# The Types and Nature of Questions vis-à-vis Students' Test-Taking Skills as Significant Indicators of Second Language Examinees' Performance on the TOEFL-ITP Reading Comprehension Sub-Test

Analiza Perez-Amurao
*Mahidol University International College*
*Salaya, Thailand*

## Abstract

*This study examines the reading performance of selected students at the Pre-College program of the Mahidol University International College (PC-MUIC) as they are required to attain a score of 520 in the TOEFL-ITP (or equivalent performance in IELTS) to enter MUIC. Specifically, this research aims to evaluate whether the reading skills that examinees possess correlate with successful performance on the Reading Comprehension sub-test of the TOEFL-ITP. Only TOEFL-ITP Reading Comprehension Sub-test performance has been considered in this study as IELTS is not taught or administered in the Pre-College program. This study makes use of descriptive qualitative-quantitative design relying heavily on the following instruments for data collection: Commercial-based test-prep texts (Reading Comprehension Sub-section), Schraw and Roedel's Levels of Difficulty (1994), the researcher's modification of said band, the respondents' scores per question type, tabulations of the respondents' scores based on the levels of difficulty of the items and the question types used in the test, focused interviews with the respondents, and retrospective journal entries of the researcher. This study aims to shed light on issues surrounding how second language learners' reading skills affect performance on standardized tests such as TOEFL. This study specifically seeks to provide MUIC PC instructors empirical data that would help them understand their own students' reading difficulties which, consequentially, will aid them address teaching-learning issues.*

**Key words:** TOEFL, reading skills, test-taking

## Introduction

In an EFL/ESL classroom, cases of a reading teacher facing the enormous and challenging task of making students love reading as the students themselves hesitate or refuse to embrace the benefits of the exercise are no longer new. Many times, a reading teacher finds himself frustrated at the end of the day as he finds it hard to make the students to even open the pages of their reading material (Snow, Griffin, & Burns, 2005). This, of course, is complicated by the fact that the teacher has no choice but to push the students even further not only because they need to learn and use a number of reading strategies at the end of the term, but also because they need to acquire reading skills and test-taking strategies for them to achieve satisfactory scores in the assessment stage.

In situations where the students' performance on high-stakes tests both culminates in one academic exercise, on one hand, and commences in another, on the other, both the learner and the teacher come to a point where obtaining the required score becomes the chief objective. In circumstances like this, the challenge of realizing the chief objective lies

© 2011 Time Taylor International ISSN 2094-1420

heavily on the shoulders of the teacher. This, of course, does not mean that the students are freed of their responsibility. This simply means that apart from the pro-active attitude expected of them being the test- takers and direct recipients of said pedagogic exercise, the academic community also looks forward to seeing the teacher take a more hands-on approach if only to moderate the task and eventually accomplish the goal (Dole, 2004).

In an effort to assist reading teachers, academicians and theorists who are monitoring this field for future academic discourse, this study was conducted to investigate and consequently provide corroborated data on learners' reading and test-taking skills as revealed by their scores in the Reading Subtest of their complete practice test. This paper conceptually aims to validate the importance of drawing on appropriate reading strategy albeit indirectly carried out in this study as correlated to achieving satisfactory scores in the said standardized test within the Reading Comprehension context.

This paper aims to establish a premise underscoring how the subjects' test performance directs educators to seek appropriate measures namely, but not limited to, the following diagnostic objectives: to identify the types of TOEFL-ITP Reading Comprehension questions examinees failed to understand and respond to accurately; to determine the nature of the types of questions examinees performed unsatisfactorily on; to identify the test-taking strategies examinees failed to use before and during the testing period, and; to identify the strategies that they presumably used successfully.

Likewise, this study underpins the need to attain the following pedagogic objectives: to construct effective activities based on the students' test-taking strategies that were judged to be successful (Focus on activities); to provide students with a cognitive framework for improving their test performance (Focus on self-development), and; to develop a test-taking strategies syllabus based on the diagnostic and pedagogical data that has been identified by this study (Focus on teaching materials).

The growing demand for the English language has resulted in an increase in the use of high-stakes international tests such as TOEFL, IELTS, TOEIC and the like. While more and more users of the English language are relying heavily on these tests for various reasons, including for admission to universities and for employment abroad, Andrew Cohen (2006) said that for the past 30 years, the focus of research in second language (L2) assessment was primarily on testing results. There was little attention given to what the test-takers do to arrive at the right answers and how the assessment matched the skills it was supposed to test. While there have been some studies conducted focusing on test-taking strategies, results of this researcher's inquiries revealed dearth in studies within the Thai context on L2 assessment looking into the relationship between the examinees' language skills, their test-taking strategies and the effectiveness of strategy instruction so as to aid test-takers' performance on high stakes standardized tests.

In a study of a Thai university students' performance on TOEFL, Marukatat and Bunnag (2003) found out that based on the scores

gathered from nine of the 10 Asean member-states whose TOEFL results were studied, the Thai students performed second from the last. The outcome showed how the Thai students were surpassed by their peers from the neighboring countries particularly from Burma, Vietnam and Cambodia. Various reasons were cited. In some other studies within the Thai context, one of the strongest reasons that accounted for such a poor performance was the country's lack of a strong reading culture, even in its native tongue. "It is unlikely that students have 'reading' models, a factor that may have a significant impact on the students if the teachers themselves are not seen to be readers" (Eskey & Grabe in Wisaijorn, 2008). Another reason considered very influential is that since many traditional Thai classrooms are teacher-controlled, the students have not been trained to think independently for themselves. Because of excessive emphasis on social conformity, i.e., "submissive students who do not ask questions are seen as well-behaved," the students, in turn, have become very passive, afraid of asking even the most relevant questions about their readings. Students who are "critical and analytical and who reason with their teachers are often viewed as aggressive and disobedient, and have trouble fitting into the Thai education system" (Chareonwongsak, 2007, p. 4). In situations like this, one can easily deduce how emotional factors control learners' accomplishments and actions (Alvermann & Guthrie, 1993). Research shows that when children are able to make positive associations with reading, they tend to regard the task as something rewarding and enjoyable. However, when they see negative connections between reading and what actually happens inside the classroom, "their achievement tends to suffer. These children will either avoid reading altogether or read with little involvement" (Henk & Melnick, 1995, p. 470).

On a related note, students from Mahidol University International College (hereafter referred to as MUIC) are required to attain a score of 520 in TOEFL (or equivalent performance in IELTS) to be admitted to the university. As preliminary observation based on some learners' past performance on TOEFL exams and review classes yielded, PC TOEFL instructors found out that learners have the most difficulty doing the Reading Comprehension Subtest. This problem is linked to the Thai students' ability to cope with simple communication in their own area, and in their inability to use advanced language (Fredrickson in Wisaijorn, 2006). In line with this, a study on the students' language proficiencies, test-taking strategies and strategy instruction can prove to be of immense value as it will provide the stakeholders, i.e., students, teachers, curriculum designers, practical information on how to improve the examinees' performance. This primarily explains why this study placed a singular focus on the students' performance on the Reading Comprehension Subtest.

Second language assessment research, for the most part, has focused on issues that deal primarily with the outcomes of language testing, item performance, test reliability and the like (Cohen, 2006). While results of these studies have served great roles especially useful for academicians, test constructors and even test-takers themselves, the fact

remains that L2 assessment has focused very little on the types and nature of questions L2 examinees find difficult responding to. Below are brief discussions of studies and results done in L2 assessment. These research results are presented with the aim of providing a bigger picture of what was already done and what is still needed to be done in said field.

Carol Fraser's (1999) report entitled "Lexical Processing Strategy Use and Vocabulary Learning through Reading" looked into the lexical processing strategies (LPSs) used by L2 learners as they tried to make sense of new words while reading and sought to see the effect of such strategies when learning unfamiliar lexical items. This study aimed to increase one's understanding of the role of LPS in as far as how they can help in the reassessment of the current academic practice. Results showed that generally, instruction using LPSs indicated potential for enriching one's vocabulary. The study also demonstrated that the use of some LPSs leads to retention rates higher than the others.

In 2002, Abanomey (In Cohen, 2006) conducted a test-taking strategy research exploring some features of a test format and checking on how influential the use of authentic texts is against inauthentic ones in a reading test. The inquiry looked into whether authentic texts would impact the manner in which test-takers employ test-taking strategies. It also explored the differences, if there are, between test-takers reading authentic texts and those reading inauthentic ones in their use of "bottom-up (text-based) and top-down (knowledge-based) strategies." Results showed that text authenticity did not affect the number of strategies used with them, although it did affect the manner in which examinees used the test-taking strategies.

In a related study in 2002, Liz Hamp-Lyons and Alan Davies carried out a research project with a hypothesis stating "that international English tests are biased: by that we mean that they systematically misrepresent the 'true scores' of candidates by requiring facility in a variety of English to which whole groups of candidates have not been exposed." This study primarily focused on issues that relate to assumptions about International English (IE) and World Englishes (WEs) views. Realizing that their data set was not sufficient for them to come to any conclusions, as they were only left with one data that appeared to support their WEs hypothesis, Hamp-Lyons and Davies suggested instead that a further study be done because although the "bias" on the basis of their research was "not proven," it could not be dismissed either.

In the context of a study done in 2002 involving the use of TOEFL, David Qian's (2002) investigation validated the significance of breadth and depth of vocabulary knowledge in reading comprehension within the scholastic setting. The study found out that the "dimension of vocabulary depth is as important as that of vocabulary size" in forecasting test-takers' performance on academic reading.

In a more localized research setting in Thailand, Patareeya Wisaijorn's (2008) "Strategy Training in the Teaching of Reading Comprehension" looked at the country's L2 reading situation and saw how current classroom practices have been directly influenced by the Thai

educational culture. Wisaijorn found out that having the greatest tendency to be teacher-centered and teacher-directed, the Thai education system negatively influences reading in English in most Thai classrooms such that "weak performances in reading in English indicate difficulties in fulfilling the demands of their [students'] studies."

While there have already been a considerable number of researches in L2 assessment, studies that explore on the correlation between the types and nature of questions as significant indicators of L2 examinees' performance on the TOEFL-ITP Reading Comprehension subtest remain deficient. So much so that the context in which this study was done is in Thai, a culture that has been generally labeled in a number of local researches to have a very poor reading ethic either in its own language or in a foreign one. Situated in a milieu where the education system is highly influenced by its culture, the Thai university students' reading skills seem to be performing unsatisfactorily. It is within this same vein why this research aims to pioneer investigations on Thai university students' performance on a high stakes test, particularly TOEFL. This study is anchored on the premise that by knowing and understanding the types and nature of questions and test-taking strategies used by the examinees, said variables would serve as significant indicators of their performance on the Reading Comprehension component of said test and, consequently, will aid in the curriculum and materials development.

## Method

### Design

The study made use of the descriptive qualitative-quantitative design. Using the scores per question type, this research observed and measured the behavior of the participants in relation to the specific types of questions they performed on successfully and unsuccessfully.

Purposive sampling was used as the main basis of selecting the participants. This sampling method was used because of the very nature of the investigation which is to look into the reading skills of a specific group of students placed under a bridge program in the Thai educational setting.

### Instruments

The investigation relied heavily on the following instruments for data gathering: commercial-based test-prep texts from *Barron's TOEFL Strategies*, a commercial-based test-prep material published by Barron's Educational Series, Inc. (particularly the Reading Comprehension sub-test); Schraw and Roedel's Levels of Difficulty (1994) with researcher's modification; the subjects' scores per question type; tabulations of the respondents' scores (Based on the levels of difficulty of the items and the question types used in the test); focused interviews with selected students, and; retrospective journal entries.

## Participants

This study involved three groups with a total of 43 students admitted to Level Four (4) of the PC Program during Quarter Four (4) of AY 2008. Level Four (4) students are those whose language skills have been identified to be in the Intermediate and Upper-Intermediate levels. Students in this group are those who are expected to join mainstream university classes at Mahidol University International College assuming that they get an average of D+, equivalent to 65%, in their writing, reading, and listening/speaking non-credit PC classes, and achieve a TOEFL score of 520. The selection of the students admitted to the program was primarily based on the scores they obtained after taking MUIC's College Entrance Examination. Being an international college, MUIC requires that apart from passing the entrance examination, students should also obtain a TOEFL-ITP score of 550 or above. When students fail to satisfy the admission requirements, they are recommended for enrolment at the Pre-College Program (hereafter referred to as PC), a bridge program of the MUIC which aims at improving the students' English language and mathematical skills.

## Procedure

The investigation was done by, first, extracting the questions from the Reading sub-test and tabulating them based on the subjects' correct and wrong responses. Second, using the initial tabulation, all the 50 Reading Comprehension questions were arranged according to their levels of difficulty the basis for which was Schraw and Roedel's three levels of difficulty. In their report, Schraw and Roedel validated the use of a band to identify the levels of difficulty of test items based on the success rate of the entire research population. Success rates were determined based on how difficult all the 43 respondents found each question. In this study, an item that had 30%-50% of 43 students responding to it correctly was labeled *Difficult (D)* while an item that had 50%-70% of 43 students responding to it correctly was marked *Moderately Difficult (MD).* An item that had 70%-90% of 43 students responding to a question correctly was classified as *Easy (E).*

Third, the researcher grouped the scores based on the question types and amended the band by organizing the levels of difficulty into four. The fourth level was added to accommodate scores of students that did not fall under the three levels included in Schraw and Roedel's study. Grouped this way, the scores and, implicitly, the students' reading skills were explored and analyzed. In the analysis, the examinees' test performance was discussed in light of existing language learning-teaching issues and theories and the researcher's general familiarity with the examinees' knowledge and/or lack of knowledge of test-taking skills.

## Data Analysis

To confirm research findings, the researcher also triangulated her data with two other research instruments, namely, focused interviews with selected students and retrospective journal entries vis-à-vis an ethnographic observation of said interviewees. This way, biases arising from interpretations based on pure assumption are done away with.

To fulfill one common-knowledge principle for testing which is to identify a student's areas of relative strength and weakness in subject areas, analyzing and making sense of what test results imply proves to be one of the most fundamental steps. In analyzing the results of the practice test taken by the subjects in this study, the researcher initially considered the use of basic statistical analysis. After tabulating the scores, however, she and her statistician decided that even a simple tabulation of scores and other related information triangulated with the use of other research instruments, such as focused interviews and retrospective journal entries, would be enough to provide her with the most salient data needed to arrive at significant statements in conjunction with the research objectives of said study. Focused interviews provided the researcher students' insights about the types and nature of questions they found difficult. Said interviews allowed her to look further into why they found some questions difficult to tackle. Taking advantage of the privileged position in noting down every single detail one can get from an insider's perspective, the researcher also kept journal entries to substantiate findings and interviews.

To check on the subjects' performance on the test, the researcher tabulated the scores and ranked them based on the correct and wrong responses they gave. This would allow anyone to see that the test items could be interpreted in isolation from the others despite the fact that they came in clusters. This means that even if one set of questions was based on one same passage, each test item could be interpreted independently as answers to the succeeding questions were mutually exclusive and not dependent on the previous one. Having ranked the scores, the researcher then identified the level of difficulty of each test item by using the categories used by Schraw and Roedel in their study. As discussed previously, because the categories in Schraw and Roedel's study (1994) were only comprised of three levels of difficulty, those whose scores that fell lower than what was originally cited in the 1994 study were put under another level, the result of the researcher's modification of the band. In so doing, scores that fell between the 10%-30% range were classified as *Extremely Difficult (ED).*

## Results

Results show that the spread of scores was very wide illustrating that the respondents' performance on each question type was not consistent. Because of this, the researcher deemed it more practical to offer an analysis and interpretation of results based on the correct

responses given per type of question rather than on the overall ranking of the test items irrespective of question types.

To address directly the goal of this study, the foregoing discourse is hinged on a number of diagnostic and pedagogic objectives the researcher found to be achievable. There are some diagnostic and pedagogic objectives, however, which the researcher reckoned not attainable as the data collected were not sufficient to provide substantive analysis and interpretation. A more comprehensive study on the respondents' test-taking and learning strategies may be conducted later to attain some of these goals. As such, issues surrounding the third and the fourth diagnostic objectives are not dealt with in the succeeding discussion. On the other hand, even if the first and the third pedagogic objectives are directly related to the respondents' test-taking and learning strategies, this study, nonetheless, offers to provide information on them which, although not necessarily directly related, is allied, at the very least.

The one last, but equally important, issue that readers of this paper need to be reminded about is the testing conditions under which the respondents of this study were asked to go through. In the actual TOEFL exam, test takers are asked to complete all the three sections of the exam in one meeting which normally takes a little over two hours. No break in-between test sections is allowed. In this study, however, in as much as the proctors/lecturers of the practice test tried to establish and maintain the same testing conditions observed during the actual TOEFL exam, the practice test was divided into two, the first two sections of which, namely, Listening and Structure and Written Expression, were taken in one regular classroom meeting. The Reading Comprehension section was taken the following day, during the regular classroom session, too. This set-up needs to be factored in because slight differences in testing conditions are normally anticipated to have effects on test-takers' performance and test results.

The presentation of the results of this study will be two-pronged such that the question type and the levels of difficulty of each type will be provided to give a better and clearer context.

Question type 1 is on *Facts & Details*. Questions of this type are those that require answers that are directly stated in the passage. Ordinarily, answers to these questions may be arrived at even without having to draw a conclusion about a text read.

Under the *Facts & Details* test type are four questions which were found to be *Extremely Difficult* with the top-ranking question yielding a mere 16.28% success rate among the respondents. Consistently, this item also ranked 2nd in the overall ranking of test items.

Although the test items seemed to be knowledge questions at first, their very nature actually asked examinees to go beyond recognizing details in the passage. Apart from the tabulations demonstrating items 24, 26, 27 and 31 as *Extremely Difficult* ones, ethnographic observation further revealed that the reason why the respondents found them as such was that these questions required them to synthesize and discriminate information, a higher-order thinking skill the researcher conjectured the

respondents had not acquired yet at the time of the test. Apart from the lower-order thinking skills details cited in both the stem and the options, which fell under Knowledge and Comprehension (Bloom, 1956), the test takers were supposed to take note of higher-order thinking skill points that particularly required them to do an analysis and a synthesis before the correct answer for each could be arrived at. On top of this, it is also important to note that while the examinees found said four questions to be *Extremely Difficult*, the success rate of the students on each test item varied both within the same level of difficulty and across the other levels. This means that the variation actually stemmed from the degree of specificity of each question checking on how familiar each respondent was with the language and content of each question regardless of the students' being homogenously grouped in the PC Program and the homogenous kind of instruction they received prior to taking the test. Question 24 came out with the least rate of success at a meager 16.28% making it more difficult by 10% compared to Questions 26 and 31. Questions 26 and 31 were of equal rank turning up 25.58% only of the 43 respondents who exhibited mastery of said skill at the time of the test. Question 27, with a success rate of 27.91%, indicated that knowledge of this question was slightly higher by 2.33%.

Questions 18, 45, 4, 6, 9 and 22 comprised another sub-category of a *Facts & Details* question type. These items were considered *Difficult*. The researcher gathered that although the questions merely asked for facts, generally speaking, they were not simple *Wh-* questions. She further gathered from focused interviews that additional test item descriptors also known as expanders such as "...what common characteristics *distinguished the careers* of the Mayo brothers?" added to the degree of specificity of the question that the respondents had to deal with as opposed to the usual *Wh-* question which could have gone as simple as this: "What common characteristics did the Mayo brothers have?" Following the expanded synthesis question format, Questions 18, 4, 9 and 22 appeared complex for the examinees to readily understand and answer correctly. Another test item, question 45, on the one hand, was difficult, nevertheless, because although the descriptor came out short and simple, the options required the examinees to process and eventually analyze each of them more thoroughly. The analysis required by each option made the item seem a trick question for them. Question six, on the other hand, no matter how simple and straightforward the question was, came tricky as well because the students had to be very keen on details. When asked, students said that the error may be traced back to their not having checked each distractor very carefully.

Only two *Facts & Details Questions* fell under another sub-category, namely the Moderately *Difficult Level,* with success rates of 55.81% and 69.76%. Despite having a difference of 14%, Questions 38 and 16 were both categorized as *Moderately Difficult* items only primarily because the degree of specificity of the descriptors did not appear to be as high as those of the questions found in the *Extremely Difficult* and *Difficult* test items. Although the nature of these questions did not vary

much from the previous ones (questions 24, 26, 27 and 31*)* labeled as *Extremely Difficult* and *Difficult*, the additional descriptors and expanders of these questions were simpler and fewer compared to those of the former such that the underlined words in the questions"…what purpose do the fine hairs *on the body of the bee* serve?" and "… what does marketing research *include*?"posed to be much easier to comprehend. On this, the respondents expressed their preference for simpler and shorter questions.

Questions 39 and37 came out to be another sub-category. They placed under the *Easy Level*. Although the basic aim of this question was to require the examinees to extract facts and details from the text, tabulations indicated that these two questions turned out to be *Easy* for them. The researcher's observation revealed two possible two reasons, later confirmed by the respondents during the focused interviews. First, both the stem and the options—the distracters and the best answer—were stated in simple form. Second, both questions were plain knowledge questions—those that touch on the lower-order thinking skills—asking the examinees to recall information from the text, virtually the lowest form of task test takers normally do cognitively.

The second type of question has to do with making an *Inference*. An inferential question asks an examinee to comprehend an idea or argument that is strongly implied but not directly stated in the text.

The question type with the third highest number of items included in the test—inference questions 44, 49, 3, 21, 47 and 34—were found to be *Extremely Difficult* and ranked as the second most difficult ones. Unlike questions in expanded form, those which asked the test takers to integrate additional pieces of information provided in the stem of the question, questions 44, 49, 3, 21, 47 and 34 were stated following a much simpler format. Despite the simplicity of their format, however, these questions turned out to be the second most difficult questions the examinees found. This difficulty can be attributed to the fact that "inferential comprehension questions measure interpretation. These items require one to 'read between the lines' or even 'beyond the lines' combining past knowledge and familiarity with text information. These were synthesis questions, the highest form and most challenging one based on Bloom's Taxonomy of Learning (1956).

In the case, however, of the next two groups of Inference Questions based on the levels of difficulty aptly classified as *Difficult* and *Moderately Difficult*, although considered also as *Inference Questions*, they did not turn out to be as complicated as the others for two interconnected reasons. First, the questions simply asked the test takers to infer information that could be easily traced to simple facts and details cited in the passage. Second, because of this, the options were not as abstract as was the case of the questions 44, 49, 3, 21, 47 and 34.

In other words, the nature of the inference questions such as items 42, 29, 33, 15, and 36 required the use of information that was more explicitly stated in the passage. It was a less difficult task required by the previously mentioned items, which was to merge concepts and produce something implied in the text.

The third type of question is about identifying the *Main Idea* of a given text. This is a type of question about the overall idea expressed in the passage. All this asks is the primary point which the writer is trying to convey.

Only three *Main Idea Questions* were included in this test. Because of the small number of questions, results particularly involving said test items cannot be considered conclusive. Nevertheless, the use of said items in this test was deemed significant if only to establish a baseline data unique to the respondents' learning context. Of the three *Main Idea Questions* included in this test, only one item turned out to be *Extremely Difficult,* such that only 25.58% got the correct answer. This question, item 12, was a main idea question which turned out to be difficult. Apart from having to synthesize the meaning of said question, students ultimately had to make connections between the text and a related subject matter. In so doing, the test takers were actually asked cognitively to use the information they read about in another related context or situation. Interview results revealed that they were not used to making and/or establishing connections and patterns the way they were asked to do on the test. Part of the difficulty may also be attributed to the students' having been so used to simple Main Idea Questions such that requiring them to make associations seemed too difficult a task for them to fulfill at the time. This confirms one of the findings of Wisaijorn's study (2008) concerning features of Thai education saying that "students may find it difficult to develop skills in creative thinking, independent and alternative learning, questions and/ or discussion."

Interestingly, the gap between Question 12 (*Extremely Difficult)* and Question 1 (*Difficult*), and Question 1 (*Difficult*) and Question 50 (*Moderately Difficult* ranged from approximately 13.95% to 13.96%, a consistent difference across three levels of difficulty. What can be seen as a give-away in Question One was that the format of the question was the most common the test takers normally encountered especially in their practice tests, seat works, and assignments. It did not come out as a surprise that they found the same line of questioning easy to handle. Also, the examinees found Question 50 shown above clear-cut and, therefore, easy. Both the stem and the options were precise such that matching the test item and the answer did not give the test takers a hard time.

*Organization* is the focus of the fourth type of question. This question asks a test-taker to determine how the ideas in one paragraph relate to the ideas in another paragraph. In this sub-category, Question 11 turned out to be *Difficult*. It may be assumed at first that the item, question 11, looked easy to handle, as it only required the examinees to identify the exact location of the information in question. Tabulations disclosed, however, that this seemingly easy question was actually extremely difficult for the examinees in general. Students' reactions showed that this difficulty can be traced primarily to how the stem was crafted. While it mainly tested a test taker's knowledge of the organization of the text, it turned out more complicated than expected because of its expanded format no matter how simple the question was. It required the

examinees to look for some pieces of information considered crucial to determining the exact location of the data in question. The students said that the test item seemed too long a string of a question. As such, the researcher gathered that "for the expansion" and "of the practice" expanded what could have been a simple one. Such a difficulty was very similar to what the test takers commented on regarding test items 18, 4, 6, 9 and 22.

If students could not identify and/or establish the relationships between and among the parts within a long string of a question, finding the correct answer to it then was simply impossible. This explains as well why this particular test item, although it virtually asked test takers to perform the same task similar to what the students did on Question 32, proved to be more difficult.

Another sub-category of a test type about *Organization* is the *Difficult Level*. Although slightly different in terms of percentages, questions 43 and 32 were found to be *Difficult.* The two test items below, of course, are not exactly of the same format. Question 43 straightforwardly asked the examinees to identify the possible source of the passage in question thereby requiring them to make some judgment and eliminate the distracters in favor of the best answer. Question 32, on the other hand, simply asked the test takers to map out the passage to enable them to figure out where exactly a piece of information in question could be found. Different the questions might had been, the examinees found both of them difficult. What appeared to be more interesting, however, was that Question 43 turned out to be more difficult than Question 32, the reason for which might be attributed to the nature of the former. It did not only simply make the test takers think about the plausibility of all of the options. It also required them to do the following: One, to give meaning/s to each of the options; two, to judge which of the options made the most sense, and; three, to eliminate the distracters so as to choose the best one among them. Required to do all three sub-tasks to answer a single item, the students admitted that the steps seemed too daunting a task for them to do. This result came out to be in synch with the researcher's observation of the students' performance on the test.

The fifth type of question is on *Referents*. Typically, this type of question asks an examinee to determine which noun a pronoun refers to and/or vice versa. Although of different percentages, Questions 23 and 20 both emerged as *Easy* items. This could be attributed to two main reasons: one, the students have always had this type of question in their reading quizzes hence their level of familiarity with it was high, and, two, this type of question fell on the knowledge category of questions in reference to Bloom's Taxonomy. The use of knowledge questions in this part of the test, items that checked on and activated the students' lower-thinking skills, required students to make a simple matching of the term in question and its referent.

The sixth question type asks about *Tone & Purpose.* A question about the 'tone' requires the examinee to identify the emotion the writer is

trying to show in the text whereas a question about the 'purpose' asks what the author is trying to do in the passage.

Although what can be said about this type of question could not be taken conclusively as it was the only item of this kind in this test, giving it a serious thought may greatly help future studies. It ranked, however one puts it, as the 3rd most difficult question making it an *Extremely Difficult* item with only 20.93% success rate. Similar to how the examinees performed on items that tested their ability to make good inferences, which were classified either as *Difficult* or *Moderately Difficult*, this particular question that checked on the students' ability to sense the Tone and Purpose of the text and/or author proved to be one item they needed to put more attention to. This runs consonant with the researcher's observation and interview results that because the respondents were not in the habit of reading texts in English—aside from what was given them within the classroom setting—they had very limited background knowledge. Having rich background knowledge, on the other hand, could have been very helpful in their understanding of the author's intention in the text they read. Schema theory states that "reading comprehension is an interactive process between the text and the reader's prior background knowledge" (Adams and Collins, 1979; Rumelhart, 1980). This leads to an understanding that their schema at the time of the test was not sufficient to help them understand better the texts they read.

What seemed to have compounded the situation was the fact that the test-takers not only had to sense the tone and purpose behind the text. They also had to provide a reason, a task that required them to approach the item employing their cognitive skills. In so doing, they needed to give meaning to each of the options requiring them to synthesize each before eliminating the distracters. To add to that, the stem itself of question 19, which used highly specific terms such as "legally binding agreements", if not comprehensible and familiar enough to the test-takers could have been a major source of difficulty.

The last test type is on *Vocabulary.* These items basically ask for word meanings doing which can be done in a number of ways, namely, using structural clues, understanding meanings from word parts, and finding definitions from context clues.

Based on at least three consecutive actual TOEFL exams administered in the past, the PC TOEFL lecturers observed that *Vocabulary* questions, followed by *Facts and Details*, had the most number of questions in the test. It is actually a trend in practice tests that holds true even in the actual TOEFL-ITP exam. This explains partly why *Vocabulary* questions also had the most number of questions in the *Extremely Difficult* category. Of greatest importance, however, in the discussion of the examinees' performance on this test are the following: (1) the word "consistent" turned out to be the most difficult word yielding a 13.95% success rate, and; (2) eight vocabulary items out of 21 questions occupied the *Extremely Difficult* category.

Based on their levels of difficulty and starting from the most difficult were vocabulary items that ranked respectively as first, with

13.95% success rate (consistent), second, with 16.28% success rate (integrated and outlets), fourth, with 23.26% success rate (queries), fifth, with 25.58% success rate (detect), sixth, with 27.91% success rate (contributions and clustered), and seventh, with 30.23% success rate (stationary).

Words such as "subsequent" (question 10) and "absorb" (question 14) ranked 10th giving the examinees a success rate of 37.21%. The word "dedication" (question 5) with a success rate of 39.53% ranked 11th. All the three words were classified as *Difficult*.

The word "narrows" (question 25), with a success rate of 51.16%, ranked 16th whereas the phrase "accounts for"(question 17), with a success rate of 65.12%, ranked 20th. Based on the students' scores, both vocabulary items were *Moderately Difficult*.

Three vocabulary items turned out to be easy for the test-takers, namely, "statistics" (question 35), "hues" (question 41), and "founded" (question 7) with the first two terms ranked as 23rd at 79.07% and the third word ranked 24th at 81.4%.

## Discussion

In the tabulation showing the overall ranking of the test items contrasting its difficulty against the rest of the items, a combination of six different types of questions forms part of what the respondents found to be *Extremely Difficult.* On top of the list is *Vocabulary* having only a total of six students responding to it correctly, i.e., the item was *easy,* the equivalent percentage of which is 13.95%. Based on Schraw and Roedel's levels of difficulty, said question came out to be an *Extremely Difficult* one for 86.04% or 37 students of the entire TOEFL population for Quarter 2-2009. Still based on same band, the other types of questions the subjects found extremely difficult were *Facts & Details*, *Tone & Purpose*, *Inference*, *Main Idea*, and *Organization* although not necessarily in this exact order. This does not mean though that these types of questions were no longer classified again under any of the three other levels of difficulty, namely, *Difficult, Moderately Difficult,* and *Easy.* On the contrary, because of the huge number of questions for the entire Reading Comprehension Section, most of the question types were reclassified across the 50-item Reading Comprehension Section. To illustrate, although *Vocabulary* topped the rank within the *Extremely Difficult* level, the respondents found other *Vocabulary* test items to be *Difficult*, *Moderately Difficult*, or *Easy* depending on the degree of specificity of the question.

What appeared critical, in as far as vocabulary items were concerned, was that the examinees' performance, generally speaking, posed as a reminder for students to pay more attention to their vocabulary, if only to better their scores. Evidently, with eight vocabulary items dominating the *Extremely Difficult* category, it may be safe to assume that poor vocabulary must have contributed to their not having answered correctly most other items albeit these items followed different formats and tested other reading skills. Steven Stahl, in his 1999 study,

categorically stated how strongly correlated reading comprehension and vocabulary are. Cynthia and Drew Johnson (2004) further asserted that "Limited vocabularies prevent students from comprehending a text." This, however, is not surprising as it can be taken as a direct consequence of the students' not being fond of reading texts in English.

With the findings considered unique to this specific group of students, it is nevertheless equally interesting to know what the general levels of difficulty each type of question had. Although all of the 50 items performed differently as the questions were mutually exclusive from each other, establishing and studying the general pattern the questions took is worth a look. As some question types appeared several times within the *Extremely Difficult* category, their sequencing was done based not on the frequency of the items, but on when each type appeared the first time within said category. As such, *Vocabulary* items topped the list as the most difficult item followed by *Facts and Details*. The third on the list was *Tone and Purpose*, fourth of which was *Inference*, fifth was *Main Idea*, sixth was *Organization*, and last was *Referents*.

While this study upholds the unquestionable assistance every language teacher can extend to language learners, the findings of this study nevertheless espouse the pedagogical merit of looking into the very specific circumstances surrounding language learning issues that are not always the center of attention in the everyday classroom.

In view of the current practices in most reading and TOEFL preparation classes, this study underscores the following points:

First, the subjects' main difficulty was on vocabulary primarily because (1) their lexis was insufficient, and (2) they either had no or lacked knowledge of the fact that word meanings change in different contexts.

Second, questions in expanded format, especially those that required the respondents to use higher-order thinking skills such as synthesis and analysis, proved to be difficult as combining new ideas to form a new whole involved cognitive and meta-cognitive tasks they could not do with ease.

Third, inferential questions, generally speaking, were difficult to tackle not only because of the examinees' poor vocabulary, but also because of their limited schema making them unable to make sense of what the texts indirectly said.

Fourth, making associations between a text and a related subject matter outside the passage was a difficulty which may be attributed to their poor world/background knowledge.

Just as the study offers its findings to be beneficial to some identified sectors in the field of language teaching, it also provides a number of suggestions that may be undertaken in exploring other language teaching-learning possibilities.

This study recommends the following undertakings relating to vocabulary enrichment: students need to expand their vocabulary both through explicit vocabulary instruction and sustained outside reading; students have to be taught about the "pervasiveness of contextual

variation in meaning" (Nagy, n.d.) and be trained how to recognize these changes so as to raise their level of awareness.

To address the students' need for exercises that address the higher-level thinking skills, this study proposes that more synthesis and analysis exercises be given to the students to develop in them the ability to put ideas together to create a sound outcome. Teachers should demonstrate how to tackle a question in expanded format.

Given the fact that schema development or enriching the students' background knowledge remains to be a challenge, this study supports the move to get students more exposed to texts with implied meanings. Teachers should help learners identify implied meanings in texts first, by demonstrating, and, second, by allowing them to try it on their own until they arrive at a desired level of reading writers' implications. Furthermore, students should be more exposed to readings rich in world meanings aimed at deepening their world/background knowledge. Teachers should be able to demonstrate to them how world/background knowledge can be used to better understand various texts.

As there was insufficient data gathered, limitations were met which did not allow the researcher to study the learners' test-taking strategies. This paper instead suggests that another investigation looking into said area through the use of verbal protocol be done in the future to further confirm results of this study; as propelled by the constraints encountered while gathering data, test items used were not authentic TOEFL questions, and, as such, some possible differences might have led to disparity in findings as opposed to using genuine TOEFL questions. A replicate study may be done in the future instead using, if possible, real TOEFL test items. With respect to the above recommendation, should a replicate study be done in the future, it is important to note that exactly the same testing conditions be observed so as to achieve an experience closest to that of, if not the same as, the actual TOEFL exam.

Lastly, this study recommends that findings and conclusions of this study be taken as something unique to the experience of the research participants and not an absolute representation of the entire PC population or of the MUIC students taking TOEFL as part of the university's admission requirements.

In sum, this study confirms the importance of paying close attention to some very specific language difficulties second language learners encounter especially when said issues are left unattended outside the classroom context, providing learners with no support system they so need.

## References

Abanomey, A. in Cohen, A. (2006). The coming of age of research on test-taking strategies. *Lawrence Erlbaum Associates, Inc.*, *3*(4), 307-331.

Adams, M., & Collins, A. (1979). *A schema-theoretic view of reading. New directions in discourse processing.* Norwood, NJ: Ablex.

Alvermann, D. E., & Guthrie, J. T. (1993). Themes and directions of the National Reading Research Center, *National Reading Research Center*, *1*, 1-11.

Andrew, C. (2006). The coming of age of research on test-taking strategies. *Lawrence Erlbaum Associates, Inc.*, *3*(4), 307-331.

Bloom, B.S. (Ed.) (1956). Taxonomy of educational objectives, the classification of educational goals- Handbook I. *Cognitive Domain.* NY: McKay.

Chareonwongsak, K. (2007, July 7). Ten dimensions for new Thai thinking skills. *Bangkok Post*.

Cohen, A. (2006). The coming of age of research on test-taking strategies. *Lawrence Erlbaum Associates, Inc.*, *3*(4), 307-331.

Dole, J. (2004). The changing role of the reading specialist in school reform. *The Reading Teacher*, *57*(5), 462-471.

Fraser, C. (1999). Lexical processing strategy use and vocabulary learning through reading. *Studies in Second Language Acquisition*, *21*, 225-241.

Hamp-Lyons, L., & Davies, A. (2002). Bias revisited. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *4*, 97-108.

Henk, W., & Melnick, S. (1995). The reader self-perception scale (RSPS): A new tool for measuring how children feel about themselves as readers. *The Reading Teacher*, *48*(6), 470-482.

Johnson, C., & Johnson, D. (2004). The importance of vocabulary development. In *The Wordly Wise 3000 Teacher's Guide for Books 1-5*. The Educator's Publishing Service.

Marukatat, S., & Bunnag, S. (2003, May 28). Region can learn from Thailand? *Bangkok Post*.

Nagy, W. (n.d.). *On the role of context in first-and second-language vocabulary learning*. Cambridge: Cambridge University Press.

Qian, D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning Journal*, 52(3).

Rumelhart, D.E. (1994). Toward an interactive model of reading. In R.D. Rudell, M.R. Rudell & H. Singer (Eds.), *Theoretical models and process of reading* (4th ed.) Delaware: International Reading Association.

Schraw, G., & Roedel, T. D. (1994). Test difficulty and judgment bias. *Memory & Cognition*, *22*, 63-69.

Snow, C. E., Griffin, P., & Burns, M. S. (2005). *Knowledge to support the teaching of reading*. NJ: John Wiley and Son.

Wisaijorn, P. (2008). Strategy training in the teaching of reading comprehension: Does it work for students whose first language is NOT English? *Chulalongkorn University Language Institute*.

## About the Author

Analiza Perez-Amurao obtained her AB-BSE in ELT from the Philippine Normal University in 1992. In 2006, she finished her MA in English Language and Literature Teaching at the Ateneo de Manila University. That same year, she obtained her Postgraduate Diploma in TESOL from RELC-Singapore. She currently teaches and acts as Coordinator at the Preparation Center for Languages and Mathematics at the Mahidol University International College, a leading state university in Thailand. An article she wrote on the theme "Innovation in Education" placed 5th in the recently concluded 2010 SEAMEO-Australia Press Award.

© 2011 Time Taylor International ISSN 2094-1420