英語科之內涵分析:認知及閱讀能力分析**。行政院國家科學委員會專題研究成果報告（編號：NSC94-2411-H-003-024），未出版。

APPENDIX

Definition and Sample Question for Each Reading Skill

Reading skill	Definition
LM	Ability to locate or identify specifically stated facts.
CD	Ability to interpret the pro-forms, the elliptical expressions, and the lexical cohesions.
DM	Ability to recognize markers that signal the sequence of events, markers that signal discourse organization, or markers that signal the writer's point of view.
FV	Ability to identify the functional value of the sentence or the whole paragraph. Types of functional value include defining, classifying, asserting, exemplifying, instructing, apologizing, and so on.
TO	Ability to identify the principle by which the text is organized and recognize how the ideas hang together.
PS	Ability to recognize the presuppositions underlying the sentences or text. Presuppositions can be divided into two groups: the background knowledge and/or experience that the writer expects the reader to have, and the opinions, attitudes, or emotions that the writer expects the reader to share or to understand.
IF	Ability to identify the meaning that is not explicitly stated but can be inferred.
PD	Ability to predict what is likely to come next and what is not.
RO	Ability to combine information from various parts of the text and put it together in a new way (e.g., by calculating).
IT	Ability to identify a restatement of a sentence or a passage.

ST	Ability to recognize the writer's tone, mood, voice, attitude, or the text style.

Note. LM refers to comprehending literal meaning. CD refers to recognizing and interpreting cohesive devices. DM refers to Interpreting discourse markers. FV refers to recognizing functional value. TO refers to recognizing text organization. PS refers to recognizing presuppositions. IF refers to recognizing implications and inferences. PD refers to predicting. RO refers to reorganizing. IT refers to interpreting. ST refers to recognizing style and tone.

大學學科能力測驗及指定科目考試
英文閱讀測驗題目之評鑑

摘要

本研究的主要評鑑民國93年至97年大學學科能力測驗及指定科目考試的英文閱讀測驗題目。本研究主要探討以下三個研究問題：一、大學學測及指考的英文閱讀測驗題目評量哪些閱讀技巧？評量這些閱讀技巧的題目各佔百分之多少？二、學測及指考的應試者針對評量不同閱讀技巧的題目的表現為何？三、哪一項閱讀技巧最能在這五年的學測及指考中一貫的鑑別出高分組與低分組？閱讀測驗題目的分類法主要採用 Nuttall (2000) 對於閱讀技巧及問題類型的分類。兩位英文領域的專家，將134題英文閱讀測驗題目個別分類至十一個閱讀技巧中。研究結果顯示，大學學科能力測驗的英文閱讀測驗題目，主要評量以下六項閱讀技巧：一、詮釋(39.24%)，二、字面上的理解(25.32%)，三、再組織(18.99%)，四、理解言外之意及推論(7.59%)，五、辨別句子或文章的功能價值(6.33%)，六、解釋關聯詞語(2.53%)。指定科目考試的英文閱讀測驗題目，除了上述這六項技巧外，尚有還有評量「辨別文章風格及作者論調」這項技巧。這七項技巧出現的百分比呈現如下:詮釋(40%)、理解言外之意及推論 (18.18%)、再組織(16.36%)、字面上的理解(12.73%)、解釋關聯詞語(5.45%)、辨別句子或文章的功能價值(3.64%)，及「辨別文章風格及作者論調」(3.64%)。此外，學測應試者對於評量「字面上的理解」的題目表現最好，但對於「辨別句子或文章的功能價值」的題目表現最差；指考應試者對於評量「辨別句子或文章的功能價值」的題目表現最好，但對於「辨別文章風格及作者論調」的題目表現最差。普遍來說，應試者在學測的表現比應試者在指考的表現來得好。另外，在這五年的學測及指考中，沒有一項閱讀技巧被發現最能夠一貫的鑑別出高分組及低分組。

關鍵字：閱讀理解 布魯姆分類 Nuttall 閱讀技能分類 參與性 構念效度