國立臺灣師範大學英語學系

博　士　論　文

Doctoral Dissertation

Department of English
National Taiwan Normal University

教師命題過程與學生答題過程研究

Exploring Teachers' Test-constructing Processes and Students' Test-taking Processes

指導教授：程玉秀博士

Advisor: Dr. Yuh-show Cheng

研 究 生：曾繁萍

Fan-ping Tseng

中 華 民 國 一百零三 年 八 月

August, 2014

中文摘要

本文旨在研究以下三個問題：第一，高中教師如何命一份英文學科能力測驗的模擬試題？資深教師與新手教師在命題時的考慮點有何不同？第二，高中學生如何回答英文學科能力測驗模擬試題的題目？高程度學生與低程度學生的答題策略有何不同？第三，學生答題時的考慮點與教師命題時的考慮點是否一致？

四位高中英文教師及四十八位高中學生參與此研究。教師的任務是要命一份英文學測模擬試題，內含詞彙測驗、綜合測驗、及閱讀測驗等共二十八題選擇題；學生的任務則是要回答教師所命的模擬試題題目。所有參與者在執行任務時，必須要進行有聲思考法，以作為本研究的主要分析資料。

本研究主要結果如下。首先，資深教師與新手教師在命題時的考慮點略有不同；資深教師的命題考量較以學生為中心，而新手教師的命題考量則較符合評量上的命題原則。此外，資深教師所命的試題並沒有優於新手教師；而且，在四位教師所命的題目中，有不少試題是被專家評定為有瑕疵、不合適，並需要修正及改進的。

其次，學生在作答不同類型題目時，大致上會採用不同的策略。然而，學生在作答三種類型的題目時，均有使用「消去法」。此結果顯示，消去法乃學生在本研究最常使用的答題策略。另外，高程度學生比低程度學生較常使用字彙及文法知識和演繹思考法來作答；而低程度學生比高程度學生較常利用「猜測法」來回答任何類型的題目。

研究也發現，學生的答題考慮點與教師的命題考慮點大不相同，兩者的一致率只有 33%。此外，學生的想法和新手教師的想法較一致，而和資深教師的想法較不相同。高程度學生在綜合測驗的答題考慮點上，和教師們的命題考慮點出入較大；而低程度學生在閱讀測驗的考慮點上，和教師們的考慮點不一致性較高。

關鍵字：試題命製、試題命製過程、答題過程、策略運用、字彙測驗、綜合測驗、閱讀測驗、有聲思考法

ABSTRACT

This study aims to investigate three research questions. First, how did experienced and novice teachers construct mock tests for the Scholastic Ability English Test (SAET)? Second, how did higher- and lower-proficiency students take those mock tests? Third, were students' considerations for answering the tests consistent with teachers' test-constructing considerations?

Four senior high school teachers and forty-eight senior high school students participated in this study. All participants were asked to do think-aloud while performing their tasks. The teachers were asked to construct twenty-eight items of multiple-choice questions on vocabulary, cloze, and reading comprehension. The students were asked to answer the questions constructed by the teachers.

Major findings of this study are summarized as follows. First, the experienced teachers and novice teachers seemed to make different types of considerations in constructing their tests. The experienced teachers took more student-oriented factors into account while the novice teachers took more test-construction principles into consideration. Despite their different considerations in test-constructing processes, the two experienced teachers did not seem to produce better test items than the two novice teachers. All four teachers had constructed some items that were deemed poor, problematic, or inappropriate from the authority's perspective.

Second, students generally used different strategies when answering different types of questions. However, they seemed to use the strategy of "elimination" very frequently on three types of tests. In terms of the proficiency levels, higher-proficiency students tended to use their vocabulary knowledge, grammar knowledge, and deductive reasoning more frequently than lower-proficiency students in answering the items. On the other hand, lower-proficiency students tended to use the strategy of "guessing" more frequently than higher-proficiency students across three types of questions.

Third, students' considerations for answering test items clashed with teachers' test-constructing considerations to a great extent; the overall consistency rate between them was only about 33% in this study. Furthermore, students generally thought in a way more congruent with novice teachers than with experienced teachers. In addition, higher-proficiency students' considerations clashed more with teachers' considerations on cloze items while lower-proficiency students' considerations clashed more with teachers' considerations on reading comprehension questions.

Key words: test-construction, test-constructing process, test-taking process, strategy use, vocabulary test, cloze test, reading comprehension test, think-aloud

ACKNOWLEDGEMENTS

I owe a great deal to many people who have given me love, support, and assistance during my long years of doctoral study at NTNU. Without them, I would never have got this degree, nor would I have been able to complete this dissertation.

My first sincere gratitude goes to my esteemed advisor, Dr. Yuh-show Cheng, for her clear guidance and valuable advice for me during these years of writing this dissertation. If it had not been for her constant and timely encouragement, I would have given up, not to mention finishing this research project. I am truly grateful for everything she had done for me at so many critical moments.

I am also indebted to my exceptional committee members, Dr. Chiou-lan Chern, Dr. His-nan Yeh, Dr. Vincent Wu-chang Chang, and Dr. Hsueh-ying Yu, for their insightful comments and constructive remarks on my dissertation. Their valuable suggestions have contributed a lot to the improvements of this dissertation.

I would also like to thank all the excellent professors who have taught me during my studies at NTNU. Thank you for leading me into the research field, and thank you for nurturing me so that I can become what I am now. These eminent professors are Dr. Vincent Wu-chang Chang, Dr. Yuh-show Cheng, Dr. His-nan Yeh, Dr. Chiou-lan Chern, Dr. Hsi-chin Chu, Dr. Howard Hao-jan Chen, Dr. Wen-Ta Tseng, Dr. Ho-ping Feng, Dr. Chyiruey Chen, Dr. Chien-Jer Charles Lin, and Dr. Miao-Hsia Chang.

My deepest appreciation also extended to all the high school teachers and students participating in this study. Their precious time and effort devoted to this research work helped make this dissertation a reality. I would also like to thank Dr. Hsueh-o Lin and Dr. Jung-Han Chen for their critical comments on the test items used in this study. I greatly appreciate their generous help.

I owe a debt of gratitude to my family for their love and unfailing support throughout my study. I would like to thank my dearest mother, sister, brother, sister-in-law, and niece, who took the sweet burden of taking care of my aging father most of time while I was away working on my dissertation. Their unselfish sacrifice gave me the extra time to take on such a demanding research project, and thus contributed a lot to the completion of this dissertation.

I would like to express my profound gratitude to my husband, son, and daughter for their faithful love and support during my long years of study at NTNU. Although they did not quite understand what I have been working on these years, they still support me wholeheartedly. Without their company and encouragement, I would never have the strength and motivation to finish this dissertation. Dear, I love you all.

Finally, may all glory be to God alone! *Soli Deo gloria!* To my Lord in heaven, I dedicate this dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Motivation and Background

Tests seem to play a major and prominent role in Taiwan's high school language classrooms. As an English teacher in a senior high school in Taiwan, I find both myself and my students constantly facing language tests of all kinds in a semester, such as class quizzes, weekly tests, midterms, finals, etc. Among these different tests, midterms and finals are considered most important by students, since these are formal, school-required exams, the results of which profoundly affect their academic records. Thus, students would work very hard to prepare for the exams. After they take the exams, the results are always analyzed quantitatively, with different scores presented to teachers and school authorities for comparison. However, I don't think the raw scores show us enough information about the students' understanding of what is being tested on the exams. In other words, students' test scores may not faithfully represent their true language abilities; there might be some other factors, such as guessing, involved. This assumption aroused my interest in how students take the exams. I think that the results of the investigation into students' test-taking process would add substantial meanings to test scores.

The school-required exams (i.e., midterms or finals) are also taken seriously by teachers, who are mostly responsible for preparing the exams. In my school, the formal exam is prepared by one teacher alone. Most of the time, the teacher who is assigned to construct a formal exam is under great pressure, and I am no exception. As far as I am concerned, constructing a formal exam is no easy task, and the test-constructing process is a laborious one. Since I am always struggling through the test-constructing process, I am curious about how other teachers go through such a process and what factors they take into consideration when they construct a test.

My inquiries into how teachers construct a test and how students take a test led me into the ample and diverse research of the language testing field. The blossoming language testing research has covered many aspects of the testing practice, such as test types, test validity, test-scoring methods, test-takers, and test-taking processes, to name just a few. Taken together, the abundant research body seems to center around two major themes: one on issues concerning "tests," and the other on issues regarding "test-takers" (the recipients of tests). Yet tests are not born in a vacuum; instead, they are produced or written by teachers or researchers (the initiators of tests). But, to my surprise, research on the part of the initiators of tests or on how tests are constructed receives little attention, leaving not only a missing piece to the testing field but also a large, potential area for further investigation.

Bachman (2000), in his state-of-the-art article on language testing, concludes that he believes "there are two areas in which language testing and language testers must continue to grow and develop: the professionalization of the field, and validation research" (p.18). I think my present research well accords with Bachman's (2000) arguments. For one, Bachman's first prediction, *the professionalization of the field*, has two major thrusts: "the training of language testing professionals; and the development of standards of practice and mechanisms for their implementation and enforcement" (p.19). I believe the results of my study on teachers' test-constructing processes might shed some light on teacher education curriculum, which is in line with Bachman's (2000) focuses. For another, the results of my research on students' test-taking processes might contribute some fruits to the test *validation research*, which is one of Bachman's (2000) major concerns.

Taken together, this study is motivated by the desire to resolve the puzzle in my teaching career as well as the possibility of bridging the research gap in the language testing field.

Statement of the Problem and Research Rationale

The present study consists of three aspects: (1) an investigation into how teachers construct tests, (2) an exploration into how students take tests, and (3) a comparison between teachers' test-constructing considerations and students' considerations for answering the tests constructed by the teachers.

In the whole testing research, the factor of test-constructors doesn't seem to receive its due attention compared with the issues concerning tests and test-takers. Yet, as Jafarpur (2003) points out, in a program where test-construction is one individual teacher's responsibility, the role of the test-constructor is more important than in a program where test-construction is carried out by a committee. I agree with Jafarpur's (2003) observation, since test-construction is usually one teacher's responsibility in many of Taiwan's English teaching contexts, especially in middle schools. Thus, research on test-construction might add insights into the testing practice in Taiwan. There have been some studies investigating teachers' test-constructing skills (e.g., Carter, 1984; Coniam, 2009), and some describing the training courses on test-construction (e.g., Kirschner, Spector-Cohen, & Wexler, 1996; Johnson, Becker, & Olive, 1999). Jafarpur (2003) took it further, exploring the test-developer as a facet of test variance. Among those studies, though Johnson, Becker, & Olive (1999) and Coniam (2009) have reported some teachers' reflections on their test development process after finishing the test items, no study, to my knowledge, has directly examined teachers' test-constructing process by using the think-aloud method. Therefore, I think it is worthwhile to have a thorough investigation and analysis of teachers' test-constructing process through the think-aloud method.

While investigating test-constructing process, I also hope to examine possible test constructor effect, especially length of teaching years, in test variance. Jafarpur's (2003) examination of teacher-produced tests shows that there was a test constructor

effect on the performance of test-takers using multiple-choice reading comprehension tests that had no specifications. Jafarpur's (2003) results aroused my interest in the test constructor effect on test variance using tests with specifications. Based on the research findings of rater effect on performance tests, I assume test constructor may have an effect on variance of multiple-choice tests. For example, Brown (1995) explored the influence of rater backgrounds (native/nonnative; with/without industry and teaching experience) on assessments in an oral test of Japanese for tour guides. Her results showed that there were significant differences in ratings awarded for some individual criteria, though there were no significant differences between different types of rater in terms of the overall grade awarded. In another study, Lim (2011) examined new and experienced raters' performance longitudinally over multiple time points in writing assessment. The results showed that novice raters, who initially differed in performance from their experienced counterparts, learned to rate appropriately relatively quickly, and that raters were able to maintain rating quality over time. Since studies such as Brown (1995) and Lim (2011) have suggested that there exists rater effect even with rating criteria provided, by analogy, it is reasonable to assume that there might also be test-constructor effect involved in test development even with test specifications provided. Therefore, in addition to investigating teachers' test-constructing processes, I want to explore how the tests produced by novice teachers differ from those produced by experienced teachers as well.

Unlike the scant studies on test-constructing processes, research into test-taking processes has received more attention, and has been recognized as part of the construct validation research (Anderson et al., 1991; Bachman, 2000; Cohen, 2006). Cohen (1984) was probably one of the early researchers in exploring L2 test-taking strategies through verbal report data. Later on, more studies followed the trend, and have shed some light on how students were actually thinking while they were taking

tests. For instance, Nevo (1989), Anderson et al. (1991), and Rupp, Ferne, and Choi (2006) investigated students' test-taking process in multiple-choice reading comprehension tests, while Storey (1997), Sasaki (2000), Yamashita (2003), and Moghaddam (2010) examined students' test-taking process in cloze tests. Among these studies, Yamashita (2003) also compared the test-taking processes of skilled readers and those of less skilled readers, and the results did show that the two groups adopted different information in answering the gap-filling cloze test. These test-taking process studies on L2 learners are valuable in that they have caught and described part of the cognitive processes while L2 learners were taking their tests. Yet, L2 learners are unique in each context, and their test-taking processes might vary from culture to culture. Since there have been no published studies, to the best of my knowledge, examining Taiwanese EFL learners' test-taking processes, I would like to investigate the actual processes of how Taiwanese high school students take their English tests. Moreover, motivated by Yamashita (2003), I would also like to examine whether there is any difference between test-taking processes of high-proficiency students and those of their low-proficiency counterparts.

In his pioneering study on L2 test-taking process, Cohen (1984) stated that "the purpose of such research has been to explore the closeness-of-fit between the tester's *presumptions* about what is being tested and the *actual* processes that the test taker goes through" (p.70). Later, Nevo (1989), also commented that "the examiners' assumptions regarding what they test and their expectations from the respondents often do not match the actual processes which the respondents undergo during testing" (p. 200). Both researchers have pointed out the phenomenon that test takers may answer test items in ways different from what test constructors have expected. Nevertheless, it is a pity that after two decades, there is still little research empirically examining the degree of fitness between test-constructors' considerations and

students' test-taking considerations. One of such rare studies is Gierl (2001), which compared cognitive representations of test developers and those of students on a mathematics test. As far as I know, no such similar comparison has been conducted in the language testing field. Hence, I think a comparison between EFL teachers' test-constructing considerations and EFL students' considerations for answering tests would be a research line worth pursuing.

Given the motivation and research rationale stated above, I would like to investigate EFL teachers' test-constructing processes and EFL students' test-taking processes by using the think aloud method. In addition, I will also compare teachers' test-constructing considerations and students' considerations for answering the test items produced by the teachers.

Purpose of the Study

The purpose of the study is threefold: (1) to investigate how EFL teachers construct tests; (2) to examine how EFL students take tests; and (3) to explore whether there is any match or mismatch between teachers' test-constructing considerations and students' considerations for answering the test items produced by the teachers.

The present study was situated in a context I am familiar with. In other words, the participants in the study were Taiwanese senior high school teachers and students, and the tests they constructed or took were mock tests for the Scholastic Ability English Test (SAET). The reasons for selecting the SAET as the research tool are as follows. To begin with, the SAET is a nationwide test administered by the College Entrance Examination Center (CEEC). All Taiwanese senior high school students are familiar with the test since they have to take either the SAET or the Department Required English Test (DRET), the other important test administered by the CEEC, to enter university or college. Since the SAET, usually held in January, occurs prior to

the DRET, held in July, most students will take the SAET, and some will skip the DRET if they are admitted to the universities they want with their SAET grades. The statistics released by the CEEC also indicate that the number of students taking the SAET is usually much larger than that of students taking the DRET[1]. Thus, the SAET is considered a very important college entrance exam by both the senior high school teachers and students. I believe that studies on examining how teachers and students treat the SAET would yield more valuable fruits than those examining other tests or exams.

Although I regard the SAET as a crucial entrance exam for Taiwanese senior high school students, the present study probed into the processes of how teachers constructed "mock tests" of SAET and how students answered them. The main cause for this substitution of mocks tests for the real test is that I do not have the legitimate access to the real SAET, which is usually prepared by the CEEC committee. Hence, it is impossible for me to employ the real SAET as my investigation instrument. Despite this, I think the mock test of SAET can still serve a good purpose for this study for the following three reasons. First, my main research goal is to explore the processes of test-constructing and test-taking, not to examine the purpose of the test itself. Regarding this, even though the SAET and its mock test may serve different purposes, the former being an achievement test and the latter more like a diagnostic one[2], the differences between them would not influence my study to a great extent. Thus,

[1] The numbers of students taking the SAET and the DRET in the years of 2011, 2012, and 2013 are as follows:

(year)	2011	2012	2013
Number of students taking the SAET	146,302	154,560	150,030
Number of students taking the DRET	82,164	75,839	65,966

sources: http://www.ceec.edu.tw/AbilityExam/SatStat/學測歷年報名人數 1030103.pdf
http://www.ceec.edu.tw/AppointExam/DrseStat/102DrseStat/指考歷年報名人數 1020613.pdf

[2] A diagnostic test, according to Hughes (2003), is used to identify learners' strengths and weaknesses. Therefore, I think an SAET mock test serves the purpose of a diagnostic test because it shows students what their weaknesses are in preparing for the SAET.

having no access to the real SAET, I consider its mock test a good substitute. Second, among the ready-made SAET mock tests on the market, many of them, as I observe, are constructed by individual teachers instead of by a committee. This happens to resemble the testing practice in most English classrooms in Taiwan, where a test is usually prepared by an individual teacher alone. Consequently, I think investigating how an individual teacher construct a mock test may reveal more of the true testing practice than exploring how a committee prepares a formal test. Third, it is also a common practice that students often take several mock tests before sitting for the SAET. Therefore, the use of mock tests as research tools in the present study will not seem peculiar to students or teachers since they are quite familiar with mock tests. Given the above three reasons, I think the mock tests of SAET are legitimate research instruments in the study.

Research Questions

To achieve the above-mentioned research purposes, the present study addressed the following three research questions:

1. What considerations do teachers take into account when they construct mock tests for the Scholastic Ability English Test (SAET)? How do the tests constructed by novice teachers differ from those constructed by experienced teachers?
2. What strategies do students use to answer the SAET mock tests? How do the higher-proficiency students use strategies to answer the test items differently from the lower-proficiency students?
3. Are students' considerations for answering the SAET mock tests consistent with teachers' test-constructing considerations?

Delimitations

This study focused merely on the reading part of the mock test for the SAET

though the original formal test of the SAET consists of two sections—reading and writing. The reasons for narrowing down my research scope to reading are as follows. First, reading section of the SAET accounts for 72 points out of 100, while writing section accounts for only 28 points. To make my study more focused and compact, I decide to explore only the reading section, which is the major part of the SAET. Second, the reading section of the SAET is tested in the format of multiple-choice questions, one that is favored in many large-scale tests in Taiwan because of its reliable and rapid, economical scoring. As many scholars (e.g. Heaton, 1988; Weir, 1990; Cohen, 1994; Hughes, 2003) have indicated that good multiple-choice questions are notoriously difficult to construct, I think it is highly worthwhile to investigate how teachers construct the multiple-choice reading questions.

Significance of the Study

The significance of this study can be discussed from three perspectives: (1) design of language testing research, (2) teacher education curriculum, and (3) language testing research in Taiwan.

To begin with, the design of this study is pioneering in that it is the first study to explore EFL teachers' test-constructing process by using think aloud method and that it is also the first study to investigate the match and mismatch between EFL teachers' test-constructing considerations and EFL students' considerations for answering the tests constructed by the teachers. Therefore, the design of this study not only adds a new research direction to the language testing research as a whole, but also offers a new dimension of examining test validation in particular.

Secondly, the results of this study will shed new light on teacher education curriculum, in particular, on the testing course for EFL teachers. For one, the study examines the processes of how novice and experienced teachers construct a test; the findings may have implications for test construction guidelines. Teachers in training

might also benefit from this study as they would know in detail how a test is constructed from scratch. For another, the analyses of teachers' test-constructing processes and students' test-taking processes will help teachers know what students are thinking about while taking tests and whether the students are thinking in line with them. Thus, the results can familiarize teachers with EFL learners' test-taking behavior or strategies, giving them the background to derive pedagogical implications in their own teaching practice.

Lastly, this study is unique in the language testing research in the context of Taiwan. Up to date, there seems to have been no published study examining teachers' test-constructing considerations by using the think aloud method in Taiwan. It is hoped that the present study will generate more similar studies to help portray Taiwanese EFL learners' test-taking behavior and EFL teachers' test-constructing considerations. Most importantly, if the present study, which used the SAET mock test as the research tool, receives some critical acclaim in the testing field in Taiwan, it is hoped that a study using the real SAET as the research tool can be conducted by the CEEC committee some day. Such studies, I believe, will help improve the quality and validity of the SAET, one of the standardized tests in Taiwan.

CHAPTER TWO

LITERATURE REVIEW

In this chapter, research concerning the following themes will be reviewed. First, I will give an overview of language testing research. Second, I will review studies on how students take tests and on how teachers construct tests. Third, I will review research into the relationship of teachers' test-constructing processes and students' test-taking processes. Finally, I will review literature concerning the technique of verbal report in language testing.

Overview of Language Testing Research

Language testing research, a well-established branch of applied linguistics, has evolved and expanded through the years. Bachman (2000), in his state-of-the-art article, chronicled the major developments of testing research in the last two decades of the 20th century and also predicted the future directions for testing research in the 21th century. To gain a rough understanding of the whole testing research and to situate the present study in the testing field, I will briefly review Bachman (2000) in the following.

According to Oller (1979, cited in Bachman, 2000), language testing research, from the mid-1960s through the 1970s, was dominated by the hypothesis that language proficiency consisted of a single unitary trait, and the research methodology used was often a quantitative and statistical one. Then, the 1980s saw the influence of second language acquisition (SLA) research on testing research. Research in SLA spurred language testers to investigate not only a wide range of factors on language test performance (e.g., Douglas & Selinker, 1985; Chapelle, 1988; Hale, 1988), but also the strategies involved in the process of test-taking itself (e.g., Cohen, 1984). It was during this period that research on test-taking process began to emerge. Toward the end of 1980, language testers were challenged by Pienemann et al. (1988) to

explicitly take into consideration language learners' developmental sequence in the design of language tests and in the interpretation of test scores.

Testing research in the 1990s witnessed expansions in five major areas: (1) research methodology; (2) practical advances; (3) factors that affect performance on language tests; (4) performance assessment; and (5) ethical issues. Each of the five areas will be summarized briefly below.

Methodological approaches employed in language testing research in the 1990s have become increasingly sophisticated and diverse. Newer and more powerful quantitative methods, such as criterion-referenced measurement (Lynch & Davidson, 1997), generalizability theory (Bachman, 1997), item response theory (Pollitt, 1997), and structural equation modeling (Kunnan, 1998), have superseded classical norm-referenced reliability coefficients and exploratory factor analysis. Moreover, qualitative approaches have also been applied to language testing research. They include expert judgments, introspective and retrospective verbal reports, observations, questionnaire and interviews, text analysis, conversational analysis, and discourses analysis (Banerjee & Luoma, 1997).

Concerning practical issues, testing research agenda began to see advances in the areas of cross-cultural pragmatics (e.g., Hudson et al., 1992; 1995), languages for specific purposes (e.g., Douglas, 2000), computer-based assessment (Gruba & Corbel, 1997), and a renaissance in research into the testing of vocabulary (e.g., Read, 2000) and the development of new kinds of vocabulary tests (e.g., Nation, 1990; Laufer & Nation, 1999).

Regarding factors affecting performance on language tests, research has mainly focused on characteristics of the testing procedure (e.g., Fulcher, 1996; Riley & Lee, 1996), characteristics of test takers (e.g., Hill, 1993), and the test-taking process (e.g., Storey, 1997). A number of the test-taking process studies have used qualitative

methodologies mentioned above, such as verbal reports, questionnaires, and discourse analysis.

"Performance" assessment (McNamara, 1997), or "alternative" or "authentic" assessment (Herman et al., 1992; Wiggins, 1993) in the 1990s, has been spurred largely by widespread dissatisfaction with standardized multiple-choice tests in the communicative language teaching context, and by the developments in task-based language teaching and assessment (Norris et al., 1998). Performance assessment measures include classroom observation, portfolios, conferences, journals, questionnaires, interviews, self- and peer- assessment, group oral assessment, etc (Brown, 1998).

Ethical issues in the 1990s included research into washback on instruction (e.g., Alderson & Wall, 1993; Wall & Alderson, 1993), ethics of test use (e.g., Lynch, 1997; Shohamy, 1997), and professionalization of the testing field (e.g., Stansfield, 1993), which includes two interrelated activities: professional training and a code of practice (Davies, 1997).

After reviewing the major developments of testing research in the last two decades of the 20th century, Bachman (2000) also suggested some future directions for testing research in the 21th century. He believes that "there are two areas in which language testing and language testers must continue to grow and develop: the professionalization of the field, and validation research. However, rather than being two disparate directions,…these are two virtually related areas that lie on the same path" (Bachman, 2000, p. 18).

According to Bachman (2000), the professionalization of language testing has two major thrusts: (1) the training of language testing professionals; and (2) the development of standards of practice and mechanisms for their implementation and enforcement. Bachman (2000) further argues that "we will need not only to develop

standards of professional competence in language assessment, but also to become more active advocates for the inclusion of such standards in the standards for the training and certification of language teachers" (p. 20).

In regard to validation research, Bachman (2000) believes that the research in the past decade into factors and processes that affect language test performance and test scores will continue to blossom. In addition, "the debate over methodological issues has …moved from an overly simplistic view of the incompatibility of quantitative and qualitative approaches to a greater appreciation of their complementarity and of the necessity for including a range of approaches in our research agendas" (Bachman, 2000, p. 22).

In conclusion of his article, Bachman (2000) voices again that the professionalization of our field and validation research will continue to be vital to language testing. Bachman (2000) believes that:

> Language testing will grow as a profession in the twenty-first century to the extent that it effectively marshals the resources at its disposal to continue to vigorously investigate the validity of the inferences we make on the basis of test scores and the fairness of the uses we make of these scores. Validity and fairness are issues that are at the heart of how we define ourselves as professionals, not only as language tester, but also as applied linguists. (p. 25)

After reviewing Bachman's (2000) overview article of the testing research, I think my present study well-fitted into the future directions mentioned in Bachman (2000). On one hand, my research focus on teachers' test-constructing processes is in line with research into "the professionalization of the field." On the other hand, my research concern of students' test-taking processes helps contribute to the "validation research." It is against this backdrop that the present study unfolds.

Studies on Students' Test-taking Process

In this section, we will review several verbal report studies on students' test-taking process in reading tests, the focus of the current study.

Early Attempts

Cohen (1984) is one of the early efforts examining test-taking process by using verbal report data. The main purpose of Cohen (1984), which described the results of five student course papers, was to discuss methods for obtaining verbal report data on L2 test-taking strategies and to report on some types of the findings obtained. The verbal report methods explored in Cohen (1984) included think-aloud and self-observation (i.e., introspection, immediate retrospection, and delayed retrospection). The data obtained EFL and ESL students' test-taking strategies on cloze tests and multiple-choice reading comprehension tests. The results, in general, showed that not all of the students read the entire cloze passage or reading passage before answering the test items although they were requested to do so. In terms of cloze tests, it was found that some students did not use the context to find clues for filling in the blank, and that students would use the strategy of translation in doing cloze tests. Moreover, when not knowing how to fill in a blank, poor students would leave it blank, and better students would make guesses. In terms of multiple-choice tests, students reported either reading the questions first or just part of the article and then looking for the corresponding questions. Moreover, students would use a strategy of matching material from the passage with material in the item stem and in the alternatives. In sum, Cohen (1984) has demonstrated some ways how verbal report data can be obtained, and the paper concluded that "there is value in striving for a closer fit between how test constructors intend for their tests to be taken and how respondents actually take them" (Cohen, 1984, p. 79).

Following Cohen (1984), we will first, in the following, review other verbal report studies on multiple-choice reading comprehension tests (e.g., Nevo, 1989; Anderson et al., 1991; Rupp, Ferne & Choi, 2006), and then review those on cloze tests (e.g., Storey, 1997; Sasaki, 2000; Yamashita, 2003; Moghaddam, 2010).

Studies on Multiple-choice Reading Comprehension Tests

Nevo (1989) examined students' test-taking strategies on a multiple-choice reading comprehension test by adopting the methods of immediate introspective verbal report and retrospective report. Forty-two Hebrew students studying French participated in the study, and they were asked to complete a multiple-choice test on four reading passages (two in Hebrew and the other two in French). An innovation of Nevo (1989) is that a checklist of fifteen strategies was provided for students to facilitate their reporting of strategy use after completing each item of the test. The results showed that there was a transfer of strategies from L1 (Hebrew) to L2 (French), and that the most frequently used strategies in both languages were *returning to the passage* and *clues in the text*. It was also found that in L2, students used more strategies which did not lead to the correct answer than in their L1. Finally, the major contribution of Nevo (1989) to the verbal report method is that by providing a checklist, it is possible to obtain feedback from students about their strategy use on an item-by-item basis.

Anderson et al. (1991) presented the results of an exploratory study that examined three types of information (test-taking strategies, item content, and item performance) in the investigation into the construct validity of a reading comprehension test. The participants were twenty-eight Spanish-speaking students, and they were asked to produce retrospective think-aloud protocols while taking an English reading comprehension test, which contained forty-five multiple-choice questions. The results were as follows. First, there was a statistically significant association between students' reported strategies and the three question types determined by the test developers. Second, students' strategy use was significantly related to item difficulty and to item discrimination. More specifically, five strategies were worthy of note in the study. First, the strategy *stating failure to understand*

occurred more frequently on inference test items, was used fewer times on easy items, and was used more times on items that discriminated well among those students who scored high on the test. Second, *paraphrasing* occurred more frequently on items asking students to identify the direct statement of the passage, and was used more times on items classified as acceptable in terms of discrimination. Third, *guessing* was reported more times on inference items, fewer times on easy items, and occurred about as often on acceptable and rejected items in terms of item discrimination. Fourth, *matching the stem with a previous portion of the text* was reported fewer times on items directed at identifying the main idea, reported fewer times on easy items, and reported fewer times on acceptable items. Fifth, *making references to time allocations* was reported fewer times on inference questions and more times on acceptable items. This study has thus showed us the value of test-taking protocols, along with other data sources, in the investigation of construct validity of a reading comprehension test.

Rupp, Ferne, and Choi (2006) examined test-takers' use of strategies on a multiple-choice reading comprehension test, with a purpose of investigating the equivalence of reading processes and strategy use in testing and non-testing reading conditions. The participants were ten ESL adult learners, who were first asked to verbally report their test-taking process in a semi-structured interview and then were asked to do concurrent think-aloud. The results showed that reading processes in a test condition were strikingly different from those in a non-testing context. Moreover, the construct of reading comprehension was shown to be assessment specific and was fundamentally determined through item design and text selection. In terms of learner strategies, the study presented three findings. First, learners viewed responding to multiple-choice questions as a problem-solving task rather than a comprehension task. Second, learners selected a variety of unconditional and conditional response strategies to deliberately select choices. Third, learners combined a variety of mental

resources interactively when determining an appropriate choice. In sum, the authors concluded that their findings support the development of response process models that are specific to different item types, the design of further experimental studies of test method effects on response processes, and the development of questionnaires that profile response processes and strategies specific to different item types.

Studies on Cloze Tests

Storey (1997), employing the methods of concurrent think aloud and immediate retrospection, investigated twenty-five Hong Kong EFL students' test-taking process in a 13-item, multiple-choice, discourse cloze test. The purpose of the study was to provide introspective validation of the testing technique and the test items by assessing observed test-taking behavior against a predicted model of ideal performance. The results revealed that different items entailed varying degrees of construct validity. Some students were found to have used theoretically expected reading processes, while others merely considered information at the within-sentence level. Although there was a mismatch between the theoretically assumed processes and the actual processes applied by some test-takers (such as use of the strategies of *elimination* and *surface matching*), the items were capable of generating construct-relevant processing, and the test was judged to have a good degree of construct validity.

Sasaki (2000) investigated how content schemata activated by culturally familiar words might have influenced students' test-taking processes in a cloze test. Sixty Japanese EFL students were divided into two groups, each completing either a culturally familiar or an unfamiliar version of a cloze test. The participants were asked to produce immediate retrospective protocols while taking the test, and then to recall the passage after they had completed the whole test. The results showed that those who read the culturally familiar cloze text tried to solve more items and generally

understood the text better, which resulted in better performances than those of the students who read the unfamiliar text. The paper concluded it has demonstrated the merits of using multiple data sources for investigating students' test-taking processes, and that the results also support the claim that cloze tests can measure higher-order processing abilities.

Replicating Sasaki's (2000) experiment, Moghaddam (2010) examined the effects of cultural schemata on Iranian students' test-taking processes in a cloze test. The participants were 116 Iranian university students, who were divided into two groups, each completing either a culturally familiar or a culturally unfamiliar version of a cloze test. They were asked to develop retrospective protocols of their test-taking process and recalls of the cloze passage. Similar to the findings of Sasaki (2000), the results of Moghaddam (2010) showed that students who read the culturally familiar cloze text generally understood the text to a greater extent and resulted in a high score in comparison with those who read the unfamiliar text. Both Sasaki (2000) and Moghaddam (2010) suggested that cultural schemata has certain effect on students' test-taking processes in cloze tests.

Yamashita (2003) compared skilled and less skilled readers in their processes of taking a gap-filling cloze test. Twelve Japanese EFL students (six skilled and six less skilled) were required to complete a 16-item gap-filling test while thinking aloud about their test-taking processes; afterward, they were interviewed informally by the researcher. The results demonstrated that both skilled and less skilled students used text-level information more frequently than other types of information (such as clause-level, sentence-level, and extra-textual information). However, the skilled readers used text-level information more frequently than the less skilled readers. In sum, the gap-filling test generated processes that made readers utilize text-level constraints, and overall differentiated well between skilled and less skilled readers.

We have reviewed several studies on students' test-taking processes or strategies so far. Although those studies were conducted for different purposes (e.g., validation, comparison of L1 and L2, comparison of skilled and less skilled readers, or the effects of cultural schemata), they all employed the method of verbal report in their experiment. It can be seen clearly that verbal report has been widely utilized as a means of collecting qualitative data. As Sasaki (2000) comments well, "The product- and process-oriented data complemented each other, providing insights that could not have been gained in the absence of one or the other" (p. 107). In the present study, I will also use verbal report to examine Taiwanese students' test-taking process in a reading test.

Studies on Teachers' Test Construction

"Classroom teachers are in the front line of introducing students to formal learning, including assessment" (Leighton, et al., 2010, p. 7). That is, the first test students take in class is usually made by their teachers, and it is also their classroom teachers that prepare them for the formal, large-scale tests. Therefore, teachers' "assessment literacy" (Stiggins, 1991) is very important. According to Stiggins (1991), "teacher *assessment literacy* [emphasis in the original] is characterized by understanding what it takes to produce high-quality achievement data for both classroom and large-scale tests, scrutinizing achievement data and not accepting it at face value, and being sufficiently confident to ask questions about technical information and complicated summaries of test scores" (Leighton, et al., 2010, p. 9). Although assessment literacy is important, it is a pity that many teachers do not seem to be equipped with a solid grounding in the basic knowledge of assessment principles or practices (Leighton, et al., 2010). Many educators have also noted that, for teachers, producing good tests is a demanding task (Davidson & Lynch, 2002). The general inadequacies of teachers' knowledge of test-constructing skills can be shown in the

studies reviewed in the following.

Training in Teachers' Test Construction

To begin with, Carter (1984), almost three decades ago, investigated teachers' competence in test item development by asking them to identify and write specific items aimed to measure particular reading skills (main idea, detail, inference, and prediction), and by interviewing them about their test-constructing perceptions and processes. The results showed that teachers had more difficulty in identifying and developing items tapping higher-level reading skills (i.e., inference and prediction) than in identifying and writing items to test lower-level cognitive skills (i.e., main idea and detail). The interview data suggested that teachers felt insecure about their knowledge of basic principles for item writing and that they might possess a limited repertoire of test-constructing skills. Based on these results, Carter (1984) argued for an emphasis on the testing course in preservice and inservice teacher education.

To equip teachers with test-construction principles and to improve their test-writing skills, many teacher education programs began to include language testing courses. Kathleen M. Bailey and James D. Brown have reported the results of two questionnaire surveys (Bailey & Brown, 1996; Brown & Bailey, 2008) of instructors of language testing courses worldwide. They found that the contents of the language testing courses were quite diversified, covering topics such as hands-on experiences, general topics, item analysis, descriptive statistics, test consistency, and test validity. In spite of these diverse topics which could enhance teachers' assessment literacy, some preservice or inservice teachers did not take the testing course either as an elective or as a requirement (Brown & Bailey, 2008). Consequently, it is possible that some classroom teachers might be ill-prepared to face the challenges of classroom assessment because they do not have the opportunity to learn to do so.

Among the various topics a testing course might offer, "test-writing," one of the

hands-on experiences in Brown and Bailey's study (2008), is of vital importance to classroom teachers. Although the existing programs or textbooks on how to write tests are many, research on how teachers go through the test-constructing process is quite few. Studies relevant to the issue in question are Kirschner, Spector-Cohen, and Wexler (1996), Johnson, Becker, and Oliver (1999), and Coniam (2009), which will be reviewed respectively.

Kirschner, Spector-Cohen, and Wexler (1996) reported on a college teacher education workshop on the construction of EFL tests and materials. The workshop was held due to the discovery by the EAP course coordinators that the reading tests produced by their staff had two major problems. First, many of the teacher-produced tests did not reflect the main hierarchy of ideas of the text; and second, many of the test questions were confusing to students due to poor wording. To help improve the teachers' test-writing skills, a four-session workshop was launched. The workshop consisted of a presentation of theoretically motivated criteria for writing test items, a description of the stages in the test writing procedure, and hands-on experience in the construction of a reading comprehension test. In two sessions of the workshop, the participating teachers, in groups, were asked to (1) produce a 10-question reading comprehension test; (2) have another group comment on their test, and (3) revise their original test based on the comments and the criteria checklist. Although Kirschner, Spector-Cohen, and Wexler (1996) did not describe in their paper the actual test-constructing processes the teachers went through or the teachers' attitude toward the workshop, their study did show us what a test-writing workshop would be like and the need for holding such a workshop for inservice teachers.

Johnson, Becker, and Oliver (1999) reported on a second language testing course in an M.A. TESOL program in the US. In the course, seven student teachers not only learned about some theoretical issues related to language testing, but were also asked

to write, administer, and analyze a reading test of their own making. Again, Johnson, Becker, and Oliver (1999) did not describe the test-constructing processes the student teachers underwent, but they presented the student teachers' responses (through interviews with the instructor) to the test-writing exercise. In general, the student teachers all agreed that the experience of the test-writing task was valuable in that it gave them a meaningful integration of theory and practice. They were also surprised to have found that they were unable to predict which items would be good discriminators; therefore, they found the test analysis to be both interesting and informative. Johnson, Becker, and Oliver (1999) even conducted a follow-up survey two years later, asking the class members whether they had retained the information about testing and whether they had used their knowledge of testing in their teaching. One member responded that he did not remember a lot of the information they had learned, but he did remember the exercise of constructing and evaluating a test. Another member commented he had learned that it was difficult to make a "perfect" test. The study of Johnson, Becker, and Oliver (1999) suggests that the hands-on experience of writing, administering, and analyzing a test is meaningful to student teachers, and they argue for more weight on the test-construction training for student teachers in language testing pedagogy.

Coniam (2009) reported on a study investigating the quality of teacher-produced tests for EFL students and the effects of a training program in test development. The participating teachers, enrolled in a part-time M. A. in ELT program in Hong Kong, were asked to produce reading comprehension tests in real life for their own students, proceeding through the stages of test specification, moderation, item analysis, and test refinement. Subsequently, the participants were encouraged to complete an open-ended questionnaire, which asked them to reflect on their test development process and to examine their test data for quality in terms of classical test statistics.

The results showed that the majority of items produced by the participants could not be deemed “good” from a classical test measurement perspective. In other words, the quality of those teacher-produced tests could not be considered of high quality, which was beyond the participants’ expectations. In terms of the effects of the testing program, comments from the questionnaires revealed that the program was quite effective in raising participants’ awareness of test principles and insights into the processes of test development and analysis. One participant’s comment was worth mentioning that “frankly speaking, I seldom consciously consider qualities of a ‘good’ test. However, I now realize that constructing or evaluating a good test is a complex task involving both art and science” (Coniam, 2009, p. 232). The study concluded that to improve the quality of teacher-produced tests, all areas of training in test development and production requires attention.

So far, we have reviewed three studies (Kirschner, Spector-Cohen & Wexler, 1996; Johnson, Becker & Oliver, 1999; Coniam, 2009) concerning training in test-construction. These studies have shown us the different stages teachers went through in test development (such as test writing, moderation, analysis and refinement), the examination of the quality of teacher-produced tests, and the participants’ responses (through interviews or questionnaires) to the test-writing training programs. Nevertheless, we have not been informed of any firsthand account of the considerations that teachers had taken into while constructing their test items. In other words, we have not learned about teachers’ verbal reports of their own test-constructing processes. As a matter of fact, no study, to my knowledge, has directly examined this issue through verbal report analysis. Thus, the present study will be a good complement to literature on training of teachers’ test-writing skills.

Studies on Test Constructor Effect

The notion of investigating teachers’ test-constructing process shifts our attention

from test-writing training to test-constructors. The investigation of whether there is a constructor effect in test variance is worth exploring, especially in contexts where test-construction is usually teachers' solitary work, such as Hong Kong (Coniam, 2009) and Taiwan. Jafarpur (2003) has provided a good starting point for researching test constructor effect.

The purpose of Jafarpur (2003) was to explore the relative impact of the test constructor on the performance of test-takers using multiple-choice reading comprehension tests that had no specifications. Six EFL test constructors and 335 Iranian university students were involved in the study. Each of the six test constructors was asked to independently write eleven multiple-choice comprehension items on a pair of passages assigned to them at random. The test constructors were not informed of the sources of the passages or of the intentions of the study. Neither were they provided with any specification or assistance on how to assess reading skill. They just formulated the stem and the choices on the basis of their own renditions of the passage. Afterward, the original items on the passages and the items produced by the six constructors were trialed on 335 Iranian EFL learners. The results showed that there was unintended variation in the test-taker's scores when they took different sets of items designed by different item writers. Therefore, the results suggested that there was probably a facet associated with the test constructor that explained some of the variation in the test-takers' performance.

Since Jafarpur (2003) suggested a test constructor effect on the test-takers' performance of tests that had no specifications, I assume there might also exist a test constructor effect on test variance using tests with specifications. This assumption is based on the research findings of rater effect on performance tests, for example, Brown (1995) on a speaking test, and Lim (2011) on a writing tests.

Brown (1995) explored the influence of rater backgrounds (native/nonnative;

with/without industry and teaching experience) on assessments in an oral test of Japanese for tour guides. Assessments of fifty-one test candidates made by thirty-three raters were compared in order to determine what effect raters' background has on assessments that were made on linguistic and real-world criteria. The results showed that although there were significant differences in ratings awarded for some individual criteria, there were no significant differences between types of rater in terms of the overall grade awarded.

Lim (2011) examined new and experienced raters' performance longitudinally over multiple time points in a writing test. The study used operational data from the writing section of the MELAB, an international exam of English proficiency, to investigate the rating quality of six novice raters and five experienced raters over three time periods of twelve to twenty-one months. Rating quality was operationalized in terms of rater severity and consistency. The results were as follows. First, novice raters, where initially differing in performance, learned to rate appropriately relatively quickly. Second, raters were able to maintain rating quality over time. Third, rating volume and rating quality might be related to each other.

Brown (1995) and Lim (2011) have clearly shown that there exists rater effect even with rating criteria provided. Therefore, by analogy, it is reasonable to suggest that there might also be a test constructor effect on students' test performance even with test specifications provided. Motivated by Lim (2011), I want to examine possible test constructor effect, especially constructor's length of teaching years, in test variance. In other words, I want to explore how the tests produced by novice teachers differ from those produced by experienced teachers.

Research into the Relationship Between Test-constructing and Test-taking Processes

Literature in language testing has suggested that there might be a mismatch

between test-developers' assumptions about what they test and the actual processes test-takers have gone through (Nevo, 1989). There have been studies (MacKay, 1974; Haney & Scott, 1987; both cited in Cohen, 1994) showing that students may get an item wrong for the right reasons or right for the wrong reasons. Therefore, research aiming to investigate "the closeness-of-fit" between test-constructors' assumptions and students' actual test-taking processes is vital to the validity of test results. Nevertheless, such research is quite rare. To the best of my knowledge, Gierl (2001), which examined a mathematics test, has explored this issue to a certain extent.

Gierl (2001) examined whether Bloom's *Taxonomy of Educational Objectives: Cognitive Domain* provided item writers with an accurate model for anticipating the cognitive processes used by elementary school students on a large-scale achievement test in mathematics. Thirty Grade 7 students, divided into high math achievers and low math achievers, were asked to think aloud as they solved problems on the math test, which consisted of eighteen multiple-choice items. Major findings of this study were as follows. First, Bloom's taxonomy does not provide an accurate model for guiding items writers to anticipate the cognitive processes used by students, since the overall match between the responses expected by the item writers and the responses observed from the students was only 53.7%. Second, the match score between the expected and the observed responses differed for the high and low math achievers and also differed across the two content areas (numeration and operations/properties) measured on the test. Item writers were able to anticipate the processes used by high achievers more readily than by low achievers.

Gierl (2001), though examining a math test, has offered us an example of how to do the "closeness-of-fit" research into the relationship between test-constructors' assumptions and test-takers' performances. He compared the test-developers' categorization of item types by using Bloom's taxonomy with students' think-aloud

protocols. He did not collect the item writers' think-aloud protocols probably because the test was a large-scale achievement test, which might be written by a committee instead of individuals so that the test writers' think aloud protocols were unavailable. Nevertheless, seeking a fair judgment, I think research into a match between test-constructors' considerations and test-takers' considerations should better be done in a way in which think-aloud protocols of the two parties are compared and contrasted. It is with this goal in mind that I design the present study in this fashion: collecting the test-constructors' think-aloud protocols and the test-takers' think-aloud protocols, and then examining the correspondence between the two sets of verbal report data.

Verbal Report in Language Testing

The main methodology used in the present study will be verbal report. In this section, I will review relevant literature to verbal report, explaining what verbal report is and describing how to do verbal report.

Ever since Ericsson and Simon (1984) published their classic book *Protocol Analysis: Verbal Reports as Data*, the 1980s and early 90s have seen a keen interest in applying the method of verbal protocol analysis in many areas of research, such as psychology, education, and cognitive science (Ericsson & Simon, 1993). In doing verbal reporting, a person has to verbalize his or her thought processes while completing a given task, and the validity of verbal protocol analysis is based on the notion that "individuals had privileged access to their experiences" (Ericsson & Simon, 1993, xii), and that the information in their verbal reports is trustworthy (Park, 2009). In other words, "verbal protocol analysis is a methodology which is based on the assertion that an individual's verbalizations may be seen to be an accurate record of information that is (or has been) attended to as a particular task is (or has been) carried out" (Green, 1998, pp. 1-2).

Generally speaking, in doing verbal reporting, subjects are required to perform at least two tasks: *processing* the designated task and *reporting* the processes. Given the fact that verbal protocols may be gathered under varying circumstances where the possibility of the temporal separation between task processing and reporting exists, several researchers (e.g., Cohen, 1987, 2006; Ericsson & Simon, 1993; Green, 1998) have categorized verbal reports in different ways. For example, Cohen (1987, 2006), in describing the ways of gathering language learners' test-taking strategies, classified verbal reports into three categories: (1) *Self-report*: learners' descriptions of what they do, characterized by generalized statements (e.g., questionnaires or interviews on general test-taking behaviors); (2) *Self-observation*: the inspection of specific, contextualized language behavior, either introspectively (e.g., stimulated recall or immediate retrospection) or retrospectively (e.g., questionnaires or interviews on a specific test-taking instance); (3) *Self-revelation*: stream-of-consciousness disclosure of thought processes while the information is being attended to (e.g., concurrent think aloud).

Green (1998) categorized verbal reports in another way, focusing on three variations of the verbal report procedure: Forms of report, temporal variations, and procedural variations. Green's (1998) categorization can be summarized in Table 1.

Table 1.
Three Variations of the Verbal Report Procedure

Forms of Report	Temporal Variations	Procedural Variations
Talk aloud	Concurrent	Mediated Non-Mediated
	Retrospective	Mediated Non-Mediated
Think aloud	Concurrent	Mediated Non-Mediated
	Retrospective	Mediated Non-Mediated

As Table 1 shows, there are two forms of verbal report. According to Green (1998), talk aloud report includes information that is already encoded in verbal form, and the information roughly corresponds to words in the mind, or thoughts that might be spoken. On the other hand, think aloud report includes information already encoded in verbal form, plus information that may not originally have been encoded in verbal form. This non-verbal information must be transformed and then verbalized. Therefore, think aloud report includes more information than talk aloud report does. In terms of temporal variations, each form of report can be divided into concurrent (simultaneous) report and retrospective (subsequent) report. Furthermore, in view of procedural variations, verbal reports can be gathered in either mediated or non-mediated way. Taken together, Green's (1998) categorization yields eight different verbal report methods, and researchers can adopt one that is suitable for their research purposes.

In categorizing verbal reports, Ericsson and Simon (1993) drew a distinction between three levels of verbalization: Level 1 verbalization is simply the vocalization of covert oral encodings, which involve no intermediate processes and no additional oral encodings. Level 2 verbalization concerns descriptions or explications of the thought content. Level 3 verbalization requires the subject to explain his thought processes or thoughts. In interpreting Ericsson and Simon's (1993) levels of verbalization, Park (2009) refers to Level 1 verbalization as *talk aloud*, Level 2 as *think aloud*, and Level 3 as *retrospection*. Park's interpretation, in my opinion, makes Ericsson and Simon's (1993) categorization of verbal reports more comparable to Cohen's (1987, 2006) and Green's (1998).

Verbal protocol analysis has now become a popular research tool in the fields of language learning and testing. Numerous books and articles (e.g., Cohen, 1987; Faerch & Kasper, 1987; Green, 1998; Gass & Mackey, 2000) have been published to

introduce verbal protocol analysis in language research, and many studies (e.g., Afflerbach & Johnston, 1984; Block, 1986; Pritchard, 1990) have employed verbal reports as a method of gathering the mental processes that language learners use. In language testing, the technique of verbal reporting is generally used in two areas: (1) test validation research, and (2) studies on test-taking processes or strategies.

Validity is one of the crucial terms in language testing, and it investigates whether a test measures the ability or skill that it claims to measure. In other words, validation research seeks to explore "the closeness-of-fit between the tester's *presumptions* about what is being tested and the *actual* processes that the test taker goes through" (Cohen, 1984, p. 70). If there is a close match between the two, then the test may be said to measure what it purports to measure; that is, the test has good validity. In test validation, qualitative validation techniques, such as verbal reports, can provide information on the content of the test, the properties of the test tasks, and the processes involved in test taking and assessing (Banerjee & Luoma, 1997). Therefore, verbal reports have increasingly played a vital role in the validation of assessment instruments. For example, Anderson et al. (1991) explored the use of think aloud protocols and the other quantitative data to examine the construct validity of a reading comprehension test. Moreover, Norris (1991) used verbal protocol analysis as part of the process for validating the use of multiple choice questions. As quantitative techniques can provide information only on the psychometric properties of the test, verbal reports can "provide evidence other than the test developer's statements about what is being assessed in the test, and how" (Banerjee & Luoma, 1997, p. 276). That is exactly the major contribution verbal reports make to the validation research. In addition, as "students may get an item wrong for the right reasons or right for the wrong reasons" (Cohen, 1998, p. 91), test developers would never know the truth behind students' test scores and the true validity of the test without students' verbal

protocols.

Verbal reports are also commonly used in studies examining students' test-taking processes or strategies. In describing test-taking strategies, Cohen (2006) even regards verbal report as an important tool since it has become a primary research tool for this kind of endeavor. The technique of verbal reporting has been extensively applied to almost every aspect of language testing—reading (e.g., Nevo, 1989; Storey, 1997; Sasaki, 2000), listening (e.g., Buck, 1991; Ross, 1997), speaking (e.g., Cohen & Olshtain, 1993; Swain, 2001), and writing (Lay, 1982; Raimes, 1985; Arndt, 1987). The abundant verbal report studies have helped us gain a better understanding of the test-taking processes and the strategies students employ while doing the designated tasks. Verbal protocol analysis is even applied to examining rater behavior in speaking tests (e.g., Orr, 2002) and in writing tests (e.g., Weigle, 1994).

Although I divide the verbal report studies into the above two areas based on the researchers' major purposes, the test validation research and the test-taking strategy research are in fact two areas that complement each other. Cohen (2006), in summarizing the 25 years of research on test-taking strategies, states that "research on test-taking strategies can serve as a valuable tool for validating and refining notions about the test-taking process (p. 325)" and that "empirical research on test-taking strategies can provide valuable information about what tests are actually measuring" (p. 325). In sum, studies employing verbal protocol analysis can serve dual purposes at the same time—revealing students' test-taking process or strategies on the one hand, and verifying the validity of the test on the other hand. Although verbal protocol analysis is not without its criticism (e.g., Nisbett & Wilson, 1977; Storey, 1997), it has lent itself to becoming a vital technique in the qualitative research.

CHAPTER THREE

METHODOLOGY

This section depicts the participants, instruments, and procedures for collecting and analyzing data in the present study.

Participants

The participants in the study included four Taiwanese senior high school English teachers, who constructed the mock tests for the Scholastic Ability English Test (SAET), and forty-eight Taiwanese senior high school students, who took the tests the teachers constructed.

Four incumbent senior high school English teachers were invited to participate in the study. Two of them are novice teachers, and the other two are experienced teachers. For a more striking difference, the "novice" teachers in this study referred to those teaching English no more than three years, and the "experienced" teachers referred to those who have been teaching English for at least ten years. All of the four recruited teachers met this requirement during their participation in the study. The four teachers were asked to sign a research consent form (see Appendix A), fill out a background questionnaire (see Appendix B) before participating in the study, and answer a feedback sheet (see Appendix C) after finishing their task in the study.

As for recruitment of the forty-eight student participants, a convenient sampling was adopted due to administrative constraints. That is, they were all recruited from a senior high school in northern Taiwan where I was teaching at the time of data collection. The procedures for recruiting student participants are as follows. First, 144 students[3] were asked to take a shortened version of intermediate level of Foreign

[3] Among the 144 students, 124 belonged to three intact classes, which were taught by me. Thus, these students took the FLPT in their regular English classes. The remaining twenty students, who were interested in my study but belonged to two other classes taught by my colleagues, took the FLPT during the school lunch break under my supervision.

Language Proficiency Test (FLPT, see Appendix D). Second, based on the 144 students' scores on the shortened version of FLPT, twenty-four among the top twenty-six students and twenty-four among the bottom thirty-two students (i.e., the potential targeted research participants) were invited to participate in the present study[4]. The top twenty-four students served as the higher-proficiency group and the bottom twenty-four students as the lower-proficiency group. Based on the forty-eight students' scores on the three English achievement exams (i.e., two midterms, and one final) held in the previous semester, students in the higher-proficiency group did have a better English proficiency than those in the lower-proficiency group. The students' scores are tabulated in Table 2.

Each of the two proficiency groups was further divided into four subgroups, with six students in each subgroup, who then took different SAET mock tests (Forms A, B, C, and D). The division principle I adopted was the S-shape grouping. First, I ranked the students in both proficiency groups by their FLPT scores, from high to low. Then, the S-shape grouping principle was adopted to put students in each group into four subgroups (see Table 2). In this way, each of the forty-eight students was assigned a code name. For instance, the student with the code H01A means that the student belonged to the higher-proficiency group, and took Form A of the SAET mock test.

The forty-eight student participants, males and females included, were all second-year senior high school students. Their average age was seventeen. Although it is usually the third-year students that take the SAET, some second-year students also take the mock test of the SAET in my school and in other schools as well. Therefore, I regard the second-year students as legitimate participants for the present

[4] Ideally, I would like to invite the top twenty-four and the bottom twenty-four students as my participants, but some of them, especially those ranking at the bottom were reluctant to participate. Therefore, only those who met my criteria and were willing to participate were recruited as the student participants in this study.

study. Each of the forty-eight student participants were asked to sign a research consent form (see Appendix E) before participating in the study.

Table 2.

Participants' FLPT Scores and Exams Averages

Higher-proficiency			Lower-proficiency		
No.	FLPT score	Exams average	No.	FLPT score	Exams average
H01A	32	91.7	L01A	13	75.0
H02B	31	98.3	L02B	13	41.3
H03C	30	86.0	L03C	13	65.3
H04D	30	89.0	L04D	13	53.0
H05D	29	93.9	L05D	13	28.7
H06C	28	90.3	L06C	13	31.0
H07B	28	91.3	L07B	13	42.7
H08A	27	94.3	L08A	13	61.7
H09A	27	89.7	L09A	12	36.3
H10B	27	86.7	L10B	12	49.0
H11C	27	68.3	L11C	12	47.7
H12D	27	87.0	L12D	12	43.7
H13D	27	82.3	L13D	12	47.3
H14C	26	93.3	L14C	12	40.7
H15B	25	88.0	L15B	11	41.7
H16A	25	65.3	L16A	10	30.7
H17A	24	69.0	L17A	10	38.4
H18B	24	70.7	L18B	10	30.7
H19C	24	63.0	L19C	10	22.0
H20D	23	67.7	L20D	10	23.3
H21D	23	77.0	L21D	9	60.3
H22C	23	72.0	L22C	9	28.7
H23B	22	91.0	L23B	9	39.0
H24A	21	85.3	L24A	8	64.7
(average)	26.25	82.96		11.33	43.45

Note. The full score of FLPT is 40. The full score of Exams average is 100.

Prior to the study, the participants had received at least seven years of English instruction in school, with three years in elementary school, three years in junior high school, and one and a half years in senior high school. The English instruction they received in elementary school mainly focused on communicative skills, such as

listening and speaking, while the English instruction in high schools mostly centered upon vocabulary, grammar, and reading. Although the participants had received much formal English instruction, their English proficiency, in general, is not very high as their scores on the FLPT suggest (see Table 2).

Instruments

The instruments used in the study included a background questionnaire, a feedback sheet, a shortened version of intermediate level of Foreign Language Proficiency Test (FLPT), two sets of materials for test construction, and four mock tests for the Scholastic Ability English Test (SAET).

Background Questionnaire

The background questionnaire (see Appendix B) was intended to gather basic information about the four teachers' teaching and test-constructing experiences. It is written in Chinese with twenty-one items. The first seven items ask the participants to report their personal background, such as their educational background and years of teaching experiences. The next three items ask the teachers to reflect on their familiarity with the SAET. Then, the following nine items probe the teachers' test-construction experiences, asking about their test-construction frequencies in school and about their skills in test construction. The last two open-ended questions ask the participants to describe their attitude toward test construction and the problems they had encountered in test construction.

Feedback Sheet

The feedback sheet (see Appendix C) is intended to gather the four teachers' opinions about their task in the study. It is written in Chinese with seven items. The first five questions ask the teachers to express their opinions about the materials based on which they constructed tests items, and about the test types (i.e., vocabulary, cloze, and reading comprehension) they were requested to construct. The last two

open-ended questions ask the teachers to describe the impact of adopting think aloud method in the test-constructing process on them, and their opinions about participation in the present study.

Foreign Language Proficiency Test

A shortened version of intermediate level of Foreign Language Proficiency Test (FLPT, see Appendix D) was used to measure the potential student participants' English proficiency. The test, taken from Tsai (2008), consists of forty items of multiple-choice questions, with ten items on vocabulary and idioms, twenty on grammatical usage, and ten on reading comprehension.

The reasons for adopting the shortened version of FLPT as a proficiency measure in the present study are as follows. First, the test was modified by Tsai (2008) from a standardized test developed by the Language Training and Testing Center (LTTC) in Taiwan. The modified test, according to Tsai (2008) in her pilot study, was proved to have good internal reliability (α = .79) and good convergent validity (r = .57, p< .001). Therefore, the modified test is considered a reliable measuring instrument. Second, Tsai (2008) used this shortened version of FLPT to measure Taiwanese senior high school students' English proficiency levels. Since the participants in the present study were also senior high school students, it is considered appropriate to use the same test to measure the participants' English proficiency in the present study.

Two Sets of Materials for Test Construction

Two sets of materials (Material A and Material B, see Appendix F) were provided by the researcher for the teachers to construct mock Scholastic Ability English Test (SAET). The teachers were asked to construct twenty-eight multiple-choice items based on the materials provided. Specifically, they were required to construct fourteen items (five vocabulary, five cloze, and four reading comprehension items) based on the contents of Material A, and another fourteen items

on Material B.

There are two reasons for providing designated materials for teachers to construct tests. First, if not given the same materials, the four teachers would probably construct their tests based on different materials, which would add another variable to the present study. Since teachers' considerations for selecting texts for test construction are not the major concern in the study, it would be a reasonable practice to provide teachers with designated materials. In fact, teachers' text selection considerations deserve another study. Second, concerning the research questions in the present study, only by asking teachers to construct tests based on the same materials can we compare the test items produced by different teachers and their test construction considerations on a more equal footing. Due to the above two reasons, I, the researcher, took the responsibility of selecting the materials for test construction in the present study.

Each set of the two materials includes a vocabulary list of forty words, a short passage for cloze test, and a longer passage for reading comprehension. The two sets of materials were carefully selected from the past SAETs, and were designed to be comparable in difficulty levels. My criteria for selecting the words and passages are as follows. As for the words in the two lists, I first collected all the words occurring in the vocabulary section of SAET from the year 2010 to 2012. Then, I deleted words that are beyond Level 4[5] or are not included in the vocabulary list compiled by the CEEC. Finally, I carefully chose eighty words among the remaining ones, divided them equally into two lists, and made them as comparable as possible in terms of word length, part of speech, and word level.

[5] The College Entrance Examination Center (CEEC) has compiled an English reference word list for the two major standardized English tests in Taiwan, namely, the SAET and the DRET. The list contains 6,480 words in American English, and these words are classified into six levels, with Levels 1-4 suitable for the SAET, and Levels 1-6 for the DRET.

Table 3.

Comparison of Material A and Material B

	Material A	Material B
Vocabulary list:		
Word length (average letters)	7.975	7.650
Part of speech (number of words)		
Verb	12	12
Noun	12	12
Adjective	8	8
Adverb	8	8
Word level (number of words)		
Level 1	1	1
Level 2	7	8
Level 3	15	15
Level 4	17	16
Cloze test passage:		
Topic	Listening	Being on time
Length (words)	153	166
Difficulty index	9.5	8.7
Source	SAET in 1999	JCEE in 2000
Reading comprehension passage:		
Topic	An old leather shoe	Dress code
Length (words)	304	301
Difficulty index	7.8	8.8
Source	SAET in 2001	SAET in 2002

As for the passages for cloze test and reading comprehension, they were also selected from the past JCEE or SAET from the year 1999 to 2002. The main reason I selected the used passages as the materials in my study is that the used passages were already piloted and tested, and thus are considered reliable. Another benefit of using the old passages is that I could compare the items which the participants constructed with those appearing in the past SAETs. However, there is one disadvantage in using the old passages: the participants might have seen them before, and might copy the test items on the original SAETs. To avoid such a disadvantage, I intentionally chose

the passages from ten years ago so that the experienced teachers might forget seeing them and the novice teachers might have a slim chance of viewing them[6]. The passages were also controlled for their topic, length, and difficulty level, the last of which is determined by the Flesch-Kincaid grade level formula. The details of the two sets of materials are presented in Table 3.

Four Mock Tests for the Scholastic Ability English Test

The major instruments used in the study were four mock Scholastic Ability English Test (SAET), the multiple-choice part only. The four mock tests were constructed by the four teachers and were taken by the forty-eight students. The formal SAET usually contains fifty-six items of multiple-choice questions. In this study, each mock test, though, contains just twenty-eight items: ten vocabulary items, ten cloze items, and eight reading comprehension questions. The procedures for producing the four mock tests are as follows.

First, each of the four teachers was asked to construct a test consisting of twenty-eight items based on the two sets of materials provided: fourteen items on Material A, and fourteen items on Material B. Then, the four tests were reshuffled into four SAET mock tests (Forms A, B, C and D, see Appendix G) in the following way as illustrated in Figure 1.

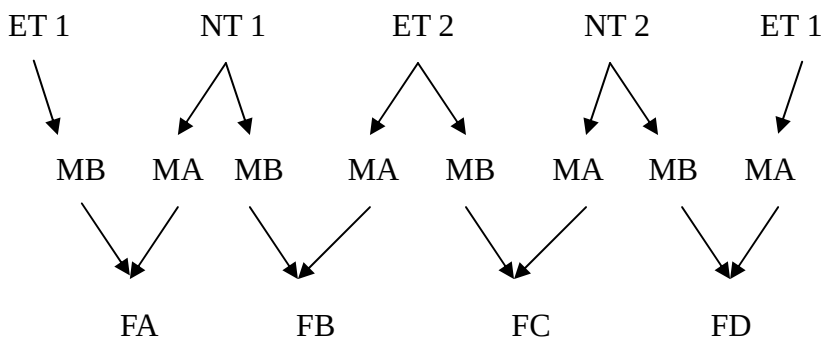
As Figure 1 shows, Form A is composed of fourteen items constructed by Novice Teacher 1 based on Material A and fourteen items by Experienced Teacher 1 based on Material B. The other three Forms were also produced in a similar way. The detailed characteristics of the four SAET mock tests (Forms A, B, C, and D) are as follows. There are twenty-eight items on each test: ten vocabulary items, ten cloze items, and eight reading comprehension items. On each form, fourteen items (including five

[6] It turned out that when asked about whether they have seen the passages in the two sets of materials, the four teachers all claimed that they had not seen those passages prior to the study.

vocabulary items, five cloze items, and four reading comprehension items) were written by an experienced teacher based on one material (A or B), while the remaining fourteen items were written by a novice teacher based on the other material (B or A). In other words, each form contains items constructed by both experienced and novice teachers on both of the two materials. There are two reasons for such arrangements. First, by combining test items written by different teachers into one test, participants can answer test items constructed by two teachers in one testing session, which helps to facilitate data collection and analysis procedures. Second, by adopting four similar forms in this study, the results I obtained to answer the research questions would be considered more reliable and convincing.

Figure 1.

Procedures for Producing Four Forms of Tests



Note. ET=Experienced teacher; NT=Novice teacher; M=Material; F=Form

Data Collection Procedures

In the present study, I adopted verbal report as the major means to collect data. All of the participants were asked to produce concurrent think-aloud protocols or retrospective protocols while they were doing their tasks. In addition, the discussion sessions at the end of each student's task also served as supplementary data. Data collection procedures consisted of two phases: collection of teachers' verbal reports and collection of students' verbal reports.

Collection of Teachers' Verbal Reports

Four senior high school English teachers were recruited to participate in the study. I met with them individually and explained to them what they should do in the study. The procedures are as follows. First, they were asked to sign a research consent form (see Appendix A). Then, they were required to fill out a background questionnaire (see Appendix B). Afterwards, they were introduced to the tasks they needed to accomplish in the study: to construct a mock test for the SAET based on the two sets of materials provided, and to produce concurrent think-aloud protocols while constructing the test. I explained to them the technique of verbal report, demonstrated how to do verbal report, and gave them chances to practice with the technique. After they were familiar with the technique, they were given a digital recorder, and were encouraged to produce concurrent think-aloud protocols (either in English or in Chinese, their mother tongue) while constructing each test item. As it has been noted that test-constructing process is usually long and laborious, the four teachers were given more time and freedom to construct their tests. That is, they were allowed to construct tests in any place, such as their home or offices, within a longer period of time. They were reminded to turn on the digital recorder whenever they started to construct the test and think aloud. After they finished their tasks, they gave their mock tests to me along with the digital recorders. Then, they were asked to fill out a feedback sheet (see Appendix C), which explored their opinions about the tasks in the study. Dates for collecting teachers' verbal report data are presented in Appendix H.

One issue worthy of note is that the act of formulating the test items might be the output of the accumulation of a lot of brainstorming. In a sense, the concurrent think-aloud protocols might not be able to cover the whole test-construction process. Therefore, the teachers were asked to produce retrospective protocols for each item as well. Finally, after the teachers finished constructing their tests, I had informal

interviews with the two experienced teachers (ET 1 and ET 2) to discuss the unclear points in their verbal report protocols. Since the two novice teachers' (NT 1 and NT 2) verbal report protocols were quite clear, and their descriptions of the rationales for constructing each item were detailed, I did not have further discussions with them.

Additionally, in order to evaluate the qualities of the tests the four teachers constructed, I invited two university professors experienced in English test construction and familiar with SAET to review the tests. The reviewers' comments also helped to triangulate the data.

Collection of Students' Verbal Reports

Forty-eight senior high school students selected from a larger sample who had taken the shortened version of FLPT were recruited to participate in the study. I arranged forty-eighty time slots to meet with the forty-eight students individually in school to collect their verbal report data (see Appendix H). All of the data collection procedures were carried out between March and June in the year 2013.

I met with the forty-eight students individually, and in each meeting session, the procedures went as follows. First, I informed the participant of the tasks he or she needed to perform in the study. Second, I asked the student to sign a research consent form (see Appendix E). Then, I introduced to the student the technique of verbal report, demonstrated how to do verbal report, and gave the student several minutes to practice doing think-aloud. I demanded that the student identify the clues or reasons why he or she chose a certain answer to each question in his or her verbal report. After the student was familiar with the think-aloud technique and my requirement, he or she was asked to start taking the assigned test (one of the four Forms), while at the same time, producing concurrent think-aloud protocols (either in English or in Chinese, their mother tongue) for each test item. There was no time limit for the student to take the test and to think aloud. They answered the test items at their own pace.

While the student was taking the test and doing think-aloud, I sat beside the student, recorded his or her verbal report with a digital recorder, and also took notes of his or her test-taking process. Since not all of the students did concurrent think-aloud successfully, I asked some of them to produce retrospective verbal report when they finished taking the test if I thought the students did not provide enough clues to their thinking process in their concurrent think-aloud protocols. At the end of each session, I asked the student about his or her opinions on the think-aloud technique, and about his or her feelings about participating in the study. In average, each session for collecting the student's data lasted about seventy minutes.

Data Analysis Procedures

There are three major sets of data in this study: (1) teacher-produced SAET mock tests, reviewers' comments on the tests, and the teachers' background questionnaire and feedback sheet results; (2) teachers' concurrent and retrospective think-aloud protocols, and (3) students' mock tests scores, and their concurrent and retrospective think-aloud protocols. Procedures for analyzing these data are as follows.

The first set of data, concerning the teacher-produced tests, were analyzed in the following way. First, I compared and contrasted the four mock tests. Then, I used the reviewers' comments as supplementary data to evaluate the qualities of the four tests. The results of the teachers' background questionnaires and feedback sheets were also presented and discussed item by item.

The two sets of verbal report data are the crucial parts in the study. The think-aloud protocols were analyzed as follows. Regarding the teachers' verbal reports, I transcribed all four teachers' think-aloud protocols verbatim on my own. Then, I tried to sort out the teachers' considerations or primary concerns for constructing each item. Afterwards, I read carefully through those considerations and formulated patterns for the teachers' test-construction considerations. Concerning the students'

verbal reports, I first transcribed all the forty-eight students' think-aloud protocols verbatim on my own. Then, I tabulated and summarized the students' strategies to answer each test item. Finally, I generalized and formulated the students' test-taking strategies based on their think-aloud protocols. Since the present study is exploratory and pioneering in nature, there is no appropriate existing patterns or schemes to describe my data. Therefore, I used an iterative process, examining the protocols again and again until I developed suitable patterns to describe the data.

In terms of the students' mock test scores, they were presented for reference but were not further analyzed. There are two reasons for not including the test scores as a main source of data. First, the test scores are not the major concern in the present study, which is qualitative in nature and focuses on the students' test-taking processes. Second, without going through any validation procedures, the reliability and validity of the four sets of tests were uncertain, so were the test scores.

In order to compare teachers' test-constructing considerations and students' considerations for answering the test items, I first figured out teachers' and students' primary considerations for each item based on their protocols, and then compared their considerations item by item to see whether they matched or not.

CHAPTER FOUR

RESULTS AND DISCUSSION ON TEACHERS' TEST CONSTRUCTION

This chapter reports the findings related to the first research question (RQ1) in the present study: What considerations do teachers take into account when they construct mock tests for the Scholastic Ability English Test (SAET)? How do the tests constructed by novice teachers differ from those constructed by experienced teachers?

There are four sets of results in this chapter, and they are presented in the following sequence: (1) the results of the teachers' background questionnaires; (2) the results on teachers' test construction processes and considerations based on think-aloud protocols; (3) the results of the teachers' feedback sheets; and (4) the analyses of the four teacher-constructed SAET mock tests. Finally, this chapter concludes with a general discussion on the test-construction performances of the two experienced teachers and of the two novice teachers.

Results of Teachers' Background Questionnaires

Two experienced teachers and two novice teachers who have been teaching English in senior high schools in northern Taiwan were invited to participate in the study. The four teachers' results to the first nineteen items (items 1, 2, 4, and 7 excluded) on the background questionnaire are summarized in Table 4.

As shown in Table 4, all four teachers have taken language testing courses in university, and they all stated that they were familiar with the old SAETs and the guidelines for constructing SAET. Furthermore, they all considered cloze items easy to construct. While three teachers (ET 1, ET 2, and NT 1) considered reading comprehension questions difficult to construct, NT 2 considered vocabulary items difficult to construct.

Table 4.

Results of Teachers' Background Questionnaires

Item	Question	ET 1	ET 2	NT 1	NT 2
3	Gender	F	F	F	M
5	Have you taken any language testing courses?	Yes	Yes	Yes	Yes
6	Number of years of teaching English in senior high school	13	14	1	2
8	Do you have experiences of constructing SAET mock tests?	No	No	Yes	No
9	Degree of familiarity with the old SAETs: (A) very familiar (B) familiar (C) unfamiliar (D) very unfamiliar	A	B	B	B
10	Degree of familiarity with guidelines for the SAET: (A) very familiar (B) familiar (C) unfamiliar (D) very unfamiliar	B	B	B	B
11	Frequency of constructing classroom tests: (A) often, every week (B) sometimes, when there is a need (C) seldom, using ready-made tests	A	A	B	B
12	Frequency of constructing English midterms or finals in school: (A) three times every semester (B) at least once every semester (C) at least once every school year	A	B	B	B
13	Number of teachers constructing midterms or finals in school: (A) one (B) two (C) more than three	A	A	A	C
14	Are the midterms or finals reviewed before being taken by students?	No	Yes	No	Yes
15	Does your school do item analysis of midterms or finals?	Yes	Yes	No	No
16	Period of time for constructing an English midterm or final exam: (A) one to three days (B) one week (C) two weeks	A	C	B	B
17	Types of questions difficult to construct: (A) vocabulary (B) cloze (C) reading comprehension (D) none	C	C	C	A
18	Types of questions easy to construct: (A) vocabulary (B) cloze (C) reading comprehension (D) none	B	B	B	B, C
19	Evaluation of your own test-constructing skills: (A) very good (B) good (C) ordinary (D) poor (E) very poor	B	C	D	B

Among the four teachers, only NT 1 had the experience of constructing SAET mocks tests while teaching in a cram school. However, she did not construct a whole test; she was given the task of constructing cloze tests and reading compression tests.

When asked to evaluate their own test-constructing skills, both ET 1 and NT 2 stated that they possess good skills. ET 2 thought that her test-constructing skill were ordinary, and NT 1 thought that her skills were poor.

As for the two open-ended questions on the background questionnaire that inquired their perceptions about constructing an exam paper and difficulties in constructing test items, the four teachers' responses are as follows.

Experienced Teacher 1 (ET 1)

When ET 1 was assigned the task of constructing a midterm or a final exam, she considered it a pleasure. She wanted to leave the impression on students that the teacher took the test-constructing task seriously, so she would make her test a little difficult so that the diligent students would obtain pleasure and a sense of achievement while taking the test constructed by her. When asked about test-constructing difficulties, ET 1 responded that she seldom experiences difficulties while constructing tests. Moreover, she often reads different kinds of articles and keeps abreast of the current events to help students gain more knowledge beyond their textbooks.

Experienced Teacher 2 (ET 2)

When ET 2 was assigned the task of constructing a midterm or a final exam, she thought it was just a routine. What she needed to take into account while constructing that test was the students' proficiency level. One difficulty ET 2 has experienced in constructing tests was that she was not sure to what extent of difficulty she could make her test based on the designated texts.

Novice Teacher 1 (NT 1)

When NT 1 was assigned the task of constructing a midterm or a final exam, she felt very happy, because she had the opportunity to re-examine the contents and the effectiveness of her teaching by constructing a midterm or a final exam. One difficulty NT 1 has experienced in constructing tests was that her students knew too little vocabulary, which limited the scope and variation of her test items and test types.

Novice Teacher 2 (NT 2)

When NT 2 was assigned the task of constructing a midterm or a final exam, he felt that the task was a pleasure, not a pressure, because he was very interested in testing and evaluation. One difficulty NT 2 has experienced in constructing tests was that when the test covered too little material, he would have fewer words to use in his test.

Analyses of Teachers' Think-aloud Protocols

This section presents the results of the analyses of the four teachers' think-aloud protocols. In analyzing the protocols, I used the induction method and tried to formulate each teacher's test-constructing process in general. Then I tried to describe each teacher's considerations when producing the three different types of test items.

Construction of Vocabulary Test Items

The four teachers were asked to construct five vocabulary items based on Material A, and five items on Material B. Each of the lists on both materials contained forty words (see Appendix F). The following sections report the four teachers' construction processes and considerations.

The Construction Processes and Considerations of Experienced Teacher 1 (ET 1)

Test-constructing process

When asked to construct five vocabulary items, ET 1 first decided on the parts of speech she wanted to construct items with. She wanted to construct items on words

with different parts of speech. She ended up constructing only one item on nouns and two items on adjectives on Material A, because she thought that nouns were usually tested in cloze items. Besides, she thought that the item on adverbs should not be put at the outset of the five vocabulary items.

Then, ET 1 examined the list word by word, checking the words that she wanted to test on. She admitted that when she saw a particular word, she would like to test on it. For example, when she saw the words "considerably, capacity, and dominant," she wanted to test on them. In fact, the three words were all the keys (i.e., the correct options) in her items on Material A. Moreover, she liked to choose frequently-tested words (e.g., *considerably* and *voluntary*) as her keys. She also took students' English proficiency into consideration when selecting the key. For example, she thought the word "squeeze" would be difficult for students, so she did not include an item on that word although she wanted to test on it.

After the keys were chosen, ET 1 began to write stems for the words. In writing a stem, she would consult online resources or her own item pool. For example, when she wanted to write a stem for the key "contract," she consulted her item pool, and revised and adopted her old item containing this word. In writing the stem, she would revise it again and again. Sometimes, when she couldn't write a proper stem for the word she had selected as key, she would discard the word, and chose another one as key.

After finishing the stem, ET 1 began to think of three distractors for each item. She did not choose words on the list as her distractors. She probably misunderstood my instructions that all four options (one correct option and three distractors) should come from the wordlist. Thus, she wrote her own distractors for each item. When writing the distractors, she would consult online resources as well. She would also put the distractor in the stem to see whether it fitted the stem. When she couldn't think of

a proper distractor, she would move on to writing the next item.

After she finished constructing the five vocabulary items, she would pay attention to the occurring frequency of a certain option (A, B, C, or D), and made each option appear equally frequently.

One uniqueness about ET 1's test-constructing process is that after she finished constructing the items, she would take a further look at the items she produced and revise them one or two weeks later. ET 1 explained that later revision was her test-construction routine, and it often helped her find out the flaws in her items she had not detected before.

Test-constructing considerations

ET 1's test-constructing considerations were related to three perspectives: the key, the stem, and the distractors.

Construction of keys. In selecting a key (i.e., the correct option of an item) from the wordlist, ET 1 made the following considerations:

(1) The keys are frequently-tested words. ET 1 claimed to be very familiar with the old SAETs; thus, she would select those frequently-tested words as her keys.

(2) The keys are polysemous. ET 1 preferred to use words with multiple meanings as her keys. For example, each of the words "contract, capacity, and collapse" has at least two different meanings. Moreover, she would test on the meaning of a polysemous word which students were less familiar with. For instance, the word "contract" has two meanings: one is "an official agreement between two parties," which is familiar to students, and the other is "to get an illness," which is unfamiliar to students. ET 1 chose the latter meaning of contract as the testing point of the key.

(3) The original chosen keys can be replaced with other words in the process of writing stems for the words. When she was unable to come up with a proper stem

for the chosen word, ET 1 would select another word as the key.

Construction of stems. In writing a stem, ET 1 took the following issues into consideration:

(1) The stem should not be too long; instead, it should be concise. Yet, she thought that items on one material should be more difficult than those on the other material. Thus, the stems she wrote on Material B were generally longer than those on Material A.

(2) The stem can include some important phrases, so that students can have more chances to learn them. For example, ET 1 put two phrases in her stems. One is "make both ends meet" in the fourth item on Material A (i.e., Recently, with oil and electricity prices going up, the commodity prices increased _____. Thus, more and more people barely make both ends meet.) The other is "on a ~ basis" in the fourth item on Material B (i.e., School on a tight budget had difficulty continuing hiring guards for the campus safety and some generous parents responded to that by donating money on a _____ basis.)

(3) The stem should contain sufficient clues to avoid ambiguity. For example, in writing the third item on Material B, ET 1 added the phrase "in the end" to the stem to justify the correct option "originally" and to falsify the distractor "unfortunately." The item is presented as follows.

ET 1: B3[7]

The demanding teacher always gives students a lot of homework and tests. _____, students complain of that. In the end, they adapt to the ways he teaches.

(A) Tentatively (B) Alternatively (C) Unfortunately (D) Originally

(4) Current events can be a topic of the stem. ET 1 wrote two stems about current events on her test. The first one is "Despite its _____ of no more than 12

[7] The code "ET 1: B3" refers to the third vocabulary item on Material B constructed by ET 1.

passengers, the van had 16 students in it." (ET 1: A3) The second one is "Without warning, the bridge _____, causing the cars to fall into the river. The tragedy was attributed to the erosion of the foundation and a lack of maintenance." (ET 1: B1)

Construction of distractors. Concerning the distractors, ET 1 made the following considerations:

(1) Distractors should be of the same part of speech as the correct option.

(2) One distractor should not be used twice on the same test. Thus, among the five vocabulary items on Material A or B, there were twenty different words for the twenty options.

(3) Words that resemble the keys in form are chosen as distractors. ET 1 liked to test students' ability to distinguish two words similar in spellings. For example, "capability" was the distractor to "capacity" (ET 1: A3), and "released" was the distractor to "revised" (ET 1: B5). She prepared her distractors this way to "mislead students on purpose[8]."

(4) Distractors are prepared in the way that they belong to the same semantic field as the key or the words in the stem so as to improve the power of the distractors. For example, the distractors in the first item on Material B (ET 1: B1) were prepared in the following way: Since the correct option was "collapsed," ET 1 thought of "constructed," which was the opposite to "collapsed." Then, because the stem included the word "bridge" and she thought "suspended bridge[9]" is a kind of bridge, ET 1 came up with "suspended" as the distractor. Finally, since a road can be "extended" and students might think that a bridge could be extended as well, the distractor "extended" was chosen. The item under discussion is as

[8] The words in quotation marks were quoted from ET 1's think-aloud protocols.
[9] Here ET 1 used the expression "suspended bridge" to refer to a "suspension bridge."

follows.

ET 1: B1

Without warning, the bridge _____, causing the cars to fall into the river. The tragedy was attributed to the erosion of the foundation and a lack of maintenance.

(A) suspended (B) collapsed (C) constructed (D) extended

(5) Distractors may collocate with the word in the stem. For example, "barrier" is the correct option to the stem (ET 1: B2) "According to the investigation, the language _____ caused the air crash. The misunderstanding of the pilot's spoken English was to blame." ET 1 wrote three distractors which all collocate with the word "language," namely, "acquisition," "requirement," and "resistance."

(6) Distractors are designed in the way that those students who think in Chinese would consider appropriate. For instance, two distractors to the first item on Material A (ET 1: A1) are "current" and "fashionable," which do not collocate with the word "language" in English. However, the expression "current language" and "fashionable language" are considered acceptable in Chinese. The item under discussion is as follows.

ET 1: B1

English is a(n) _____ language, serving as a necessary tool to communicate with people with diverse nationalities.

(A) current (B) dominant (C) accurate (D) fashionable

The Construction Processes and Considerations of Experienced Teacher 2 (ET 2)

Unlike the other three teachers, ET 2 did not provide many protocols of her real test-constructing process. She just described the rationales or principles of designing each item. It seemed that she was not accustomed to the technique of think-aloud, so she did not produce many concurrent think-aloud protocols. Compared with the other three teachers, her protocols were much shorter. Yet since she had already finished her task of producing two sets of tests, it was impossible for me to ask her to reproduce her concurrent think-aloud protocols. Therefore, I had to accept this flaw and tried to

reconstruct her test-constructing process by using her retrospective think-aloud protocols and my follow-up interviews with her.

Test-constructing process

In the initial stage of constructing the vocabulary items, ET 2 examined the list word by word, listing each word's part of speech. Then, she grouped those words according to their parts of speech. She tried to distribute her five items equally among the four different parts of speech, namely, verb, noun, adjective, and adverb. Afterwards, she decided on the keys that she wanted to test on in each group. Then, she began to write stems for the keys. If she found it hard to write a stem for a certain key, she would replace that word with another word.

In writing the first item, ET 2 would make it much easier for students to answer. She emphasized that she would not like to intimidate students with a very difficult item placed at the outset of a test. She wanted to instill confidence in students.

After finishing writing a stem, ET 2 would select three distractors from the list whose parts of speech are the same as that of the key. Finally, she would pay attention to the occurring frequency of a certain correct option (A, B, C, or D), and she particularly tried to avoid two Cs in her tests because she thought that students often guessed option C more often than they guessed the other options.

Test-constructing considerations

Unlike the little information on her test-constructing process, ET 2 provided many test-constructing considerations, which can be discussed from three perspectives: the key, the stem, and the distractors.

Construction of keys. In selecting a key, ET 2 would choose one among the group of words sharing the same part of speech. She would then select a word which is more frequently-tested, and whose meaning is quite different from the others. For example, among the eight adverbs (briefly, considerably, continuously, directly,

famously, immediately, obviously, and scarcely) on the list of Material A, the word "scarcely" has a negative meaning while the other seven does not. She selected "scarcely" as her key for the adverb item. She wanted to test students whether they know the negative meaning of this word.

Construction of stems. In writing a stem, ET 2 took the following issues into consideration:

(1) The stems should cover topics of different aspects of life. The five stems should not focus on one field only, but should include diverse topics.

(2) Students' life is a good topic for the stem. ET 2 wrote a stem about students' transferring to another school because of parents' jobs (ET 2: A5), which she thought would be familiar to students. She also wrote another item about psychology (ET 2: B3), which, she thought, was related to students' life education course. Current events are also good topics for the stem. ET 2 constructed two items related to current events (ET 2: A2, ET 2: A4, and ET 2: B1).

(3) The words in the stem should avoid rural-urban divide in case some students might not understand the stem.

(4) The key and words in the stem should collocate with each other, so that the clue is sufficient for students to single out the correct option. For example, in the fourth item on Material A (ET 2: A4), only the key "measures" collocate with the stem "take immediate _____ to put his policy into practice, " while the three distractors (figures, influences, and contracts) do not. ET 2 put a lot of emphasis on testing students' knowledge of collocations.

Construction of distractors. Concerning the distractors, ET 2 made the following considerations:

(1) The four options in an item should begin with different letters. That is, no two

options begin with the same initial letter. Despite ET 2's emphasis on this point, There is still one item (ET 2: B5) that includes two options (confess and collapse) beginning with the same letter c.

(2) The word level of the distractors is taken into account[10]. ET 2 usually selected two words with higher levels, and two with lower levels. For example, in the second item on Material A (ET 2: A2), "accurate" and "dominant" are higher level words, and "current" and "practical" are lower-level words.

(3) Among the four distractors, one or two should be unattractive. ET 2 chose the distractors this way on purpose. She wanted students to delete two unlikely options quickly, and to choose one between the remaining two or three options. For example, the distractor "briefly" (ET 2: A3), according to ET 2, is very unlikely to be the correct option to the stem "Informed of his admission to his ideal university, Patrick could _____ control his joy and let out a cry." Therefore, it should be deleted first by students.

(4) Distractors should belong to the same semantic field as the key or the words in the stem. For example, the distractor "contracts" (ET 2: A3) was chosen because it belongs to the same semantic field as the word "policy" in the stem.

(5) Distractors should be chosen to "mislead those students who like to think in Chinese in answering English questions[11]." For example, according to ET 2, the distractor "approaches" (ET 2: B1) was a trap for students to fall into. This item is presented as follows.

ET 2: B1

With the release of its new smart phones, the manufacturer Nokia _____ 160% more app download numbers than Apple.

(A) boasts (B) revises (C) maintains (D) approaches

[10] In her protocols, I didn't find any clues as to how ET 2 determine the levels of the words she chose. She probably did so based on old memories or consulted some references.

[11] The words in quotation marks were quoted from ET 2's think-aloud protocols.

The Construction Processes and Considerations of Novice Teacher 1 (NT 1)

Test-constructing process

In the beginning, NT 1 read out loud the words on the list one by one. Then, she labeled each word with its part of speech, and divided those words into four groups based on parts of speech, namely, verbs, nouns, adjectives, and adverbs. Next, she selected four words from each group, with the four words similar in one aspect. For example, the four options "rapidly," "gradually," hardly," and "urgently" (NT 1: B3) are about time or speed. Afterwards, she chose one word from the four pre-selected words as the key.

With the keys, NT 1 began to write the stems. Unlike ET 1, who referred to online resources or her item pool, NT 1 didn't consult any resources. She just think of sentences on her own. She was careful in providing sufficient clues in the stems. So after she wrote a stem, she would revise it again and again until she found it appropriate. NT 1 also admitted that when she saw a certain set of words, she would like to make a sentence about them. For example, when she saw the words "preserve, maintain, confess, and freeze," she wanted to make a sentence about people's relationship, and she did so as well (see NT 1:B5).

Finally, NT 1 stated that she would like to make equal distribution of the occurring frequency of the correct option A, B, C, or D. Yet, there are still three correct option Cs and no As among her Material A vocabulary items.

Test-constructing considerations

NT 1's think-aloud protocols were the second shortest among the four teachers. When she constructed vocabulary items, she put much emphasis on the choice of distractors. The following are the considerations she made about selecting distractors.

(1) The four options should be of the same part of speech, similar in meaning and usage. For example, the four options in the fourth item on Material B are all

nouns related to abstract objects, and three of them (i.e., exception, approach, and solution) can collocate with the preposition "to." The item under discussion is as follows.

NT 1: B4

The _____ to the regulation on speed limit is allowed when the ambulance is on its duty.

(A) exception (B) approach (C) solution (D) technique

(2) The words that are similar in meaning to the key make good distractors because "they can confuse students easily[12]." For example, the words "preserve" and "maintain" (NT 1: B5) are quite similar in meaning. NT 1 expected that the two words would make students confused, and thus they would choose the wrong answer.

(3) To make distractors more attractive, they should share with the words in the stem the same semantic field. Like the two experienced teachers, NT 1 used this method to "set a trap[13]" for students to fall into. For example, the distractor "violate" is related to the word "law" in the stem "As a good Taiwan citizen, we should _____ whatever is against the law." (NT 1: A3) When students saw the word "law," they might choose the option "violate" if they did not understand the whole sentence, since "violate the law" is a collocation students might have learned before.

The Construction Processes and Considerations of Novice Teacher 2 (NT 2)

Test-constructing process

NT 2 had the longest think-aloud protocols among the four teachers. He was very careful in designing his test items, and was also very successful in doing concurrent think-aloud. The process of how he made vocabulary items are presented as follows.

Like ET 2 and NT 1, NT 2 began his task by labeling each word's part of speech.

[12] The words in quotation marks were quoted from NT 1's think-aloud protocols.
[13] The words in quotation marks were also quoted from NT 1's think-aloud protocols.

He then decided that he would test on words of each part of speech, namely, verb, noun, adjective, and adverb. He emphasized that the four options in an item should belong to the same part of speech.

Next, NT 2 further divided the words in one part of speech into several sub-groups based on his personal judgment. For instance, he divided all the nouns on the list of Material A into four sub-groups: abstract nouns, nouns concerning emotions, figures, and objects. He then chose the words in a sub-group as his options for an item so that the four options would be similar in some way. Afterwards, NT 2 referred to the English reference word list compiled by CEEC to find out the levels of the four options he selected. Among the four options, he usually chose the word with the higher level on the word list as his key.

After deciding on the key, he started writing a stem for that word. He was very careful in writing the stem, paying special attention to the usage of the key. He consulted many dictionaries in writing a stem in order to make the sentence grammatically appropriate. In the process of writing the stem, he would revise it again and again until it looked flawless.

NT 2 had pre-arranged the positions of the correct options for the five items, and he also controlled the occurring frequency of options A, B, C, and D. For example, on Material A, the correct options were set as DACBD. He then put each of the correct options in its slot, and randomly distributed the other distractors to the remaining slots.

Test-constructing considerations

NT 2's test-constructing considerations can be discussed from three aspects: the key, the stem, and the distractors.

Construction of keys. NT 2 was very prudent in selecting the keys. In addition to the rigorous key-selection procedure mentioned above, he also avoided testing on one

word which had appeared somewhere else on the same test. For example, he first selected "capacity" as his key. But later he found that the word also appeared in the cloze passage on Material B. Therefore, he replaced the word "capacity" with "decrease," and used "decrease" as his key (NT 2: A2).

Construction of stems. In writing a stem, NT 2 took the following issues into consideration:

(1) The stem should not be long. One or two sentences would suffice.

(2) The stem should contain sufficient clues to avoid ambiguity.

(3) The topics of the stems should be related to current events or any subject that students are familiar with.

(4) The words in the stem should avoid rural-urban divide and digital divide in case some students might not understand the stem.

(5) The stem should not contain difficult grammatical structures. For example, the original stem "…to punish those who drive cars after having drinks" was revised as "…to punish those who drive cars after they have drinks." NT 2 used the simpler adverbial phrase to replace the more difficult participial phrase.

(6) The stem should not provide extraneous grammatical clue to students lest they use collocation knowledge to select the keys. For instance, in writing the stem for the key "decrease" (NT 2: A2), NT 2 first came up with the phrase "on the decrease." But, he was aware that only the key fitted the structure "on the ~" while the other distractors couldn't. As a result, he revised the stem as "Because of the _____ of birth rate,…" to eliminate the possibility that students used this grammatical clue instead of their vocabulary knowledge to answer this item. The item under discussion is as follows.

NT 2: A2

Because of the _____ of birth rate, the country is now facing the problem of

having fewer and fewer young people to support it.

(A) decrease (B) addition (C) capacity (D) routine

Construction of distractors. As for the distractors, NT 2 made the following considerations:

(1) Distractors had better be similar in meaning to their correct option. That was why NT 2 went through a rigorous procedure of selecting a group of options for an item.

(2) Distractors should be equally attractive. NT 2 tried hard to achieve this goal by revising the stem again and again. If his chosen distractors were weak in elicitation when put in the stem, NT 2 would replace them with new ones. For instance, the option "contest" was replaced with "relative," "legend" with "technique," and "gossip" with "division" in the third item on Material B (NT 2: B3), which is presented as follows.

NT 2: B3

Miranda filled in the blank of _____ with the word "nurse," which means she worked as a nurse.

(A) occupation (B) relative (C) technique (D) division

A Summary of teachers' test-constructing process and considerations

Based on the four teachers' think-aloud protocols, their test-constructing processes were similar in the following ways:

In the initial stage of test-construction, ET 2, NT 1 and NT 2 would label each word's part of speech, and put the words into different groups. ET 1, on the other hand, did not go through this process. She just checked the words that she wanted to test on. Then, all the teachers selected words that they wanted to test on. They chose one or two words from each of the four part-of-speech groups as key(s). Moreover, ET 2, NT 1, and NT 2 would choose the key along with three distractors at a time for one item.

After the key and the distractors were in place, the teachers began to write the stems. All the teachers would revise the stem again and again until they thought the

stem was appropriate. ET 1 was different from the other teachers in this stage because she would write the stem first, and then think of the other three distractors for the stem. ET 1 was the only teacher that wrote her own distractors while the other three teachers selected the distractors from the word lists.

After the five items were produced, all the teachers would modify the positions of the keys to make them occur equally frequently.

As for the four teachers' test-constructing considerations, the major points can be summarized in Table 5. As Table 5 shows, although the four teachers had their own considerations in constructing vocabulary items, they still shared some ideas in common. For example, all the four teachers thought that the stems should contain sufficient clues, and that the distractors should be of the same part of speech as the key. In addition, the two experienced teachers had one similar consideration in choosing the keys: they thought that the keys had better be frequently-tested words. The two novice teachers also had one similar consideration in designing the distractors: they thought that the distractors could be similar in meaning to the key.

Furthermore, the two experienced teachers and NT 1 shared two similar considerations in producing the distractors. They thought that the distractors should share the same semantic field as words in the stem, and that the distractors could be designed in a way to attract those students who think in Chinese. On the other hand, the two experienced teachers and NT 2 also shared two similar considerations. They thought that the stems could contain current events, and that the distractors should not be used twice on the same test.

Taken as a whole, it is found that the two experienced teachers made seven similar considerations in constructing vocabulary items while the two novice teachers made only three similar considerations. In other words, the two experienced teachers were more in line in considerations with each other when constructing vocabulary

items than the two novice teachers were with each other.

Table 5.

Teachers' Considerations in Constructing Vocabulary Items

Considerations	ET1	ET2	NT1	NT2
The key				
is frequently-tested	√	√		
is polysemous	√			
does not appear in other places on the same test				√
The stem				
contains sufficient clues	√	√	√	√
is not too long	√			√
is about current events	√	√		√
is related to students' life		√		√
covers diverse topics		√		
includes important phrases	√			
should avoid rural-urban divide		√		√
should avoid digital divide				√
does not contain difficult grammatical structures				√
does not provide extraneous grammatical clues				√
Distractors				
are of the same part of speech as the key	√	√	√	√
share the same semantic field as words in the stem	√	√	√	
are designed in a way to attract those students who think in Chinese	√	√	√	
are similar in meaning to the key			√	√
are similar in form to the key	√			
are not used twice on the same test	√	√		√
word levels are taken into account		√		√
collocate with the word in the stem	√			
begin with different letters		√		
should be equally attractive				√
should not be equally attractive		√		

Construction of Cloze Test Items

The four teachers were asked to construct five cloze items based on a passage in

Material A, and five items on a passage in Material B (see Appendix F). The teachers were asked not to rewrite the two passages. All they had to do was locate five testing points (or blanks) in each passage and construct four options for each blank.

The Construction Processes and Considerations of Experienced Teacher 1 (ET 1)

Test-constructing process

ET 1 considered it easy to construct cloze tests in this task since she did not need to rewrite the passage. In the construction process, she first read the passage slowly, and while reading, she circled the testing points she wanted to test on. As she read along, she had some question types in mind; that is, she knew what points (e.g., noun, verbs, transition, phrase, etc.) she was going to construct an item on. Then, after several times of examining the passage, she decided on five blanks in the passage she wanted to test on. Later, if she found it hard to write distractors for a particular blank, she would replace it with a new one.

Afterwards, ET 1 began to write three distractors for each item. In constructing the distractors, she would consult other references for ideas or to confirm the grammar and correct usage of a distractor. If she could not come up with good distractors at that time, she would put the item aside for later revision.

As in constructing vocabulary items, ET 1 took a second look at the cloze items she produced one or two weeks ago. At this stage, she revised some distractors and completed each item with three distractors.

Test-constructing considerations

ET 1's test-constructing considerations can be discussed from two perspectives: the testing points (the blanks) and the distractors.

Choice of the words to be removed and tested. In terms of choosing the testing points, ET 1 took the following issues into consideration.

(1) The first sentence should not contain any testing points, and should be left intact.

ET 1 thought that the first one or two sentences gave students the main idea of the passage, and that they should not contain any blanks.

(2) Two testing points or blanks should not come too close.

(3) Nouns are often tested in cloze items. ET 1 mentioned the tendency she observed in the past SAETs. Therefore, she constructed noun items more often than the other three teachers did.

(4) Difficult points make good cloze blanks. ET 1 liked to test on difficult parts of the passage because she thought that it would be more challenging to students.

Construction of distractors. Concerning the distractors, ET 1 made the following considerations.

(1) Frequently-tested words or phrases make good distractors. For example, in constructing distractors to compete with the key "rather than" (ET 1: A2), ET 1 came up with "instead of" and "regardless of," which are two frequently-tested phrases on students' tests.

(2) Words or phrases that resemble the key in spelling are good distractors. ET 1 put those words or phrases in an item on purpose. She wanted to test students' ability in distinguishing words with similar forms. For example, the distractor "expectations" was written along with the key "exceptions" (ET 1: A3), and the distractor "filled with" was in place with the key "filled out" (ET 1: B5).

(3) Distractors can be the phrases that students are confused about. For instance, ET 1 constructed the distractors "to be" and "to have" for the stem (ET 1: A4) in order to "confuse students" because she found that many students could not tell the difference between the two phrases. The item is as follows.

ET 1: A4

For example, John Steinbeck is said __4__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) being (B) to be (C) to have been (D) to have

(4) Distractors should not be too strong in elicitation lest they should be considered acceptable answers as well. ET 1 emphasized the point in designing the distractors to the fifth item on Material A (ET 1: A5). She first thought of phrases like "based on" and "judging from," but she considered them ambiguous. So she gave up those two phrases, and chose "different from" instead, although it was not a participial phrase. The item is presented as follows.

ET 1: A5

Thus, __5__ what a good listener does, he may become either popular or disliked in his lifetime.

(A) depending on (B) speaking of (C) compared with (D) different from

The Construction Processes and Considerations of Experienced Teacher 2 (ET 2)

Test-constructing process

Little data was available in ET 2's protocols about the whole process of how she constructed cloze items. She did not describe how she selected the five blanks in each passage. She just mentioned her principles of designing cloze items and her considerations or rationales of constructing each item.

Test-constructing considerations

Choice of the words to be removed and tested. Regarding the testing points, ET 2 made the following considerations.

(1) Testing points should focus on phrases, grammar, and structures.

(2) Frequently-tested items make good testing points. ET 2 stated that when she saw relative pronoun "what" in the passage, she wanted to test on it because it was a frequently-tested point and students were often confused about relative pronouns.

(3) One in the five items on one passage should be difficult. ET 2 admitted that she constructed a very difficult item on the passage in Material A (ET 2: A3). In that item, she tested students' knowledge of verb tense, grammar, and contextual

meaning all at a time. She wanted students to spend more time on this item, which is shown as follows.

ET 2: A3

For example, John Steinbeck __3__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) was portrayed as (B) was told to pretend

(C) is prone to be (D) is said to have been

Construction of distractors. In terms of the distractors, ET 2 took the following issues into consideration.

(1) In designing four options for an item testing on verb phrases, the four verbs and four prepositions should be different. No two verbs or prepositions should be the same. ET 2 was careful about this in constructing the fourth item on Material A (ET 2: A4). The four options for that item were "contributed to," "originated from," "involved in," and "substituted for."

(2) The distractor can be the opposite of the correct item (i.e., testing point). For example, ET 2 constructed a distractor of "originated from," which is the opposite of "contributed to" (ET 2: A4). She wanted to test on students' knowledge of cause-and-effect relationship. The item is as follows.

ET 2: A4

No doubt his ability to listen __4__ his capacity to write.

(A) contributed to (B) originated from (C) involved in (D) substituted for

(3) Distractors can be words that collocate with the words in the passage. For instance, the two distractors "figured out" and "sent out" all collocate with "the form" (ET 2: B5). The item under discussion is as follows.

ET 2: B5

…it is usually good to arrive earlier than the appointment because there are usually forms that need to be __5__ by the patient.

(A) figured out (B) set out (C) filled out (D) sent out

(4) Distractors can be similar in form to each other. For instance, ET 2 produced two distractors "sent out" and "set out" for the fifth item on Material B (ET 2: B5).

(5) Distractors can be words or phrases that students always make a mistake on or are often confused about. For example, the distractor "Despite" was used in the item (ET 2: A1) along with other options which were all conjunctions because ET 2 found that students were often unable to tell whether the word "despite" was a conjunction or a preposition. This item is presented as follows.

ET 2: A1

__1__ they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people.

(A) Because (B) Unless (C) If (D) Despite

(6) Distractors can be designed in a way to "mislead those who think in Chinese." The distractor "was portrayed as" (ET 2: A3), according to ET 2, was used to lure those students who think in English into "a trap[14]." This item is presented as follows.

ET 2: A3

For example, John Steinbeck __3__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) was portrayed as (B) was told to pretend

(C) is prone to be (D) is said to have been

(7) One of the distractors should be designed to be less attractive. ET 2 would put a very unlikely option in her item, and she wanted students to eliminate this unlikely option and choose one answer among the remaining three options.

The Construction Processes and Considerations of Novice Teacher 1 (NT 1)

Test-constructing process

In the initial stage of constructing cloze items, NT 1 read the passage first. Then, she selected five testing points, and began to write the five cloze items. She would first place the key in a designated place, and then began to write distractors for the remaining slots. In this process, she modified or revised her distractors again and again until she found them acceptable or appropriate.

[14] The words in quotation marks were quoted from ET 2's think-aloud protocols.

Test-constructing considerations

Choice of the words to be removed and tested. In selecting the testing points, NT 1 made the following considerations.

(1) Items should focus on what students have learned or the frequently-tested points on SAETs. For example, she thought that students should know the usage of transitions, which occur on SAETs frequently, so she constructed three items on transitions among the ten cloze items.

(2) The comparative structure can be a good testing point, especially if one of the words in the structure is far away from the other. For instance, in the sentence "good listeners tend to know more and to be more sensitive to what is going on around them than other people," the word "than" is far away from "more." Thus, "than" makes a good testing point.

Construction of distractors. When constructing the distractors, NT 1 took the following issues into consideration.

(1) Distractors can be designed to "confuse students." For example, NT 2 wrote the distractor "wisely" (NT 1: B2) to confuse those careless students who took "wisely acceptable" as "widely acceptable." The latter is a correct collocation while the former is not, although both are not the key in this item.

(2) Distractors can collocate with words in the passage. For instance, in designing distractors for the first item on Material A (NT 1: A1), NT 1 wrote three distractors "with," "about," and "as" for the blank "_____ other people." All of the three distractors can collocate with the two words "other people." The item under discussion is presented as follows.

> NT 1: A1
> Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them __1__ other people.
> (A) as (B) with (C) than (D) about

(3) Distractors can be the opposite of the key. In the third item on Material B (NT 1: B3), NT 1 wrote the distractor "before" to compete with the key "after." This item is presented as follows.

NT 1: B3

A person usually tried to arrive about five minutes __3__ the invitation time, so that the host would have a little extra time to prepare for the guests.

(A) after (B) before (C) on (D) by

(4) Students' test-taking strategies can be good sources for constructing the distractors. For example, NT 1 assumed that some students would choose the option "In conclusion" when the blank appeared at the end of the passage. Therefore, when designing the distractors for the last item on Material B (NT 1: B5), she included "In conclusion" as one distractor for that item, which is presented as follows.

NT 1: B5

__5__, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be filled out by the patient.

(A) Actually (B) However (C) In conclusion (D) Unfortunately

The Construction Processes and Considerations of Novice Teacher 2 (NT 2)

Test-constructing process

NT 2 produced very detailed concurrent think-aloud protocols of how he constructed cloze items. Like the other teachers, he read the passage first. Then, he counted the number of the sentences (the first and the last sentences excluded) in the passage. He wanted his items distributed equally among those sentences. Since there were two paragraphs in the passage of Material B, NT 2 decided to construct three items in the first paragraph which contained more sentences, and two items in the second paragraph which contained fewer sentences.

Afterwards, NT 2 examined each sentence, trying to locate the potential testing

points. He located several places, and then selected among them five testing points that he wanted to test on. After the blanks were chosen, NT 2 placed the keys in the pre-arranged positions. Then, he began to write distractors for each item, and put them in the remaining slots. When he was designing the distractors, he would refer to a synonym dictionary for more ideas, or consult some dictionaries to make sure that the usage of a word for his distractor was correct. NT 2 would revise the distractors again and again until he thought they made good distractors.

Test-constructing considerations

Choice of the words to be removed and tested. In terms of testing points, NT 2 took the following issues into consideration.

(1) Leave the first and the last sentences intact. NT 2 followed this principle in constructing items on Material A, but he violated it when he constructed an item on the last sentence on Material B.

(2) Two items should not come close to each other in case there might not be enough clue to answer them.

(3) The context where each testing point locates should contain sufficient clues.

(4) The five items on a cloze passage should test on different aspects of the language. For example, on the passage of Material A, NT 2 constructed one item on transition, one on noun, one on comparative adjective, etc.

(5) A testing point should be easy for teachers to construct distractors for it. NT 2 would discard a previously-selected testing point if he found it hard to produce distractors for it.

(6) Transitions make good testing points because students have to use contextual clues to answer the item.

Construction of distractors. When constructing distractors, NT 2 took the following issues into account.

(1) The distractors should be attractive.

(2) The order of the four options is taken into consideration. For example, when making the fourth item on Material A (NT 2: A4), NT 2 arranged the four options as follows: (A) have been (B) has been (C) had been (D) having been, in the order of present tense, past tense, and past perfect tense.

(3) If the testing point is on a verb phrase (verb + preposition), the prepositions in the four options should be different. For instance, the four options NT 2 constructed for the fifth item on Material A are "resulted from," "contributed to," "consisted of," and "persisted in" (NT 2: A5).

(4) Words or phrases that contain the opposite meaning of the keys make good distractors. For example, NT 2 constructed "inclusions" to compete with "exceptions" (NT 2: A3), and "resulted from" to compete with "contributed to" (NT 2: B5).

A Summary of teachers' test-constructing process and considerations

Except for ET 2, whose test-constructing process was unknown, the other three teachers (ET 1, NT 1, NT 2) went through similar processes in constructing cloze items. They all read the passage first, and then decided on five testing points. Sometimes, they would change the testing point if they found it hard to come up with proper distractors. Next, they would pre-arranged the locations of the keys, and controlled the occurring frequency of each option A, B, C, or D. Finally, they began to write distractors for each item, and revised them again and again until they considered the distractors appropriate.

As for the four teachers' test-constructing considerations of cloze items, the major points are summarized in Table 6.

Table 6.

Teachers' Considerations in Constructing Cloze Items

Considerations	ET1	ET2	NT1	NT2
The testing points				
should not fall on the first sentence [a]	√	√	√	√
should not fall on the last sentence				√
should not come too close to each other	√			√
should test on different language aspects	√	√		√
focus on nouns	√			
focus on phrases, grammar, and structures		√		
can be frequently-tested points in SAETs		√	√	
can be a little difficult for students	√	√		
can be transitions			√	√
should be easy for teachers to construct distractors				√
should be equally distributed on the whole passage				√
Distractors				
should be frequently-tested words or phrases	√	√	√	
should be similar in form to the key	√			
should be similar in form to each other		√		
can be opposite in meaning of the key	√		√	√
can be opposite in meaning of each other				√
should contain a less attractive one	√	√		
should collocate with the words in the passage			√	
should be equally attractive				√
can be words or phrases that confuse students	√	√	√	
can be designed in a way to trick those students who think in Chinese	√	√	√	
the prepositions in each verb phrase should be different		√		√

[a] Although ET 2 and NT 1 did not mention this point in their protocols, it seems that they also took this into consideration because they did not construct any item on the first sentence.

As Table 6 shows, all the four teachers shared one similar consideration in constructing cloze items. They all thought that the testing points should not fall on the first sentence. As for the differences between experienced teachers and novice teachers, there are two considerations shared only by the two experienced teachers,

and one consideration shared only by the two novice teachers. The experienced teachers thought that the testing points could be a little difficult for students, and that the distractors should contain a less attractive one. This suggests that although the experienced teachers chose to test on difficult points, they would make one of their options easier. The only one consideration shared by novice teachers is that the testing points could be transitions. Bearing this consideration in mind, the two novice teachers ended up constructing more transitions items than the two experienced teachers did (see Table 12).

The two experienced teachers also had other considerations shared by NT 1 or NT 2. For example, the two experienced teachers and NT 1 had three considerations in common. They all thought that the distractors should be frequently-tested words or phrases, could be words or phrases that confuse students, or could be designed in a way to trick those students who think in Chinese. It seems that the three teachers, in designing distractors for the cloze items, tended to focus on words or phrases that students often made mistakes on. The two experienced teachers and NT 2 shared only one consideration. They all thought that the testing points should focus on different language aspects. On the other hand, the two novice teachers and ET 1 shared one similar consideration. They all thought that the distractors could be opposite in meaning to the key.

Taken as a whole, the two experienced teachers had seven considerations in common, and the two novice teachers shared three similar considerations. As was the case in constructing vocabulary items, the two experienced teachers were more in line in considerations with each other when constructing cloze items than the two novice teachers were with each other.

Construction of Reading Comprehension Questions

The four teachers were asked to construct four reading comprehension questions

based on a passage in Material A, and four questions on a passage in Material B (see Appendix F).

The Construction Processes and Considerations of Experienced Teacher 1 (ET 1)

Test-constructing process

ET 1 stated in her protocols that it was very difficult to construct reading comprehension questions. The general process of how she constructed reading comprehension questions is as follows.

In the beginning, ET 1 read the passage first. While reading, she examined the passage carefully to find out some points she could test on. Then, she usually came up with some general questions. For example, she would have a question on "main idea" or "best title," and a question on details. When she read along and saw some pronouns or difficult words, she would want to test on them. After she went through the whole passage, she had decided four types of questions she wanted to construct. Then, she began to write the four items.

It took ET 1 a lot of effort to construct the four items. First, she wrote the question stem. Then, she constructed the correct option. Finally, she produced the three distractors. In the item-writing process, she referred to the passage frequently in order to produce appropriate and attractive options for each item. She also modified or revised the options to avoid ambiguity.

Test-constructing considerations

ET 1's considerations can be discussed from three aspects: the selection of the testing points, the question stem, and the construction of the options (including the key and three distractors).

Choice of testing points. In terms of selecting the testing points, ET 1 mentioned that she usually did not construct items on the initial part of a passage, and that she tended to select difficult points to test on. Furthermore, ET 1 also emphasized that the

four questions should be ordered in the sequence of the testing points that appear in the passage. For example, the question testing on the points in the second paragraph should come before the question testing on the points in the third paragraph.

Construction of stems. With regard to the question stem, ET 1 was careful with the wording of the stem. For example, if she wanted to construct an item on details of the passage, she would include the phrase "according to the passage" in her question stem. According to ET 1, this was to prevent students from answering the question based on their prior knowledge rather than on the information provided in the passage.

Construction of options. In terms of writing the options, ET 1 made the following considerations.

(1) The wording of the correct option should not be the same as that of the sentence in the passage. The correct option had better be a paraphrase of rather than a replicate of the original sentence in the passage.

(2) Distractors can also contain the same words or phrases in the passage. This was to "trick" those students who did not read the passage but only skimmed for the same words in the passage.

(3) Distractors can be constructed by rewriting or modifying a particular sentence in the passage.

(4) The question stem and the options can be described with some difficult words. ET 1 thought that this could also test on students' vocabulary knowledge. For example, in the option A of the second item on Material B, ET 1 used a difficult word "dubbed" (ET 1: B2), and in the question of the fourth item on Material B, she used another difficult word "synonymous" (ET 1: B4).

The Construction Processes and Considerations of Experienced Teacher 2 (ET 2)

Test-constructing process

There was also little data about ET 2's process of constructing reading

comprehension questions. But ET 2 did mention in her protocols the process of how she prepared the four options for her item. The process is like the following: First, ET 2 produced the correct option to each item and placed it in "A" slot. Afterwards, she began to write the other three distractors. Finally, after she completed all the options in each item, she then re-arranged the positions of the correct option.

Test-constructing considerations

ET 2 mentioned some considerations of constructing her reading comprehension questions. First, in selecting testing points in the passage on Material B, ET 2 chose to test on an easy point, asking students to identify who David Smith was (ET 2: B1). Second, as for the type of comprehension questions, ET 2 thought that "the author's opinion" was a must-test type. Thus, she constructed an item asking for the author's opinion toward dress code (ET 2: B4). Third, in writing the question stems, ET 2 would use difficult expressions to test on students' vocabulary knowledge. For instance, she used "What is the thesis statement of this article?" to describe the common expression "What is the main idea of this article?" (ET 2: A1). Fourth, as for the construction of the options, ET 2 emphasized the principles of using synonyms or paraphrases of the words in the passage instead of copying the original words, phrases, or sentences in the options. Finally, she also constructed distractors by rewriting or modifying a particular sentence in the passage.

The Construction Processes and Considerations of Novice Teacher 1 (NT 1)

Test-constructing process

NT 1's test-constructing process is as follows. First, she read the passage through, trying to understand what it was about. Then, she summarized some key points in the passage that she would construct an item on. After that, she began to write each item.

In constructing each item, she wrote the question stem first, based on one of the potential testing points she located earlier. Next, she wrote the correct option and

placed it in a pre-determined place A, B, C, or D. Finally, she started to construct the three distractors one by one.

Test-constructing considerations

Like ET 2, NT 1 did not mention a lot about the considerations she made in producing reading comprehension questions. By analyzing her protocols carefully, I still found three issues she took into account when constructing the options.

(1) Options can be made by rewriting or paraphrasing the sentences in the passage.

(2) Options can be new ideas created by teachers. For example, the distractor "A successful gold seeker must have a pair of expensive shoes" (NT 1: A3) was not mentioned in the passage, and was a new idea created by NT 1.

(3) Distractors can be made by changing a word in the original sentence. For example, the distractor "Many gold seekers died of hunger and dry weather" came from the original sentence "many died of starvation and exposure to the cold weather" (NT 1: A3). By replacing the word "cold" with "dry," a distractor was created by NT 1.

The Construction Processes and Considerations of Novice Teacher 2 (NT 2)

Test-constructing process

NT 2's test-constructing process can be summarized as follows. In the beginning, NT 2 read the passage through. Then, he gave an oral summary of the passage, listing several key points in the passage. Next, he decided on four types of questions that he wanted to construct, namely, global, local, referential, and inferential questions, because he thought that the four types of questions also tested on different reading skills (e.g., skimming and scanning). However, he ended up discarding the inferential type of questions on his test because he could not find any proper pronouns on the two passages to construct inferential questions.

Then, NT 2 began to construct the question stem for the designated item type

(i.e., global, local, and referential). At this stage, he would revise the stem repeatedly until he found it appropriate. He would also consider whether it was easy to construct the options for the stem. If he could not come up with enough options for a stem, he would give up the original stem and think of a new one.

After the stem was constructed, NT 2 began to write the correct option, and then the three distractors. He would also revise the four options to make sure they were acceptable. After the four options were constructed, NT 2 would re-arrange the locations of the four options, with the shortest one placing in "A" slot, and the longest one in "D" slot. He also controlled the occurring frequency of each option A, B, C, and D, making each option occur once. Finally, NT 2 would take a second look at the four items for minor modifications.

Test-constructing considerations

NT 2 only mentioned some considerations in constructing the distractors, as presented below.

(1) Distractors can be created by making a new sentence with words or phrases in the passage. For example, the distractor "To convince readers that the woman who dropped the shoes must be a millionaire" (NT 2: A1) was made based on the word "millionaire" in the passage.

(2) Distractors can be created by making a new sentence with synonyms of words in the passage. For example, in the distractor "The journey required its participants to carry necessities with them" (NT 2: A2), the word "necessities" was used as the synonym of the word "supplies" in the passage.

(3) Distractors can be constructed by rewriting or modifying a particular sentence in the passage.

A Summary of teachers' test-constructing process and considerations

The four teachers had similar processes in constructing reading comprehension

questions. They all read through the passage first, and then located some key points of the passage they could focus on. They also decided first what types of questions they wanted to construct, and then they began to write the question stem. Next, they constructed the correct option, followed by three distractors.

Among the four teachers, NT 2 produced longer protocols than the other three teachers. He not only introduced his test-constructing principles and rationales, but also described his test-constructing process in detail. He was also the only one that emphasized the order of the length of the four options and the location of the correct option.

In terms of the four teachers' test-constructing considerations of reading comprehension questions, the major points are summarized in Table 7.

Table 7.

Teachers' Considerations in Constructing Reading Comprehension Questions

Considerations	ET1	ET2	NT1	NT2
The testing points				
should not fall on the initial part of a passage	√			
should be difficult	√			
should be easy		√		
should cover different types (i.e., global, local, referential, or inferential)				√
The question stem				
should be specific in terms of wording	√			
should contain difficult words or expressions	√	√		
Options				
can be paraphrases of the sentences in the passage	√	√	√	
can contain synonyms of the words in the passage		√		√
can contain the same words in the passage	√			√
can contain new ideas created by teachers			√	
can be made by changing a word in the sentence			√	
should be ordered in length, with the shortest one first				√

As Table 7 shows, there are no similar considerations shared by all the four

teachers. It seems that the four teachers had their own opinions about constructing reading comprehension questions. In spite of this, there is one consideration shared by the two experienced teachers; namely, the question stem should contain difficult words or expressions. There is still one consideration shared by the two experienced teachers and NT 1: they all thought that the options could be paraphrases of the sentences in the passage. The two novice teachers, however, shared no similar considerations when constructing reading comprehension questions.

Results of Teachers' Feedback Sheets

The four teachers were asked to fill out a feedback sheet (see Appendix C) after they finished the task of constructing SAET mock test while doing think-aloud. The results of the first five items are presented in Table 8.

Table 8.

Results of Teachers' Feedback Sheets

Item	Question	ET 1	ET 2	NT 1	NT 2
1	Regarding this task, which material is difficult for you to construct tests on? (A) Material A (B) Material B	B	B	B	B
2	Regarding this task, which type of tests is the most difficult for you to construct? (A) vocabulary (B) cloze (C) reading comprehension	C	C	C	C
3	Regarding this task, which type of tests is the easiest for you to construct? (A) vocabulary (B) cloze (C) reading comprehension	B	A	B	B
4	Regarding this task, which of the following is the most difficult for you to construct? (A) vocabulary in MA (B) vocabulary in MB (C) cloze in MA (D) cloze in MB (E) reading in MA (F) reading in MB	F	F	F	E
5	Regarding this task, which of the following is the easiest for you to construct? (A) vocabulary in MA (B) vocabulary in MB (C) cloze in MA (D) cloze in MB (E) reading in MA (F) reading in MB	C	C	C	C

As shown in Table 8, the four teachers seemed to have similar opinions of the task. Three of them (ET 1, ET 2, and NT 1) considered it most difficult to construct reading comprehension questions in Material B, and all the four teachers considered it easiest to construct cloze tests in Material A.

These results generally corresponded to those in their background questionnaires, in which three of them (ET 1, ET 2, and NT 1) mentioned that they found it difficult to construct reading comprehension questions, and that they considered it easy to construct cloze tests (see Table 4). But there were some differences between NT 2's responses to the background questionnaire and his responses to the feedback sheet. Prior to the study, NT 2 thought that vocabulary items were difficult to construct, and cloze tests and reading comprehension questions were easy to construct. However, in doing this task, he considered it hard to produce reading comprehension questions, and he found it easiest to construct cloze tests.

As for the two open-ended questions on the feedback sheet, the four teachers' responses are as follows.

When asked about the influence of the think-aloud method on the test-constructing process, the teachers had different opinions. ET 1 thought that doing think-aloud while constructing the tests was time-consuming; in addition, she was sometimes slower in doing think-aloud than thinking and typing the items. ET 2 thought that think-aloud method helped her search for or memorize almost all ideas in test-constructing process. Since this was her first time doing think-aloud, she thought it was interesting. NT 1 thought that think-aloud method helped her get a clearer picture of the objectives of test-construction. Moreover, doing think-aloud in the test-constructing process, she could try to think in students' minds, and thus raised the validity of her tests. NT 2 stated that think-aloud was a new method to him, and it gave him a chance to reflect on his test-constructing process. He thought the

experience of doing think-aloud was fun.

When asked about the feasibility of applying think-aloud method to test-constructing process, the four teachers' responses also differ. ET 1 thought that the method was feasible, but she was not used to it and she felt it quite strange to talk to herself. ET 2 considered think-aloud an effective method, because teachers could go back to examine their test items by listening to their recorded think-aloud protocols. In contrast, NT 1 thought the think-aloud method was not feasible in the test-constructing process because it was time-consuming. NT 2 pointed out he was not sure of the feasibility of applying think-aloud method to test-constructing process since he did not understand a lot about this method. He wrote that he needed to equip himself with more knowledge about it.

Finally, the four teachers were asked to express their opinions about participating in this study. Their opinions are as follows. ET 1 thought that doing this task helped her reflect on her test-constructing principles and patterns. However, this task was not suitable to be conducted in the office. Moreover, since participating teachers were asked not to rewrite the cloze passages, she thought it unchallenging to construct cloze items.

ET 2 expressed that test-construction was a frequent task for teachers to perform; therefore, she seldom took it seriously. However, the experience of participating in this study gave her a chance to reflect on her test-constructing process and to take test-construction seriously. ET 2 also wanted to know the results of students' performance on her test items.

NT 1 regarded it an honor to be able to participate in the study when she just began her teaching career. She thought the results of this study would make a contribution to the field of English teaching. NT 2 also expressed his pleasure in participating in this study due to his personal interest in test-making. He thought the

task in the study not only challenged his capacity to bridge theory and practice but also provided him with an opportunity to reflect on some guidelines in test-construction. If given the chance, he would also like to see the results of this study, and have a discussion on the test-constructing considerations other participating teachers had made.

Analyses of Teacher-constructed SAET Mock Tests

This section presents the analysis results of the four SAET mock tests constructed by the teachers. Each mock test consists of twenty-eight items: fourteen items based on Material A, and fourteen items on Material B. Furthermore, each set of the fourteen items includes five vocabulary items, five cloze items, and four reading comprehension questions. The four mock tests are presented in Appendix I. In the following, I compared and contrasted the four mock tests in the sequence of vocabulary items, cloze items, and reading comprehension questions. The two reviewers' comments on the four teacher-constructed mock tests served as important data in this discussion. In addition, "the checklists of the appropriateness of test construction[15]" also served as important tools to judge the quality of the four mock tests constructed by the teachers. The main reason for adopting the checklists as the judgment tools is that the checklists were designed for senior high school English test-construction principles, and was thus suitable for checking the appropriateness of test-construction of SAET mock tests involved in the present study.

Analyses of Vocabulary Items

According to the testing objectives provided by CEEC, the vocabulary items of the SAET should test on students' understanding of the formation, meaning, and

[15] The three checklists of the appropriateness of test-construction was translated from a file entitled "English test-construction principles and examples in senior high school" written by Professor Vincent W. Chang, and released by *English Education Research Center*, a website under the supervision of the Ministry of Education in Taiwan. (http://english.tyhs.edu.tw/downloads/participant.pdf)

collocation of frequently-used content words in high school. Therefore, I prepared eighty content words selected from the past SAETs for the teachers to construct their items. The four teachers had to produce five vocabulary items from the 40-word list on Material A, and another five items from the 40-word list on Material B. Their vocabulary items are presented in Appendix I.

Prior to the study, I assumed that the four teachers would construct three items on verbs, three on nouns, two on adjectives, and two on adverbs, since the two lists (Materials A and B) I provided contained twenty-four verbs, twenty-four nouns, sixteen adjectives, and sixteen adverbs. But it turned out that only ET 2 and NT 2 followed my assumption. ET 1 and NT 1, on the other hand, had different ideas from me. ET 1 constructed three items on verbs, two on nouns, three on adjectives and two on adverbs. Compared with my assumption, ET 1 constructed one more item on adjectives and one less item on nouns. This echoed with her own cloze test-constructing considerations in which she stated that nouns were often tested on cloze items. Thus, ET 1 constructed one less item on nouns. As for NT 1, she constructed three items on verbs, two on nouns, two on adjectives, and three on adverbs. However, NT 1 did not mention why she constructed more items on adverbs. The distribution of the number of items on different parts of speech in the teachers' tests are presented in Table 9.

Table 9.

Distribution of Items Testing on Different Parts of Speech

	Material A				Material B			
	ET 1	ET 2	NT 1	NT 2	ET 1	ET 2	NT 1	NT 2
Verb	1	1	2	1	2	2	1	2
Noun	1	2	1	2	1	1	1	1
Adjective	2	1	1	1	1	1	1	1
Adverb	1	1	1	1	1	1	2	1

Most of the time, the teachers' choices of words in their tests were quite different since they were faced with eighty different words to choose from. Table 10 presents the words the four teachers chose as the correct options in their items.

Table 10.
Words Teachers Chose As Correct Options

Material A	Verb	Noun	Adjective	Adverb
ET 1	contract	capacity	dominant, routine	considerably
ET 2	transfer	client, measure	accurate	scarcely
NT 1	frustrate, interpret	routine	optimistic	directly
NT 2	obtain	measure, decrease	dominant	continuously
Material B				
ET 1	collapse, revise	barrier	voluntary	originally
ET 2	boast, confess	relative	temporary	urgently
NT 1	maintain	exception	liberal	narrowly, gradually
NT 2	confess, preserve	occupation	reluctant	originally

It is to note in Table 10 that among the test items on Material A, both ET 2 and NT 2 selected the word "measure" as the correct option, and ET 1 and NT 2 both chose the word "dominant" as the correct option. Among the test items on Material B, the word "confess" was chosen by ET 2 and NT 2, and "originally" was selected by ET 1 and NT 2 as the correct options. It seems that in selecting the correct options for the vocabulary items, NT 2 had more similar ideas to ET 1 or ET 2.

Although the four teachers shared few correct options on their test items, they did make several similar choices in their selection of words as distractors. Appendix J presents the words that were chosen by at least two teachers in their vocabulary items. As Appendix J shows, it is very interesting that the four teachers all chose the word "dominant" on the list of Material A in their item, although ET 1 and NT 2 used the word as the correct option, and ET 2 and NT 1 used it as a distractor. It is also found in Appendix J that ET 2, NT 1, and NT 2 often chose the same words to use in their

items while ET 1's choice of words was quite different from the other three teachers. The main explanation for this difference is that unlike the other teachers, ET 1 did not follow my instructions of adopting only the words on the lists to prepare the four options in each vocabulary item on the mock test. As a result, she used many "illegitimate" words[16] on her test. When I saw her test items and found this flaw, I did not ask her to revise the vocabulary items since she had finished all the tasks required of her. In later discussion with her, I found out that ET 1 had misunderstood my instructions. She interpreted my requirement as that she just needed to choose words from the lists as correct options and that she could use other words she liked as distractors. Since my study focused on the processes of test-construction rather than on the choice of certain designated words, I allowed such a flaw in my study.

A Critique of Vocabulary Items

In the following, I summarized a brief critique of the vocabulary items based on the two reviewers' comments on the four tests. I located the problem areas of the vocabulary items, and then presented some (not every) problematic items as examples to illustrate the problems.

Problems with the stems

There are six problems identified by the two reviewers regarding the stems of the vocabulary items constructed by the four teachers. What follows are the six problems, each of which is accompanied by an example and the reviewers' comments[17].

(1) *The meaning of the stem was not clear*.

[16] The illegitimate words refer to those which did not appear on the two wordlists provided by the researcher. Those words are listed as follows.
Not included in Material A: combat, separate, prolong, capability, evaluation, measurement, fashionable, hesitant, temporary, reckless, slightly, roughly, tentatively
Not included in Material B: suspend, construct, extend, release, acclaim, interpret, acquisition, requirement, resistance, violent, hostile, tentatively, alternatively, unfortunately

[17] It is to note that not all the six problems were mentioned by both of the two reviewers; some of them were identified by either one of them. This note applies to the other discussions on the options of the vocabulary items and on the cloze items and reading comprehension questions as well.

Ex 1. (NT 1: A3)

As a good Taiwan citizen, we should _____ whatever is against the law.

(A) violate (B) frustrate (C) resist (D) decrease

R[18]: The meaning of "whatever is against the law" was quite unclear. Moreover, *resist whatever is against the law* did not seem to be logical in the context.

(2) *The stem was too long.*

Ex 2. (ET 1: B1)

Without warning, the bridge _____ suddenly, causing the cars to fall into the river. The tragedy was attributed to the erosion of the foundation and a lack of maintenance.

(A) suspended (B) collapsed (C) constructed (D) extended

R: The second sentence was not needed.

(3) *The stem did not contain sufficient clues.*

Ex 3. (ET 2: B1)

With the release of its new smart phones, the manufacturer Nokia _____ 160% more app downloads than Apple.

(A) boasts (B) revises (C) maintains (D) approaches

R: There was no clue in the context to answer this item.

(4) *The words in the stem were more difficult than the four options.*

Ex 4. (ET 2: B3)

While _____ and situational loneliness can be a normal, healthy part of life, chronic loneliness can be a very sad and sometimes very dangerous condition.

(A) reluctant (B) liberal (C) temporary (D) sincere

R: The word "chronic" is a Level-6 word, whereas option A (reluctant) belongs to Level 4 and options B (liberal), C (temporary), and D (sincere) belong to Level 3. Moreover, the topic of the stem is about psychology, a professional field which students might not be familiar with.

(5) *The stem involved potential controversies or personal opinions*.

Ex 5. (ET 2: A2)

Taiwan Railway Administration is working hard to make their trains' arrival

[18] "R" refers to the "reviewers' comments."

time _____ so as to win people's trust back.

(A) accurate (B) current (C) practical (D) dominant

R: Use of the real name "Taiwan Railway Administration" might cause potential controversy.

(6) *The stem contained a Chinglish proverb.*

Ex. 6. (ET 2: B2)

Well goes the saying "A distant _____ is not as good as a near neighbor." That is, good neighbors are a lot more helpful when we are in need.

(A) vision (B) legend (C) dealer (D) relative

R: The Chinese proverb was not translated correctly in the stem. "A distant relative" actually means a relative who is not so close in the family relationship with you. Yet the Chinese proverb means a relative who lives far away from you is not as helpful as a neighbor who lives near you.

The frequencies of the problems concerning the stems of the items constructed by the four teachers are shown in Table 11.

Table 11.

Frequencies of the Problems with the Stem in Vocabulary Items

Problems		ET1	ET2	NT1	NT2	(Total)
1.	The meaning of the stem was not clear.	4	2	4	4	14
2.	The stem was too long.	4	1	1	0	6
3.	The stem did not contain sufficient clues.	2	1	0	1	4
4.	The words in the stem were more difficult than the four options.	4	1	2	0	7
5.	The stem involved potential controversies or personal opinions.	0	3	0	0	3
6.	The stem contained a Chinglish proverb.	0	1	0	0	1
	(Total)	14	9	7	5	35

Table 11 shows that the most frequently-occurring problem concerning the stem is that the meaning of the stem was not clear. All four teachers had this problem with their stems. It is surprising that the two experienced teachers, especially ET 1, constructed more problematic stems than the two novice teachers did. Although ET 2

stated in her feedback sheet that she found it easiest to construct vocabulary items in this task (see Table 8), she did not produce better vocabulary stems than the other two novice teachers.

Problems with the options

There are eight problems identified by the two reviewers regarding the options of the vocabulary items constructed by the four teachers. The eight problems are presented as follows.

(1) *Two options started with the same letter.*

Ex 7. (ET 1: A2)
Human beings are not the only at risk of _____ the flu this season. The furry friends can fall ill as well.
(A) contracting (B) combating (C) separating (D) prolonging

R: The words "contracting" and "combating" both started with the letter "c."

(2) *Only an option started with a vowel.*

Ex 8. (NT 2: B3)
Miranda filled in the blank of _____ with the word "nurse," which means she worked as a nurse.
(A) occupation (B) relative (C) technique (D) division

R : Only the word "occupation" started with a vowel.

(3) *Only an option started with a consonant.*

Ex. 9. (ET 1: B3)
The demanding teacher always gives students a lot of homework and tests. ___, students complain of that. In the end, they adapt to the way he teaches.
(A) Tentatively (B) Alternatively (C) Unfortunately (D) Originally

R: Only the word "Tentatively" started with a consonant.

(4) *The distractors provided extraneous clues for elimination.*

Ex 10. (ET 2: A5)
Sandy's father was assigned to run a branch company in Mainland China; thus, she had no choice but to _____ to another school there.
(A) violate (B) recite (C) transfer (D) frustrate

R: Only "transfer" is an intransitive verb collocated with the preposition "to" and

the other options are all transitive verbs. Such a grammatical clue could make the distractors easily eliminated.

(5) *One distractor was too strong in elicitation.*

Ex. 11. (NT 1: A1)
Despite continual misfortunes happening to me, I still feel _____ about my future.
(A) dominant (B) competitive (C) optimistic (D) practical

R: Distractor B (competitive) was too strong in elicitation.

(6) *One or more distractors was too weak in elicitation.*

Ex. 12. (NT 1: A4)
Most people enjoy watching famous paintings even though they can hardly _____ the painter's thoughts.
(A) transfer (B) interpret (C) measure (D) frustrate

R: Distractor D (frustrate) was too weak in elicitation.

Ex. 13. (ET 1: A5)
It is of significant importance for drivers to have the _____ maintenance of their cars.
(A) hesitant (B) temporary (C) routine (D) reckless

R: Distractors A (hesitant) and D (reckless) were too weak in elicitation.

(7) *A distractor appeared on the same test twice*.

Ex 14. (NT 1: A3, A4)
As a good Taiwan citizen, we should _____ whatever is against the law.
(A) violate (B) frustrate (C) resist (D) decrease

Most people enjoy watching famous paintings even though they can hardly _____ the painter's thoughts.
(A) transfer (B) interpret (C) measure (D) frustrate

R: The distractor "frustrate" occurred twice on the same test produced by NT 1. One occurred in NT 1's third item on Material A, and the other occurred in the fourth item on Material A.

(8) *Two options could be correct answers*.

Ex. 15. (NT1: B3)

With his parents' patience and company for years, the retarded child _____ catch up with his classmates.

(A) rapidly (B) gradually (C) hardly (D) urgently

R: The correct option to this item, according to NT 1, is option B. Nevertheless, both options A and B could be the correct options based on different interpretations of the stem.

The frequencies of the problems concerning the options of the items constructed by the four teachers are shown in Table 12.

Table 12.

Frequencies of the Problems with the Options in Vocabulary Items

Problems		ET1	ET2	NT1	NT2	(Total)
1.	Two options started with the same letter.	7	1	1	0	9
2.	Only an option started with a vowel.	4	6	5	6	21
3.	Only an option started with a consonant.	1	0	0	0	1
4.	The distractors provided extraneous clues for elimination.	0	2	0	0	2
5.	One distractor was too strong in elicitation	0	0	1	0	1
6.	One or more distractors was too weak in elicitation.	1	2	3	3	9
7.	A distractor appeared on the same test twice.	0	0	1	0	1
8.	Two options could be correct answers.	0	0	1	0	1
	(Total)	13	11	12	9	45

It is found in Table 12 that the four teachers did not seem to take the first letter of an option into consideration when they constructed the vocabulary items; therefore, they produced many options which either started with a vowel or started with the same letter as another option. ET 1 had the highest frequency of the first problem with her options. It is probably due to her consideration that "the distractors could be similar in form to the key" when she constructed vocabulary items (see Table 5). Since ET 1 took this consideration into account in constructing the distractors, it was natural that many of her options would contain the first problem mentioned above. It

is also a pity that although ET 2 thought that "the distractors should begin with different letters" (see Table 5), she still produced one item (ET 2: B5) in which "two options started with the same letter."

All four teachers also constructed many options containing the second problem that "only an option started with a vowel." This might result from the requirement of the study that they had to select words as their options from the two forty-word lists. It is probably due to this reason that the four teachers had limited choice of appropriate words for their options. If they had been given freedom to choose their own words, they might have constructed fewer items containing the second problem (i.e., Only an option started with a vowel.).

Aside from the first three problems, the two novice teachers produced more problematic options for the vocabulary items than the two experienced teachers did. In particular, the former constructed more distractors which were too weak in elicitation than the latter. Between the two novice teachers, NT 1 seemed to have constructed more problematic options than NT 2 because she was not as prudent as NT 2 in constructing the items as their think-aloud protocols showed. However, although NT 2 had the consideration of making his distractors equally attractive (see Table 5), he still produced three distractors that were too weak in elicitation from the reviewers' perspectives.

General discussion on vocabulary items

So far, I have presented the four teachers' considerations in constructing vocabulary items and the problems of their items identified by the two reviewers. Now, it is time to examine whether the vocabulary items constructed by the four teachers were appropriate by using the appropriateness checklist for vocabulary items. I used the checklist to examine each of the ten vocabulary items constructed by each of the four teachers. In order to present a concise result of the appropriateness of those

items, I awarded the appropriateness checks in the following way. If I found one vocabulary item constructed by a teacher mismatching the description of a checklist criterion, I would put an "X" in that criterion. Only when the ten vocabulary items constructed by one teacher matched the description of a checklist criterion would that checklist criterion receive a check. The results are shown in Table 13.

Table 13.
Results of the Appropriateness Checklist for Vocabulary Items

Checklist criteria	ET1	ET2	NT1	NT2
1. Are the tested keys (including distractors) the content words frequently-used in high school?	√	√	√	√
2. Is the meaning of the context in the stem clear? Does the stem provide sufficient clue?	X	X	X	X
3. Is the length of the stem appropriate (i.e., within twenty words)?	X	X	X	X
4. Is the number of letters of the distractors similar to that of the key?	X	X	X	X
5. Is the part of speech of the distractors the same as that of the key?	√	√	√	√
6. Is the difficulty level of the distractors similar to that of the key?	X	X	X	X
7. Are the options equally strong in elicitation?	X	X	X	X

The results in Table 13 seem to suggest that none of the four teachers' vocabulary items could be considered appropriate because their test items only matched two of the seven checklist criteria. For one, all the four teachers tested on the content words which are frequently-used in high school. It was simply because these words (including the keys and the distractors) were chosen from the two wordlists prepared by the researcher. For the other, all the four teachers made sure that the key and the distractors in one item were of the same part of speech. This could be attributed to their consideration in designing the distractors of the vocabulary items.

By comparing the four teachers' considerations in constructing vocabulary items (see Table 5) and the appropriateness checklist for vocabulary items (see Table 13),

we found that there are only three issues that both the teachers and the checklist have in common; namely, *the stem provides sufficient clue, the distractors are of the same part of speech as the key*, and *the options are equally attractive*. It is also found that NT 2 was the only teacher that took these three issues into consideration although his items were still evaluated as inappropriate by the second and the seventh criteria in the checklist.

At this point, we could draw several conclusions about the vocabulary items constructed by the four teachers. First, these vocabulary items could not be deemed of a good quality from the reviewers' perspectives or could not be considered appropriate with reference to the appropriateness checklist. Many of the four teachers' vocabulary items needed to be improved because they contained flaws or problems. Second, the two experienced teachers did not seem to produce better vocabulary items than the two novice teachers did. They constructed more problematic items concerning both the stems and the distractors than the novice teachers did. Finally, although the teachers (especially NT 2) had made several considerations when constructing vocabulary items, they still produced many inappropriate or problematic items. The fact that the teachers' considerations were somewhat different from the authority's checklist criteria and from the two reviewers' viewpoints could account for part of the reason for this phenomenon.

Analyses of Cloze Items

The four teachers had to construct five cloze items on a passage in Material A, and five items on a passage in Material B. Most importantly, they were required not to rewrite the passages, so that I could compare the items they constructed. Their cloze tests are shown in Appendix I. According to the testing objectives provided by CEEC, the cloze items of the SAET should test on students' command of words (including content words, function words, idiomatic expressions, transitional words, etc.)

grammar, and passage structure. I first analyzed the types of questions the four teachers constructed on their tests, the results of which are shown in Table 14.

Table 14.

Types of cloze items the teachers constructed

Material A: Item types	
ET 1	content words (verb and noun), conjunction phrase, verb form, participial phrase
ET 2	conjunction, transition, verb phrases, relative pronoun
NT 1	comparative structure, transition, preposition, verb form, relative pronoun
NT 2	transition, comparative adjective, content word (noun), tense, verb phrase
Material B: Item types	
ET 1	content word (noun), conjunction, pronoun, transition, verb phrase
ET 2	pronoun, formal subject, content word (verb), verb phrases
NT 1	transitions, content words (adverb and verb), conjunction
NT 2	transition, relative adverb, content word (adverb and verb), comparative adjective

Then, based on CEEC's guidelines and the teachers' item types, I further categorized the different item types in Table 14 into six major types: (1) content words (nouns, verbs, adjectives, and adverbs), (2) function words (pronouns, preposition, and conjunctions) (3) transitions (transitional word and phrase), (4) phrases (verb phrase, participial phrase, and conjunction phrase), (5) verb forms (including verb tense), and (6) grammatical structures (comparative structure and formal subject). The distribution of the item types each teacher produced is shown in Table 15.

As shown in Table 15, all four teachers constructed items on content word verbs. Three teachers produced items on transitional phrase (ET 2, NT 1, and NT 2), conjunctions (ET 1, ET 2, and NT 1) and verb phrases (ET 1, ET 2, and NT 2). Two teachers constructed items on content word nouns (ET 1 and NT 2), pronouns (ET 1 and ET 2), relative pronouns (ET 2 and NT 1), transitional words (ET 1 and NT 1) and verb forms (ET 1 and NT 1). In terms of constructing cloze item types, it seems that the two experienced teachers were more in line with each other than they were with

the other two novice teachers.

Table 15.

Distribution of the cloze item types teachers constructed

	Material A				Material B			
Item types	ET1	ET2	NT1	NT2	ET1	ET2	NT1	NT2
Content words								
noun	1			1	1			
verb	1			1		1	1	1
adjective								1
comparative adjective				1				1
adverb							1	
relative adverb								1
Function words								
pronoun					1	1		
preposition			1					
conjunction		1			1		1	
relative pronoun		1	1					
Transitions								
transitional word					1		1	
transitional phrase		1	1	1			1	1
Phrases								
verb phrase		2		1	1	2		
participial phrase	1							
conjunction phrase	1							
Verb forms								
tense and aspect								
verb form	1		1					
Grammatical structure								
comparative structure			1					
formal subject						1		

Based on the analyses in Table 15, it is also interesting to explore whether teachers constructed similar items given the fact that they were provided with the same passages. In Material A, there are four points of the passage that were tested by at least two teachers. The items testing on similar points are presented as follows.

Ex. 16. (Four teachers tested on this sentence: *For example, John Steinbeck is said to have been an excellent listener, yet he was hated by some of the people he wrote about.*)

For example, John Steinbeck is said _____ an excellent listener, yet he was hated by some of the people he wrote about. (ET 1: A4)
(A) being (B) to be (C) to have been (D) to have

For example, John Steinbeck _____ an excellent listener, yet he was hated by some of the people he wrote about. (ET 2: A3)
(A) was portrayed as (B) was told to pretend
(C) is prone to be (D) is said to have been

For example, John Steinbeck is said to _____ an excellent listener, yet he was hated by some of the people he wrote about. (NT 1: A4)
(A) have been (B) be (C) has been (D) become

For example, John Steinbeck is said to _____ an excellent listener, yet he was hated by some of the people he wrote about. (NT 2: A4)
(A) have been (B) has been (C) had been (D) having been

This sentence was the only one that all four teachers had constructed an item on. It seems that all four teachers wanted to test on the structure "is said to + have + pp.," so they chose this sentence as a testing point. Even though they chose the same testing point, the four teachers had different testing focuses. For example, ET 2 tested on several linguistic knowledge (e.g., vocabulary and grammar) at a time; thus, her item was considered most difficult. On the other hand, NT 2 only tested on the knowledge of "infinitive 'to' + verb root," and her item was considered the easiest among the four items. This testing point is also the only one on the cloze passage in Material A that the two experienced teachers shared.

Ex. 17. (Three teachers tested on this sentence: *In addition, good listeners are inclined to accept or tolerate rather than to judge and criticize.*)

_____, good listeners are inclined to accept or tolerate rather than to judge

and criticize. (ET 2: A2)
(A) In consequence (B) In the end (C) In contrast (D) In addition

_____, good listeners are inclined to accept or tolerate rather than to judge and criticize. (NT 1: A2)
(A) In fact (B) In addition (C) In short (D) In other words

_____, good listeners are inclined to accept or tolerate rather than to judge and criticize. (NT 2: A1)
(A) In contrast (B) As a result (C) In addition (D) For that reason

Transitional phrases are frequently-tested points on SAETs. Since there are many transitions on the cloze passage in Material A, the finding that three teachers chose the same transitional phrase "in addition" as their testing point was really interesting. Moreover, the two novice teachers had two similar testing points in Material A (the previous one and this one).

Ex. 18. (Two teachers tested on this sentence: *No doubt his ability to listen <u>contributed to</u> his capacity to write.*)

No doubt his ability to listen _____ his capacity to write. (ET 2: A4)
(A) contributed to (B) originated from (C) involved in (D) substituted for

No doubt his ability to listen _____ his capacity to write. (NT 2: A5)
(A) resulted from (B) contributed to (C) consisted of (D) persisted in

ET 2 liked to test on verb phrases on cloze tests. Among the ten cloze items, she constructed four items on verb phrases (see Table 15). NT 2 had constructed only one item on verb phrases, and this was the very item of it. In constructing the verb phrase item, both ET 2 and NT 2 made the consideration that "the prepositions in each verb phrase should be different" (see Table 6). Thus, the options they constructed had different verbs and prepositions.

Ex. 19. (Two teachers tested on this sentence: *Thus, depending on <u>what</u> a good*

listener does, he may become either popular or disliked in his lifetime.)

Thus, depending on _____ a good listener does, he may become either popular or disliked in his lifetime. (ET 2: A5)
(A) that (B) what (C) which (D) how

Thus, depending on _____ a good listener does, he may become either popular or disliked in his lifetime. (NT 1: A5)
(A) that (B) what (C) which (D) how

The two items tested on relative pronoun "what," which is not only a frequently-tested point but also a grammatical point that students are often confused about. It is very interesting that ET 2 and NT 1 had constructed exactly the same item, including the four options being placed in exactly the same positions!

The above four sets of examples show that ET 2 was more in line with the two novice teachers in constructing cloze items on the passage in Material A. However, ET 1 seemed to choose quite different testing points from the other three teachers.

With regard to the items on the passage in Material B, there are five testing points that were shared by at least two teachers. The items sharing the same testing points are presented as follows.

Ex 20. (Two teachers tested on this sentence: *For example, in the United States, it is very important to be on time for almost all occasions.*)

_____, in the United States, it is very important to be on time for almost all occasions. (NT 1: B1)
(A) Therefore (B) For example (C) However (D) At first

_____, in the United States, it is very important to be on time for almost all occasions. (NT 2: B1)
(A) In addition (B) For example (C) As a result (D) Even so

Transitions are frequently-tested items on SAETs. The two novice teachers chose the transitional phrase "for example" as their testing point, and this was also the only

item they had in common on the passage in Material B. However, both of the above two items lacked discrimination since all the students who answered these two items got the correct answers. This might suggest that not every transition makes a good testing point because some may seem too easy for students.

Ex. 21. (Two teachers tested on this sentence: *A person usually tries to arrive about five minutes <u>after</u> the invitation time, so that the host would have a little extra time to prepare for the guests.*)

A person usually tries to arrive about five minutes _____ the invitation time, so that the host would have a little extra time to prepare for the guests. (ET 1: B2)
(A) before (B) as soon as (C) after (D) as early as

A person usually tries to arrive about five minutes _____ the invitation time, so that the host would have a little extra time to prepare for the guests. (NT 1: B3)
(A) after (B) before (C) on (D) by

ET 1 and NT 1 chose the conjunction "after" as their testing point. In order to answer this item correctly, students had to use sentential or contextual clues to answer this question. Thus, this item might be considered a good cloze test item.

Ex. 22. (Two teachers tested on this sentence: *Any time later than that is considered impolite, because it <u>keeps</u> the host and other guests waiting.*)

Any time later than that is considered impolite, because it _____ the host and other guests waiting. (ET2: B3)
(A) keeps (B) makes (C) has (D) gets

Any time later than that is considered impolite, because it _____ the host and other guests waiting. (NT1: B4)
(A) makes (B) forces (C) lets (D) keeps

ET 2 and NT 1 chose the verb "keep" as their testing point to test on students' knowledge of the structure " V+ O+ OC," which is a frequently-used structure and a frequently-tested grammatical point in high school. However, one did not need to use

contextual clues to answer this question; what one needed was the knowledge of the usage of the verbs in the options. Therefore, this item might not be considered a good cloze test item.

Ex. 23. (Two teachers tested on this sentence: ***<u>However</u>, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment***…)

_____, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment…(ET 1: B4)
(A) Instead (B) Likewise (C) Therefore (D) However

_____, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment…(NT 1: B5)
(A) Actually (B) However (C) In conclusion (D) Unfortunately

ET 1 and NT 1 chose the transitional word "however" as their testing point. It is found that there were only two transitions on the passage of Material B, namely, "for example" and "however," and that NT 1 had constructed items on both of them. Moreover, it is also found that among the five cloze items in Material B, NT 1 had four similar items to the other teachers. In particular, NT 1 had two items in common with ET 1, including this item and the above-mentioned "keeps" item.

Ex. 24. (Two teachers tested on this sentence: …***because there are usually forms that need to be <u>filled out</u> by the patient.***)

…because there are usually forms that need to be ____ by the patient. (ET 1: B5)
(A) filled out (B) applied for (C) filled with (D) taken over

…because there are usually forms that need to be ____ by the patient. (ET 2: B5)
(A) figured out (B) set out (C) filled out (D) sent out

This was the only item shared by the two experienced teachers, who tested on the verb phrase "filled out" in the collocation "fill out the form." This item might be considered a good cloze test item because students had to first recognize the

collocation in which the verb phrase and the noun were placed in two clauses, and then choose the appropriate verb phrase.

One interesting finding concerning ET 2's items is that in constructing the verb phrase item (ET 2: A4) in Material A, ET 2 made sure that "the prepositions in each verb phrase were different." However, in this item (ET 2: B5) in Material B, she did not make the same consideration. The four verb phrases all contained the same preposition "out." This might be due to another consideration ET 2 took into account that "the distractors should be similar in form to each other;" thus, she constructed the item this way.

A Critique of Cloze Items

In the following, I summarized a brief critique of the cloze items based on the two reviewers' comments on the four tests. I first located the problem areas of the cloze items, and then presented some items as examples to illustrate the problems.

Problems with the choice of blanks (or testing points)

According to the two reviewers, the four teachers' choice of testing points involved the following three problems.

(1) *A blank appeared in the last sentence of the passage.* It is often suggested that the first and last sentences of a cloze passage should be left intact; no testing points should come from these two sentences. However, all four teachers had a blank (or testing point) in the last sentence of the passage. ET 1, ET 2, and NT 1 had this problem in Material A and in Material B. NT 2 had this problem in Material B. ET 1 even constructed two items in the last sentence of the passage in Material B.

> Ex. 25. (ET 1: B4, B5)
> __4__, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be __5__ by the patient.

(2) *Two blanks appeared close to each other.* NT 1 had this problem in her test on Material A, and she seemed not to be aware of this based on her protocols.

Ex. 26. (NT 1: A1, A2)
Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them __1__ other people. __2__, good listeners are inclined to accept or tolerate rather than to judge and criticize.

(3) *The blanks could be answered based either on local clues or on common sense.*

ET 1 and NT 2 had this problem of inappropriate choice of blanks, which could be answered easily by using within-sentence clues instead of by contextual clues.

Ex. 27. (ET 1: A1)
Yet listening well is a rare talent that everyone should _____.
(A) despise (B) treasure (C) ignore (D) command

R: The item could be answered based on common sense.

Ex. 28. (NT 2: A2)
Therefore, they have _____ enemies than other people.
(A) few (B) fewer (C) little (D) less

R: The testing point was too local. No contextual clues were needed to answer this item. Furthermore, it was difficult to come up with good distractors.

The frequencies of the problems concerning the choice of testing points in cloze passages are shown in Table 16.

Table 16.
Frequencies of the Problems with the Choice of Blanks in Cloze Items

Problems		ET1	ET2	NT1	NT2	(Total)
1.	A blank appeared in the last sentence of the passage.	2	2	2	1	7
2.	Two blanks appeared close to each other.	0	0	1	0	1
3.	The blanks could be answered based either on local clues or on common sense.	1	1	2	3	7
	(Total)	3	3	5	4	15

Table 16 shows that the two experienced teachers seemed to be better at

choosing the testing points for the cloze items than the two novice teachers because they had less problems with the choice of cloze blanks. Among the four teachers, NT 2 was the only one making the consideration that "the testing points should not fall on the last sentence of a passage" (see Table 6) when selecting the blanks. However, it was a pity that he violated his own rule and constructed one item on the passage in Material B. That was why he had only one problem with the choice of blanks while the other teachers had two.

Moreover, it seems that the two novice teachers were not aware that the testing points they chose could be answered based either on local clues or on common sense. Thus, they chose more of such testing points than the two experienced teachers did. The experienced teachers seemed to be wiser in this regard.

Problems with the options

The two reviewers pointed out five problems concerning the options constructed by the four teachers.

(1) *The distractors provided extraneous clues for elimination.* All four teachers had this problem in constructing their distractors. ET 2 had such problematic items in both Material A and Material B. But it seemed that she made her item this way on purpose, according to her think-aloud protocols.

Ex. 29. (ET 1: A2)
In addition, good listeners are inclined to accept or tolerate _____ to judge and criticize.
(A) as well as (B) instead of (C) regardless of (D) rather than

R: Distractors B and C could be eliminated based on grammatical rules since neither of them could take on "to."

Ex. 30. (ET 2: B3)
Any time later than that is considered impolite, because it _____ the host and other guests waiting.

(A) keeps (B) makes (C) has (D) gets

R: Only option A can take on a gerund, and the other three could be easily eliminated based on grammatical rules.

(2) *The options were not agreeable in form.*

Ex 31. (NT 1: B5)

_____, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be filled out by the patient.

(A) Actually (B) However (C) In conclusion (D) Unfortunately

R: Options A, B, and D are one-word adverbs while C is a two-word phrase.

Ex. 32. (NT 2: B1)

_____, in the United States, it is very important to be on time for almost all occasions.

(A) In addition (B) For example (C) As a result (D) Even so

R: Option C contains three words while the other options contain two words.

(3) *The options involved assessment of varying aspects of linguistic knowledge at a time (e.g., assessing vocabulary and grammar knowledge).*

Ex. 33. (ET 2: A3)

For example, John Steinbeck _____ an excellent listener, yet he was hated by some of the people he wrote about.

(A) was portrayed as (B) was told to pretend

(C) is prone to be (D) is said to have been

R: The options were too complex, involving the assessment of both vocabulary and grammar.

(4) *More than one option could be correct answers.* All four teachers had this problem. As mentioned in Ex. 17, three teachers (ET 2, NT 1, and NT 2) tested on the same transitional phrase "in addition." However, one or two of the distractors those teachers constructed were considered acceptable by the reviewers. It seemed that it was not easy for the teachers to construct good transitions items.

Ex. 34. (The correct option is *In addition*.)

_____, good listeners are inclined to accept or tolerate rather than to judge and criticize. (ET 2: A2)

(A) In consequence (B) In the end (C) In contrast (D) In addition

R: Option A was also possible.

_____, good listeners are inclined to accept or tolerate rather than to judge and criticize. (NT 1: A2)

(A) In fact (B) In addition (C) In short (D) In other words

R: Options A and D were arguably possible.

_____, good listeners are inclined to accept or tolerate rather than to judge and criticize. (NT 2: A1)

(A) In contrast (B) As a result (C) In addition (D) For that reason

R: Options B and D were arguably possible.

(5) *The distractor was too weak in elicitation.*

Ex. 35. (ET 2: B1)

People have different ideas about what exactly is being on time and being late. _____ ideas also differ from time to time, and from country to country.

(A) These (B) Other (C) Theirs (D) Our

R: Option C was too weak in elicitation.

The frequencies of the problems concerning the options in cloze items are shown in Table 17.

Table 17.

Frequencies of the Problems with the Options in Cloze Items

Problems		ET1	ET2	NT1	NT2	(Total)
1.	The distractors provided extraneous clues for elimination.	1	2	1	1	5
2.	The options were not agreeable in form.	0	0	1	1	2
3.	The options involved assessment of varying aspects of linguistic knowledge at a time.	0	1	0	0	1
4.	More than one option could be correct answers.	1	3	2	2	8
5.	The distractor was too weak in elicitation.	1	1	0	1	3
	(Total)	3	7	4	5	19

Table 17 shows that ET 2 seemed to have constructed more problematic options for cloze items than the other three teachers. Among the identified problems, the second problem occurred to ET 2 only. Concerning the options with the second problem (see Ex. 33), ET 2 stated in her think-aloud protocols that she made that item difficult on purpose because she wanted students to spend more time on it. It seems that she did not consider it inappropriate to assess the knowledge of vocabulary and grammar in one item.

Among the five problems, the fourth one had the highest frequency. All four teachers had this problem. They all constructed some options which were so strong in elicitation that they were considered correct answers as well by the reviewers. This suggests that good cloze options are difficult to construct; even the experienced teachers would produce inappropriate options.

General discussion on cloze items

After examining the qualities of the cloze items from the two reviewers' perspectives, I used the appropriateness checklist for cloze items to evaluate the appropriateness of the cloze items produced by the four teachers. The original checklist contains ten criteria, seven of them about the testing points and the options, and three of them about the passages[19]. Since the cloze passages were prepared by the researcher, and the four teachers were asked not to rewrite the passages, the three criteria about the passages were thus not used to evaluate the appropriateness of the four teachers' cloze items. Only the remaining seven criteria were included in the checklist to evaluate the teachers' cloze items. The results are shown in Table 18.

[19] The three criteria about the cloze passages are as follows. (1) Do the passages have different topics, styles, or genres? (2) Are the passages educational, informative, and interesting? Are they suitable for students' life, learning experiences, and cognitive abilities? (3) Are the passages appropriate in terms of difficulty level?

Table 18.

Results of the Appropriateness Checklist for Cloze Items

Checklist criteria		ET1	ET2	NT1	NT2
1.	Do the items correspond to the testing objectives?	X	X	X	X
2.	Does the constructor avoid producing items on the first and last sentences of the passage?	X	X	X	X
3.	Does each blank (testing point) have sufficient clues?	√	√	√	√
4.	Do the testing points focus on both global and local understanding? Do they also focus on meaning and grammar?	X	√	X	X
5.	Are the options free from spelling, usage, or grammar errors?	√	√	√	√
6.	Are the options free from controversies or inappropriate elicitation?	X	X	X	X
7.	Are the options equally strong in elicitation?	X	X	X	X

The results in Table 18 seem to suggest that none of the four teachers' cloze items could be considered appropriate because three teachers' (ET 1, NT 1, and NT 2) items matched only two of the seven checklist criteria and ET 2's items matched three of the seven checklist criteria. ET 2's items matched the fourth criterion while the other teachers' items did not because one of ET 2 items (ET 2: B4) tested on the global understanding of the passage. Most of the other teachers' items, however, were merely considered to test on the local understanding of the passage.

Concerning the first criterion, the four teachers' items were not considered to correspond to the testing objectives because their items did not test on the passage structure, which is a requirement in the objectives provided by CEEC. As mentioned earlier, all the four teachers constructed on the last sentence of the passage in Material B; thus, they violated the second criterion. As for the sixth and the seventh criteria, the options constructed by the four teachers were not of equal elicitation. Some were so strong that two in an item could be considered correct options; some were too weak to attract the students (see Table 17). Therefore, the four teachers' options could not be

deemed appropriate in terms of the two criteria.

A comparison of the four teachers' considerations in constructing cloze items (see Table 6) and the appropriateness checklist for cloze items (see Table 18) shows that there are only two issues that the teachers' considerations and the checklist criteria have in common; namely, *the testing points should not focus on the first and last sentences of the passage*, and *the options should be equally attractive*. In fact, it is only NT 2 that took these two issues into consideration. Nevertheless, NT 2's items were still considered inappropriate in terms of the two criteria.

Based on the two reviewers' comments and the results of the appropriateness checklist for cloze items, we could draw the following five conclusions. First, the cloze items produced by the four teachers could not be considered of a good quality since each of the four teachers' items were flawed in some ways as identified by the two reviewers. Second, the cloze items constructed by the four teachers could not be deemed appropriate either because they matched only two or three criteria in the checklist. Third, the two experienced teachers did not seem to produce better cloze items than the two novice teachers did because all teachers' items needed to be revised in some ways. Finally, although the teachers made many considerations in constructing cloze items, very few of the considerations were in line with the criteria in the appropriateness checklist. This suggests that the four teachers seemed not to be familiar with the cloze testing objectives in SAET although they claimed to be familiar with the guidelines for the SAET (see Table 4) and thought that it was the easiest to construct cloze items in this task (see Table 8).

Analyses of Reading Comprehension Questions

The four teachers had to construct four reading comprehension questions on a passage in Material A, and four questions on a passage in Material B. Their test items are shown in Appendix I. According to the testing objectives provided by CEEC, the

reading comprehension questions of the SAET should test on students' ability of using their knowledge about vocabulary, idioms, semantics, syntax, or pragmatics to understand both the whole passage and the details and to further analyze and make inferences about the passage. Thus, based on the objectives and the teachers' items, I categorized the reading comprehension questions into four major types: global, local, inferential, and referential. The global type of questions include those asking for the main idea or best title of the passage. The local type of questions refer to those asking for the details of the passage or the meaning of a particular word, or asking students to judge whether a sentence is true or not according to the passage. The inferential type of questions contain those that ask students to make inferences about the passage. The referential type of questions test on students' ability to identify the correct reference of a certain pronoun. The results of the categorization are shown in Table 19.

Table 19.
Distribution of the reading comprehension question types teachers constructed

	Material A				Material B			
Question types	ET1	ET2	NT1	NT2	ET1	ET2	NT1	NT2
Global								
main idea	1	1						1
best title					1		1	
Local								
details		1						
true/ not true	1	1	1	1	1	1	2	1
meaning of a word					1			
Inferential	1	1	3	3	1	3	1	1
Referential	1							1

As shown in Table 19, all teachers produced true/not true questions and inferential questions. Specifically, the two novice teachers and ET 2 produced more inferential questions than ET 1 did. Both NT 1 and NT 2 constructed three inferential

questions on Material A and one inferential question on Material B, while ET 2 produced three inferential questions on Material B and one inferential question on Material A. On the other hand, the questions constructed by ET 1 were more equally distributed among the four major types. NT 2 also produced four types of questions on Material B.

A Critique of Reading Comprehension Questions

In the following, I summarized a brief critique of the reading comprehension questions based on the two reviewers' comments on the four tests. I first located the problem areas of the questions, and then presented some items as examples to illustrate the problems.

Problems with the question stems

According to the two reviewers, the reading comprehension questions produced by the four teachers had the following problems.

(1) *The words in the question were too difficult.*

Ex. 36. (ET 1: B4)
Which of the following words is synonymous with the word "**morale**?"

R: The word "synonymous" was difficult for students to understand. The question could be revised as "Which of the following is closest in meaning to the word '**morale**' in the third paragraph?"

(2) *The format of the question was inappropriate.*

Ex. 37. (ET 2: B1)
David Smith is a(n) _____.
(A) tailor (B) consultant (C) employee (D) employer

R: The format of this item should be a question (which is conventional), not a blank. Thus, this question could be stated as "What did David Smith do?"

One reviewer commented that this question was too trivial and should not be included. There were other more important testing points in the passage. The

other reviewer commented that since it was an inferential question, all four options were arguably possible.

(3) *The wording of the question was inappropriate.*

Ex. 38. (ET 2: A3)
What is **false** about people with "gold fever?"

R: This question should be revised as "Which of the following statements is NOT true about people with "gold fever?" It is better to use a positive word than a negative word.

(4) *The sequence of the four questions on one test was inappropriate.* ET 2 and NT 2 had this problem. ET 2 began the questions on Material B with an inferential question, which should normally be placed in the end, while NT 2 ended the questions on Material B with a global question (main idea), which should come first, based on the principle of starting with easy or global questions.

(5) *Two questions tested on similar points in one passage.* Both of the two reviewers commented that the first question (What is the purpose of this article?) and the fourth question (Why did the author mention the old leather shoe?) constructed by NT 2 on Material A actually tested on similar points. Therefore, the fourth question should be replaced with a new one.

The frequencies of the problems concerning the question stems in reading comprehension questions are shown in Table 20.

Table 20 shows that all four teachers had problems of using inappropriate wording to construct their question stems. Among them, ET 2 had more of this problem. In addition, ET 2 also produced two items which had inappropriate formats. Both the two experienced teachers used difficult words in their question stems. This could result from the consideration [20] they took while constructing reading

[20] The consideration is that "the question stem should contain difficult words or expressions."

comprehension questions (see Table 7).

As for the novice teachers, NT 2 had the problems of arranging the sequence of the questions sequence and choosing the testing points. NT 1 seemed to have fewer problems in constructing the question stems although she made the least considerations among the four teachers in constructing the items.

Table 20.
Frequencies of the Problems with the Question Stems in Reading Comprehension Items

Problems	ET1	ET2	NT1	NT2	(Total)
1. The words in the question were too difficult.	1	1	0	0	2
2. The format of the question was inappropriate.	0	2	0	0	2
3. The wording of the question was inappropriate.	2	3	1	1	7
4. The sequence of the four questions on one test was inappropriate.	0	1	0	1	2
5. Two questions tested on similar points in one passage.	0	0	0	1	1
(Total)	3	7	1	3	14

Problems with the options

According to the two reviewers, the four teachers' construction of options showed the following problems.

(1) *The words in the options were too difficult.*

Ex. 39. (ET 1: B4)
Which of the following words is synonymous with the word "**morale**?"
(A) integrity (B) productivity (C) happiness (D) enthusiasm

R: Option A, integrity, was too difficult for students because it is a Level-6 word.

(2) *One of the options was too long.*

Ex. 40. (NT 2: B1)
Which of the following statements about the study by Levi Strauss and Company is NOT TRUE?
(A) 15% of employers think casual wear will make employees unwilling to

work.

(B) 85% of employees believe that they are in a bad mood for working in casual wear.

(C) 4% of employers argue that employees will produce less when they wear casual clothes.

(D) Those who welcome the policy of casual wear think they will save money by buying casual clothes instead of formal ones.

R: Option D was too wordy and much longer than the other three options, and should be concisely stated.

(3) *The option was difficult to understand.*

Ex. 41. (NT 2: A3)

What can we infer from this passage?

(A) The woman brought with her the supplies which weighed over 40 pounds.

(B) The woman who joined the gold-seeking trip followed the crowd in clothing.

(C) The woman knew the journey was full of dangers when she decided to make it.

(D) The woman could not stand the cold weather because she lost the leather shoe.

R: Option B was difficult to understand.

(4) *Two options were similar in meaning.*

Ex. 42. (NT 1: B3)

According to the research by Levi Strauss, how do most employers think about dress code?

(A) It makes their employees less productive.

(B) It's practicable only on Friday.

(C) It helps their employees work efficiently in a good mood.

(D) It has a negative impact on their productivity.

R: Options A and D actually had similar meanings. Thus, one of them should be replaced with a new option testing on a different key point.

(5) *One or more options was too weak in elicitation.*

Ex. 43. (ET 1: B1)

Which of the following is the best title for the article?

(A) The Origin of Casual Friday
(B) How to Raise the Efficiency of Office Workers
(C) From Formal Wear to Casual Clothes
(D) The Dilemma between Suits and Jeans

R: Since the word "jeans" did not appear in the passage, option D was too weak in elicitation.

(6) *The options were not agreeable in form.*

Ex. 44. (ET 1: A1)
What is the main idea of the passage?
(A) The history of gold fever.
(B) Traveling light is crucial to seeking gold.
(C) Shoes are the perfect option on display.
(D) Daily stuff, like a shoe, can tell us an amazing story.

R: Only option A was a noun phrase while the other three were complete sentences. Thus, option A should be changed into a complete sentence as well.

(7) *More than one option could be correct answers.*

Ex. 45. (ET 2: A4)
What can we infer about the shoe's owner from this article?
(A) She was a person with great courage.
(B) She was in the business of trading gold.
(C) She was a fashion queen in the 19th century.
(D) She lost her shoe while carrying supplies back and forth.

R: Options A, B, and D could be correct answers since this was an inferential question.

(8) *None of the four options was the answer to the question.*

Ex. 46. (ET 1: A4)
In the fifth line of the second paragraph, what does "**this**" refer to?
(A) The requirement for the gold seekers.
(B) The greatest adventures of the shoe owner.
(C) The Canadian government.
(D) Gold fever.

R: None of the four options was the correct answer because the pronoun "this" referred to "one ton of supplies" in the passage.

(9) *The options were not arranged in sequence of length.*

Ex. 47. (ET 1: A4)

In the fifth line of the second paragraph, what does "**this**" refer to?

(A) The requirement for the gold seekers.

(B) The greatest adventures of the shoe owner.

(C) The Canadian government.

(D) Gold fever.

R: It is conventional practice that shorter options come first, and longer ones last. Thus, option D should be placed first, followed by options C, A, and B.

The frequencies of the problems concerning the options in reading comprehension questions are shown in Table 21.

Table 21.

Frequencies of the Problems with the Options in Reading Comprehension Items

Problems	ET1	ET2	NT1	NT2	(Total)
1. The words in the options were too difficult.	2	0	0	0	2
2. One of the options was too long.	2	0	0	2	4
3. The option was difficult to understand.	0	0	0	1	1
4. Two options were similar in meaning.	0	0	1	1	2
5. One or more options was too weak in elicitation.	3	1	0	1	5
6. The options were not agreeable in form.	1	0	1	0	2
7. More than one option could be correct answers.	2	3	0	1	6
8. None of the four options was the answer to the question.	1	0	1	2	4
9. The options were not arranged in sequence of length.	7	5	8	0	20
(Total)	18	9	11	8	46

Table 21 shows that the most common problem with the teachers' options was that the teachers did not arrange their options in sequence of length. Among the four teachers, only NT 2 paid attention to this rubric and took it into consideration while constructing the options (see Table 7). The other three teachers seemed not to be

aware of this rubric and thus placed their options randomly regardless of their lengths.

The two experienced teachers constructed more unattractive options than the two novice teachers did. Moreover, they also produced more items which had more than one correct answers. On the other hand, the two novice teachers had the problem of constructing two options which were similar in meaning and the problem of constructing items that had no correct answers.

Taken together, ET 1 had more problems with the options she constructed than the other three teachers did. It seems that the two experienced teachers did not produce better options for reading comprehension items than the two novice teachers.

General discussion on reading comprehension questions

After the two reviewers had identified the problems in the reading comprehension questions constructed by the four teachers, I used the appropriateness checklist for reading comprehension questions to evaluate the appropriateness of those items. The original checklist contains twelve criteria, seven of them about the testing points and the options, and five of them about the passages[21]. Since the reading comprehension passages were prepared by the researcher, the five criteria about the passages were not included to evaluate the appropriateness of the four teachers' reading comprehension questions. Only the remaining seven criteria were included in the checklist. The results are shown in Table 22.

The results in Table 22 show that the four teachers' reading comprehension questions were considered appropriate in terms of the testing points but were deemed inappropriate with regard to the options. It seems that all four teachers had problems of constructing appropriate options because some of their options were either different

[21] The five criteria about the reading passages are as follows. (1) Are the passages educational, informative, and interesting? (2) Do the passages correspond to students' life experiences and cognitive scopes? (3) Are the passages appropriate in terms of difficulty level? (4) Does the length of the passage match the number of questions? (5) Do the passages provide enough new information based on which teachers can construct questions?

in difficulty level (criterion 6) or lacking in elicitation power (criterion 7). Furthermore, the lengths of their options were also not similar to one another (criterion 5).

Table 22.

Results of the Appropriateness Checklist for Reading Comprehension Questions

Checklist criteria	ET1	ET2	NT1	NT2
1. Do the questions involve no common sense so that one has to read the passage to answer the questions?	√	√	√	√
2. Does the design of questions focus on the understanding of the passage?	√	√	√	√
3. Do the questions test on students' abilities to make inferences?	√	√	√	√
4. Are the words in the question stems (and in the options) paragraphed, and were thus not completely similar to those in the passages?	√	√	√	√
5. Are the lengths of the options similar to one another?	X	X	X	X
6. Are the options equally difficult?	X	X	X	X
7. Does every option have elicitation power?	X	X	X	X

A comparison of the four teachers' considerations in constructing reading comprehension questions (see Table 7) and the appropriateness checklist for reading comprehension items (see Table 22) shows that there are only two issues that the teachers' considerations and the checklist criteria have in common. The first one is that *the testing points should cover different types* (similar to criteria 2 and 3), and the second one is that *the options can be paraphrases of the sentences in the passage* (similar to criterion 4). Although not every teacher took the two issues into consideration (see Table 7), their items were deemed appropriate in terms of criteria 2, 3, and 4 in the checklist. In general, the teachers' considerations and the authority's criteria were different to a large extent, which may account for part of the reason why the teachers' reading comprehension items were inappropriate in some ways.

At this point, we could draw the following conclusions about the four teachers'

reading comprehension questions, which were regarded by them as most difficult to construct in this study (see Table 8). First, the four teachers' reading comprehension questions could not be considered of a good quality as there were many problems with their items as identified by the two reviewers. Second, the experienced teachers seemed to produce more problematic reading comprehension questions than the two novice teachers because the former constructed more flawed question stems (see Table 20) and more problematic options (see Table 21) than the latter. Third, the options for reading comprehension questions constructed by the four teachers could not be considered appropriate because they did not match the criteria for the options in the checklist. Finally, the four teachers' test-constructing considerations were quite different from the authority's checklist criteria. This may suggest that the four teachers were probably not familiar with the test-constructing principles for the SAET; thus their items were generally judged to be inappropriate by the authority.

General Discussion on the Four Teachers' Test Construction Performances

We have examined the four teachers' considerations in constructing SAET mock tests, analyzed the qualities of their mock tests, and evaluated the appropriateness of their items by using the authority's checklists. At this juncture, we can have a general discussion on the four teachers' test-constructing performances.

To begin with, the two experienced teachers and the two novice teachers seemed to make different types of considerations in their test-constructing processes. It is found that the experienced teachers liked to use more difficult words or expressions to construct their items. Moreover, the experienced teachers' considerations were more student-oriented when they selected testing points or designed the options. They would construct their items on frequently-tested points or designed their options based on students' frequently-made mistakes. They would also produce some items to "trick

students into their traps," sometimes even at the expense of violating test-construction principles, but they did not seem to care about or be aware of this problem. By contrast, the two novice teachers, due to their lack of long teaching experiences, would often take test-construction principles into consideration although they would also make guesses of students' test-taking strategies and took them into account. NT 2 was the one that was very familiar with the test-construction principles and was aware of following the principles in designing the test items. In this study, it seems that the less experienced the teachers are, the more likely they are to follow the test-construction principles that they learn from teacher education courses.

Although the four teachers had made many considerations in constructing their tests, their considerations generally did not correspond to the authority's criteria for test appropriateness. Therefore, many of the items the teachers constructed were deemed inappropriate because the items did not meet the standards of the criteria. The mismatch between teachers' test-constructing considerations and the authority's criteria suggests that the four teachers' knowledge of test-constructing skills were quite inadequate because the teachers did not take the real important factors (i.e., the authority's criteria) into account when they constructed their tests. However, only NT 1 admitted that her test-constructing skills were poor while the other three teachers thought that their test-constructing skills were either good or ordinary (see Table 4).

In terms of test qualities, the mock tests which the four teachers constructed might not be considered of good qualities as there were many problems with the test items identified by the two reviewers. Moreover, it is found that the two experienced teachers did not seem to produce better test items than the two novice teachers did; in fact, the experienced teachers constructed more problematic vocabulary and reading comprehension items than the novice teachers did. It is also to note that the four teachers did not seem to be aware of the flaws or problems in their test items.

The results of the teachers' test-construction performances in this study echoed several findings of the past research in the literature. First, the four teachers' test-constructing considerations being different from the authority's criteria provided some evidence for Leighton, et al's (2010) observation that many teachers do not seem to be equipped with a solid grounding in the basic knowledge of testing principles or practices. Thus, teachers would often construct inappropriate test items. Second, both the present study and Kirschner, Spector-Cohen, and Wexler (1996) have found that teachers had problems in constructing reading comprehension items with good wording. This suggests that teachers may need especial training in writing good reading comprehension questions. Third, the result of the poor qualities of the mock tests in the present study also correspond with Coniam's (2009) finding that the teacher-produced tests were not of high quality from a classical test measurement perspective. Finally, the results of this study also lend support to the researchers' claim that good multiple-choice questions are difficult to construct (e.g., Cohen, 1994; Hughes, 2003) because problems might easily arise with either stems or options if the test constructors were not careful enough in the process of test-construction.

In terms of the possible test constructor effect (i.e., the length of teachers' teaching years) examined in this study, the results suggested that teachers' teaching experiences did not seem to play a major role in constructing good-quality tests because the experienced teachers did not produce better test items than the novice teachers in the present study. On the contrary, the novice teachers sometimes constructed items of better quality than the experienced teachers did. These results reveal that test-construction is not a task that a teacher with long teaching years or many teaching experiences is sure to perform well. It is believed that constructing a good test is more like creating a piece of art. Only those who are interested in test-construction and those who are willing to spend time and effort on it would be

likely to construct a good-quality test.

Based on the four teachers' test-construction performances, this study, in line with previous research (e.g., Carter, 1984; Johnson, Becker, and Oliver, 1999; Coniam, 2009), also argues for the need to have inservice teachers, experienced and novice teachers alike, attend seminars or workshops on test construction regularly to polish their test-construction skills. The seminars or workshops may contain programs as mentioned in Kirschner, Spector-Cohen, and Wexler (1996), in which the participants could be equipped with the principles of test-construction, have the hands-on experience of constructing a test, have other teachers comment on their own tests, and revise their original tests based on others' comments. It is hoped that by attending such workshops regularly, teachers can polish their test-constructing skills and thus improve the qualities of the test items. Moreover, given the fact that teachers are often ignorant of the problems with their own test items, it is also important for the teachers to have their own constructed-items moderated or reviewed by a professional or colleagues, especially when the test is constructed by one teacher alone or when the test is a large-scale one.

CHAPTER FIVE

RESULTS AND DISCUSSION ON STUDENTS' STRATEGES TO ANSWER TEST QUESTIONS

This chapter reports findings related to the second research question (RQ2) in the present study: What strategies do students use to answer the SAET mock tests? How do the higher-proficiency students use strategies to answer the test items differently from the lower-proficiency students?

There are two major sets of results in this chapter: the results of the students' scores on the four mock tests, and the analyses of the students' think-aloud protocols. First, I presented the students' scores on the four mock tests, and made a comparison between the performances of higher-proficiency students and those of lower-proficiency students. Second, I reported the results of the analyses on strategies adopted by both levels of students.

Results of Students' Performances on the Four Mock Tests

Forty-eight Taiwanese senior high school students, twenty-four being higher-proficiency and twenty-four lower-proficiency, were recruited to participate in this study, and they were divided into four groups, with each group taking one form of the SAET mock test while doing think-aloud simultaneously.

The SAET mock test contained twenty-eight items, with ten items on vocabulary, ten items on cloze, and eight items on reading comprehension. Each item was awarded one score, and the full score on the mock test was twenty-eight. Students' scores on the four mock tests were presented and discussed as follows.

Table 23 presents the means of students' scores on the forms of the mock tests. It is important to note that there were only six higher-proficiency students and six lower-proficiency students taking each of the four tests.

Table 23.

Means of Students' Scores on the Mock Tests

	All	HP	LP	D (H-L)
Form A	13.50	15.50	11.50	4.00
Form B	14.50	19.83	9.17	10.66
Form C	12.67	16.83	8.50	8.33
Form D	14.33	18.00	10.67	7.33

Note. All= all students; HP= higher-proficiency students; LP=lower-proficiency; D=Difference

As Table 23 shows, higher-proficiency students generally performed better than lower-proficiency students on the four forms of tests. One interesting finding about the scores is that although the higher-proficiency students who took Form A got the lowest score (15.50) among the HP groups, the lower-proficiency students who took the same Form A got the highest score (11.50) among the LP groups. As a result, the difference between HA and LA[1] (4.00) was the lowest among the four forms. But since this is a qualitative study and the number of the students taking each form was so small (i.e., only twelve students), no statistics measures were performed to examine whether the difference between HA and LA was significant.

Students' answers to the items on each Form were presented in Appendix K. Two interesting results could be seen from Appendix K. First, not every higher-proficiency student performed better on the test than their lower-proficiency counterparts, and not every lower-proficiency student performed worse than their high-proficiency counterparts. For example, H16A and H22C performed worse than some of their lower-proficiency counterparts on Form A and Form C respectively; L01A, L24A and L14C performed better than some of their higher-proficiency counterparts on Form A and Form C respectively. This phenomenon could be explained by several reasons. For one, since this was only an experiment, not a real test which might affect students'

[1] HA referred to higher-proficiency students who took Form A test, and LA referred to lower-proficiency students who took Form A test.

grade, the students might not have taken the FLPT and the Forms of tests seriously; therefore, their performance did not correspond to the proficiency level that they were put in. For another, since every test-taking experience is unique, even a higher-proficiency student may do poorly on a certain test, and vice versa.

Second, we could gain information from Appendix K about which items students performed well on, and which items they did poorly on. It is found that there were some items that all students answered either correctly or incorrectly. Similarly, there were also some items that all higher-proficiency students answered either correctly or incorrectly, and some items that all lower-proficiency students answered either correctly or incorrectly. Those items were noteworthy because their passing rates were either 100% or 0%. Those noteworthy items were listed in Table 24.

Table 24.

Items Worthy of Note on the Four Forms

	Form A	Form B	Form C	Form D
Items all students answered correctly		16		16
Items all students answered incorrectly	14		**19**	14
Items all HP students answered correctly	7, 24	10, 16, 19, 27, 28	4, 7, 12, 16, 20, **25**	7, 9, 16, 20, **24**, 26
Items all HP students answered incorrectly	14, **18**		1, 19	1, 14
Items all LP students answered correctly		7, 16		16
Items all LP students answered incorrectly	14, 20	3, 4, 6, **13**,	**6**, 9, 17, 19, 24	4, 13, 14, 23

Many items listed in Table 24 need to be revised because they were either too easy (so all students answered it correctly) or too difficult (so all students answered it incorrectly). Thus, they were deemed “unacceptable” in terms of item discrimination measures. Note that some of them (i.e., item 18 on Form A, item 13 on Form B, items

6, 19, and 25 on Form C, and item 24 on Form D) were also identified as problematic by the two reviewers.

The items listed in Table 24 can also be re-categorized based on the teachers who constructed them. The results are shown in Table 25.

Table 25.

Noteworthy Items Constructed by Four Teachers

	Form A	Form B	Form C	Form D
ET 1 on MA				1, 4, 13, 14, 23, **24**
ET 1 on MB	7, **18**, 20			
ET 2 on MA		3, 4, **13**		
ET 2 on MB			**6**, 7, 9, 16, 17, **19**, 20, **25**	
NT 1 on MA	14, 24			
NT 1 on MB		6, 7, 10, 16, 19, 27, 28		
NT 2 on MA			1, 4, 12, 24	
NT 2 on MB				7, 9, 16, 20, 26

As shown in Table 25, ET 1 constructed at least nine items that need to be improved based on students' performances on them. Similarly, ET 2 constructed at least eleven such items, NT 1 nine items, and NT 2 nine items. Since each of the four teachers constructed twenty-eight items in this task, the high number of the "unacceptable" or "inappropriate" items was significant. This result was also in line with the findings in chapter four that some of the teacher-constructed items were poor in quality and that those items should be revised.

Table 25 also reveals two interesting results. First, among the noteworthy items, the problematic items (those in boldface) identified by the two reviewers were all produced by the two experienced teachers. Second, ET 2 not only constructed the most noteworthy items among the four teachers, but also produced more problematic items than ET 1. The two results also support the previous finding that the

experienced teachers did not produce better test items than the novice teachers did.

The noteworthy items were presented and discussed as follows.

Noteworthy Items on Form A

On Form A, five items drew our attention: items 7, 14, 18, 20, and 24.

Item 7 (ET1B)[2]

According to the investigation, the language _____ caused the air crash. The misunderstanding of the pilot's spoken English was to blame.

(A) barrier (B) acquisition (C) requirement (D) resistance

All higher-proficiency students answered this item correctly, and four of the six lower-proficiency students (L01A, L08A, L17A, and L24A) also got it right. Besides, those who got the correct answer also did it for the right reason; that is, they knew the correct meaning of the word "barrier." Item 7 is thus one of the two items that had the highest passing rate (83.33%) on Form A.

Item 14 (NT1A)

For example, John Steinbeck is said to __14__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) have been (B) be C) has been (D) become

It was surprising that the passing rate for this item was zero. None of the twelve students who took Form A answered this item correctly. Among them, five higher-proficiency students and five lower-proficiency students chose option B instead of the correct option A. The ten students' reasons for choosing option B are as follows. Four students (H16A, H24A, L16A, and L24A) thought that "be a good listener" was what John Steinbeck was going to do. Three students (H08A, L09A, and L17A) thought that the verb after the infinitive "to" should be a verb root, and thus they chose "be." H09A thought that the phrase "is said to be" was a common expression, and thus chose B. H17A chose between "be" and "become," and she finally selected "be" because she thought that John Steinbeck has already been an

[2] The code ET1B means that the item was constructed by ET 1 based on Material B.

excellent listener, not becoming one gradually. L08A used the strategy of guessing.

Item 18 (ET1B)

__18__ is called being “fashionably late.”

(A) It (B) There (C) This (D) What

This item is also difficult for higher-proficiency students because none of them got it right. Furthermore, five out of the six higher-proficiency students chose option A instead of the correct option C for various reasons. H16A guessed that the answer was option A. H01A regarded “it is called” as a collocation, and H08A thought that “it is” was a common expression to begin a sentence. Both H09A and H17A thought that the blank asked for a pronoun and that “it” was an appropriate pronoun for it.

Surprisingly, one lower-proficiency student (L09A) got the correct answer but due to random guessing. However, based on the two reviewers’ comments, both options A and C were correct. This is a problematic item on Form A.

Item 20 (ET1B)

…it is usually good to arrive earlier than the appointment because there are usually forms that need to be __20__ by the patient.

(A) filled out (B) applied for (C) filled with (D) taken over

Like item 18, this item also had the lowest passing rate (8.33%) because only one higher-proficiency student got the correct answer and no lower-proficiency students got it right. It is very interesting that none of the five higher-proficiency students who got it wrong selected option D while the five lower-proficiency students who got it wrong chose option D, but for different reasons. Both L08A and L09A mistook “taken over” for “taken care of,” and thought it collocated with “patient” in the stem. L01 A thought that “taken over” was a verb phrase describing the doctor diagnosing the illness of the patient. L24A regarded “taken over” as “waiting for,” and L16A thought that “taken over” meant that the waiting number had passed.

Item 24 (NT1A)

What can we infer about the owner of the shoe?

(A) She took no more than thirty trips in order to carry her supplies.

(B) She had to endure the humid temperature for one year.
(C) She was a brave gold seeker.
(D) She hurt herself and dropped her shoe on her journey.

All higher-proficiency students and four of the six lower-proficiency students answered this item correctly out of different reasons. Three students (H17A, L08A, and L17A) used the strategy of guessing. Three students (H08A, H09A, and H16A) thought option C was correct because they located the word “brave” in the passage. Four students (H01A, H24A, L09A and L24A) chose option C by referring to the last sentence of the passage. Thus, this item is one of the two items that had the highest passing rate (83.33%) on Form A.

Noteworthy Items on Form B

On Form B, ten items drew our attention, and they are items 3, 4, 6, 7, 10, 13, 16, 19, 27, and 28.

Item 3 (ET2A)
Informed of his admission to his ideal university, Patrick could _____ control his joy and let out a cry.
(A) obviously (B) continuously (C) briefly (D) scarcely

This item seemed difficult for the lower-proficiency students since none of them got the correct option D. In addition, all of them interpreted the word as something related to “horrible” or “terrible.” As for the higher-proficiency students, three got the item wrong. One student (H02B) chose option C out of guessing, and two students (H18B and H23B) selected option A because they thought the word “obviously” was the correct answer.

Item 4 (ET2A)
President Ma Ying-jeou calls for reforms, and thus takes immediate _____ to put his policy into practice.
(A) figures (B) influences (C) contracts (D) measures

None of the lower-proficiency students answered this item correctly. Five of them (L02B, L10B, L15B, L18B, and L23B) stated that they did not know the

meaning of “measures,” the correct option to this item, and one student (L07B) who claimed to know this word gave a wrong definition of it; he defined “measures” as “means.”

Item 6 (NT1B)

The father held a _____ attitude on his daughter’s marriage. He let her decide who would be the one she could rely on for life.

(A) comfortable (B) promising (C) liberal (D) sincere

This item had the lowest passing rate (8.33%) on Form B. None of the lower-proficiency students got it right, and only one higher-proficiency student (H23B) chose the correct option C through guessing. It is also interesting that those who did not answer this item correctly chose either option B or option D, and only one student (L18B) chose option A, which was considered a weak distractor by one of the two reviewers.

Five students selected option B. Among them, one (L02B) used the strategy of guessing. Two students (H07B and L23B) interpreted “promising” as “allowing,” and they thought that “the father allowed her daughter’s marriage.” Two students (H02B and L15B) translated “promising” as “making a promise,” and they interpreted the stem as “the father gave his daughter a promise.”

Another five students chose option D instead of the correct option C. Both H15 B and L10B mistook “sincere” for “serious,” and they thought that the father held a serious attitude toward his daughter’s marriage since it was about his daughter’s life. H10B chose “sincere” because she thought that a sincere attitude includes the feeling of being considerate. H18B translated the word “sincere” correctly and chose it mainly through the meaning of the word. L07B chose the word “sincere” just because she recognized that it was one of the letter closings although she did not know the real meaning of “sincere.”

Item 7 (NT1B)

The student who had overslept _____ caught the school bus with her long hair clipped by the bus door.

(A) originally (B) expressively (C) totally (D) narrowly

This item is very interesting in that only one higher-proficiency student (H23B) got the correct answer through the process of eliminating the unlikely options while all the lower-proficiency students got this item right through random guessing because they did not know the meaning of the word “narrowly.” Three higher-proficiency students (H02B, H07B, and H10B) defined “narrowly” as the opposite of “wide,” and considered it inappropriate for the blank. Therefore, they eliminated it from their correct option candidates.

Item 10 (NT1B)

If good friends misunderstand each other, their friendship will be difficult to ___.

(A) preserve (B) freeze (C) maintain (D) confess

This item was an easy one for higher-proficiency students, all of whom answered it correctly with the right definition of the word “maintain.” Three lower-proficiency students (L02B, L10B, and L23B) also got it right, but only one (L10B) came up with the correct definition of “maintain.”

Item 13 (ET2A)

For example, John Steinbeck __13__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) was portrayed as (B) was told to pretend

(C) is prone to be (D) is said to have been

This item was difficult for lower-proficiency students because none of them answered it correctly. Only two higher-proficiency students (H02B and H18B) chose the correct option D, but they were not so sure whether their selection was correct. As for the other higher-proficiency students, two (H23B and H07B) chose option A. H23B used the strategy of guessing, and H07B thought that “was portrayed as” was the correct option. Another two students (H15B and H10B) chose option B. H15B thought that “John Steinbeck pretended to be an excellent listener” was correct. H10B

made a hypothesis that John Steinbeck was hated by other people because of the things he wrote; thus, he might “pretend” to be an excellent listener.

One reviewer commented that this item was poorly made because it tested on both vocabulary and grammar. The other reviewer stated that both options A and D could be correct options, and it is interesting that six students (H07B, H23B, L02B, L10B, L15B, and L18B) chose option A and only two selected option D as their correct answer.

Item 16 (NT1B)

__16__, in the United States, it is very important to be on time for almost all occasions.

(A) Therefore (B) For example (C) However (D) At first

All of the twelve students answered this item correctly, and for the correct explanation. The passing rate for this item was 100%, but its discrimination index was zero. Thus, this item should be deleted or revised. One interesting finding about this item is that one participant (L07B) interpreted “the United States” as “the United Kingdom.” Although this error did not affect her choice of the correct option, the fact that she did not know “what the United States is” was a serious issue.

Item 19 (NT1B)

Any time later than that is considered impolite, because it __19__ the host and other guests waiting.

(A) makes (B) forces (C) lets (D) keeps

All higher-proficiency students answered this item correctly by means of grammar knowledge. Only one lower-proficiency student (L18B) got this item right by judging from the meaning of the word “keeps.”

Item 27 (NT1B)

According to the research by Levi Strauss, how do most employers think about dress code?

(A) It makes their employees less productive.

(B) It’s practicable only on Friday.

(C) It helps their employees work efficiently in a good mood.

(D) It has a negative impact on their productivity.

All higher-proficiency students and three of the lower-proficiency students (L10B, L15B, and L23B) answered the item correctly, and all out of the correct reason. For the three lower-proficiency students who did not answer the item correctly, they all chose option D for different reasons. Both L02B and L07B said that he found clues in the passage about option D and that the other options were not mentioned in the passage. L18B selected option D through elimination because he thought that the other three options were false statements.

Item 28 (NT1B)

Which of the following statement is **NOT** the reason most employees love casual dress?

(A) It's comfortable to wear casual dress.

(B) Their bosses prefer casual dress rather than suits.

(C) Casual clothes are much cheaper than suits.

(D) It makes them work more happily and productively.

This item had the second highest passing rate on Form B (83.33%) since all higher-proficiency students and four of the lower-proficiency students (L07B, L10B, L15B, and L23B) got the correct answer B although they chose it for different reasons. Four students (H18B, L10B, L15B, and L23B) used the strategy of guessing. Four students (H02B, H07B, H10B, and H23B) stated that the statement of option B was not mentioned in the passage; thus it was the correct answer. One student (H15B) thought the statement in option B was false, and was thus the correct answer. One student (L07B) selected option B through the strategy of elimination.

Noteworthy Items on Form C

There are twelve items worthy of note on Form C. They are items 1, 4, 6, 7, 9, 12, 16, 17, 19, 20, 24, and 25.

Item 1 (NT2A)

To prevent people from drunk driving, the government should take the necessary ____ to punish those who drive cars after they have drinks.

(A) contract (B) errand (C) influence (D) measure

This item was difficult for higher-proficiency students since none of them got the correct option D. Yet, two of the lower-proficiency students answered this item correctly, one (L06C) through guessing, and the other (L22C) for the wrong definition of the word “measure,” which was defined as “successful” by L22C.

Item 4 (NT2A)

In order to _____ more information on the issue, the students decide to go to the library to find the books that will be helpful to them.

(A) resist (B) obtain (C) transfer (D) loosen

This item was easy for higher-proficiency students since all of them got the correct option B, with the right interpretations of the stem and the word “obtain.” Four lower-proficiency students also answered this item correctly, but only one student (L11C) knew the correct meaning of the word “obtain,” and the other three (L06C, L19C, and L22C) got the correct answer through guessing.

Item 6 (ET2B)

With the release of its new smart phones, the manufacturer Nokia _____ 160% more app downloads than Apple.

(A) boasts (B) revises (C) maintains (D) approaches

This item had the second lowest passing rate (8.33%) on Form C. Only one higher-proficiency student (H14C) got the correct answer A, and none of the lower-proficiency students answered it correctly. For those who got the wrong answers, seven students (H03C, H06C, H11C, L06C, L11C, L14C, and L22C) chose option D, and four selected option C (H19C, H22C, L03C, and L19C). Based on ET 2’s think-aloud protocols, option D was constructed to distract students, and it seemed that distractor D served her purpose well. However, according to one reviewer, the stem did not provide sufficient clues for students to answer this question and should be revised.

Item 7 (ET2B)

Well goes the saying “A distant _____ is not as good as a near neighbor.” That is,

good neighbors are a lot more helpful when we are in need.

(A) vision (B) legend (C) dealer (D) relative

All higher-proficiency students got the correct answer D although only three of them (H03C, H14C, and H22C) recognized the Chinese proverb in the stem. Four lower-proficiency students (L03C, L06C, L11C, and L14C) also got it right, and only two (L06C and L14C) came up with the correct meaning of "relative" in the stem. According to one reviewer, the proverb in the stem is Chinglish. Yet it is interesting that it had the highest passing rate (83.33%) on Form C. It is likely that since the test-constructor and the test-takers are all Chinese, they share the Chinglish conventions and are able to come up with or recognize this Chinglish sentence.

Item 9 (ET2B)

Faced with a grave danger of our drinking water, we all _____ need to apply a solution to this problem lest our health be affected day after day by industrial waste in the river.

(A) narrowly (B) rapidly (C) hardly (D) urgently

This item seemed difficult for the lower-proficiency students since none of them got the correct answer D. It is interesting that three of them (L03C, L06C, and L14C) chose option B, and three (L11C, L19C, and L22C) chose option C. Three higher-proficiency students (H06C, H19C, and H22C) also selected option B out of different interpretations of the word "rapidly." H06C defined it as "quickly," H19C referred to it as "rise," and H22C interpreted it as "immediately."

Item 12 (NT2A)

Therefore, they have __12__ enemies than other people.

(A) few (B) fewer (C) little (D) less

All higher-proficiency students answered this item correctly by using grammatical rules. Two lower-proficiency students (L11C and L14C) also got it right, but only L14C provided the correct grammatical rule for selecting option B while L11C got the item right by guessing.

Item 16 (ET2B)

__16__ ideas also differ from time to time, and from country to country.

(A) These (B) Other (C) Theirs (D) Our

All higher-proficiency students chose the correct option A by interpreting the context correctly. Three lower-proficiency students (L06C, L11C, and L14C) also answered the items correctly, with two of them (L11C and L14C) providing the justified reasons.

Item 17 (ET2B)

The only time __17__ is socially acceptable to be late is when going to a friend's party.

(A) that (B) what (C) which (D) it

This item had the third lowest passing rate (16.67%) on Form C. None of the lower-proficiency students answered the item correctly, and only two higher-proficiency students (L14C and L22C) got the correct option D, but with wrong reasons. L14 C chose option D because she found that the other three options were wrong. L22C selected option D because it was the only option that was not a relative pronoun. It seemed that students were not familiar with the usage of formal subject "it," and that they often could not distinguish the usage of formal subject from that of relative pronouns.

Item 19 (ET2B)

Being on time __19__.

(A) goes neither way (B) goes both ways (C) goes one way (D) goes either way

This item was a problematic one on Form C. No students got the correct option B. The passing rate for this item was zero. Seven of the twelve students (H03C, L19C, L03C, L06C, L11C, L14C, and L22C) chose option C as their correct answer, and five of them (H06C, H11C, H14C, H22C, and L19D) selected option D. According to one reviewer, option D was also an acceptable answer to this question. For those who chose option D, two (H11C and L19C) used the strategy of guessing. Three students (H06C, H14C, and H22C) thought that there were two choices to make: being on time or being late. One could choose either one. Thus, they regarded option D as the

correct answer.

Item 20 (ET2B)

…it is usually good to arrive earlier than the appointment because there are usually forms that need to be __20__ by the patient.

(A) figured out (B) set out (C) filled out (D) sent out

Although all higher-proficiency students got the correct option C, only four of them (H03C, H11C, H14C, and H19C) recognized the collocation “filled out the forms.” The other two students (H06C and H22C) chose the correct option for the wrong reasons. H06C thought that “filled out” was “not to be full of,” which meant that the clinic was not crowded with patients. H22C interpreted the stem as “the doctor’s time was filled with by patients.”

Three of the lower-proficiency students (L03C, L14C, and L22C) also answered the item correctly, but only L14C interpreted the context correctly. Both L03C and L22C mistook the meaning of the phrase “filled out” for the meaning of the verb “fill.”

Item 24 (NT2A)

Why did the author mention the old leather shoe?

(A) To discuss the trend of shoe wearing in the 1890s.

(B) To prove that men were not the only participants to seek gold.

(C) To introduce the Chilkoot Pass, the most dangerous site in Alaska.

(D) To show that a common thing like this may have some tales to tell.

All lower-proficiency students got the item wrong, and three of them (L06C, L11C, and L22C) chose distractor A, and three (L03C, L14C, and L19C) distractor C. Three higher-proficiency students (H11C, H19C, and H22C) also answered the item incorrectly, but they all chose distractor A. Both H11C and H22C said that option A contained the phrase “in the1890s,” which also appeared in the passage; thus they chose it. H19C chose option A because she found that the other three options were false statements.

Item 25 (ET2B)

David Smith is a(n) _____.

(A) tailor (B) consultant (C) employee (D) employer

This item was very easy for higher-proficiency students, all of whom got the correct option C for the correct definition of the word “employee.” But this item was a little difficult for lower-proficiency students, because three of them (L03C, L06C, and L22C) could not tell the difference between “employee” and “employer,” and thus chose the wrong option D. ET 2 knew students’ proficiency very well, so she constructed such an “unimportant” item, in one reviewer’s words, to make those lower-proficiency students “fall into her trap.” Yet, there is another flaw in this item. According to the other reviewer, all of the four options were arguably possible since it was an inference question, and there was no clues in the passage suggesting any of them was a wrong answer. But based on the students’ incorrect answers to this item, which was all “Ds,” the students’ thoughts were more in line with ET 2 than with the reviewer.

Noteworthy Items on Form D

There are eleven interesting items on Form D. They are items 1, 4, 7, 9, 13, 14, 16, 20, 23, 24, and 26.

Item 1 (ET1A)

English is a(n) _____ language, serving as a necessary tool to communicate with people with diverse nationalities.

(A) current (B) dominant (C) accurate (D) fashionable

This item was “a trap,” in ET 1’s words, for higher-proficiency students because none of them got the correct answer B. One of them (H20D) guessed option C, and the other five selected option A. Among the five students who selected option A, four (H04D, H05D, H13D, and H21D) interpreted the word “current” as “modern,” and they thought “modern language” was an acceptable expression. One student (H12D) chose option A by using the strategy of guessing.

Two of the lower-proficiency students (L05D and L13D) got the correct option B

through guessing. It is interesting that unlike their higher-proficiency counterparts, none of the lower-proficiency students chose option A; three of them (L04D, L12D, and L20D) randomly selected option D, and one (L21D) chose option C.

Item 4 (ET1A)

Recently, with oil and electricity prices going up, the commodity prices increased _____. Thus, more and more people barely make both ends meet.

(A) slightly (B) considerably (C) roughly (D) tentatively

This item was difficult for lower-proficiency students since none of them got the correct answer B. Three of them (L04D, L13D, and L21D) regarded the word "considerably" as the verb "consider," and eliminated this option instead of selecting it. Four higher-proficiency students (H04D, H05D, H12D, and H21D) answered this item correctly, but only three of them (H04D, H05D, and H21D) translated the word "considerably" correctly. H12D chose option B out of intuition.

Item 7 (NT2B)

They _____ planned to watch that film, but the tickets sold out, so they saw this film instead.

(A) gradually (B) originally (C) hardly (D) urgently

This item was very easy for students as it had the second highest passing rate (91.67%) on Form D. Only one lower-proficiency student (L20D) answered the item incorrectly through mere guessing. All the other eleven students correctly interpreted the word "originally."

Item 9 (NT2B)

Feeling guilty, the naughty boy who broke the window finally _____ that he did it and apologized for his wrongdoing.

(A) confessed (B) boasted (C) rumored (D) proposed

All higher-proficiency students got the correct answer A, but one of them (H20D) used the strategy of guessing without recognizing the correct definition of the word "confessed." Four of the lower-proficiency students (L04D, L13D, L20D, and L21D) also got the item correct, but all through mere guessing.

Item 13 (ET1A)

However, there are __13__ to that generality.

(A) attachments (B) addiction (C) exceptions (D) expectations

This item was difficult for lower-proficiency students because none of them chose the correct option C. Two of the higher-proficiency students (H12D and H20D) answered the item correctly, but only H12D gave the correct definition of the word "exceptions." H20D translated the word "exceptions" as "extraordinary."

Item 14 (ET1A)

For example, John Steinbeck is said __14__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) being (B) to be (C) to have been (D) to have

This item was the most problematic one on Form D because its passing rate was zero. None of the higher-proficiency students answered this item correctly; one (H04) selected option A by using the strategy of guessing, and five of them selected option B. Among the five students who chose option B, two (H20D and H21D) thought that item 14 tested on grammar, so they selected option B because the phrase "is said to be" was grammatically correct. Three students (H05D, H12D, and H13D) chose option B because they thought "is said to be" was a fixed expression used by people frequently.

Three lower-proficiency students (L04D, L12D, and L13D) also chose option A, two (L05D and L21D) selected option B, and only one (L20D) selected option D, all using the strategy of guessing.

It is an interesting coincidence that a similar one of this item appeared on Form A as well, and that none of the twelve students answered that item correctly, either.

Item 16 (NT2B)

__16__, in the United States, it is very important to be on time for almost all occasions.

(A) In addition (B) For example (C) As a result (D) Even so

All twelve students answered this item correctly, thus the passing rate being 100%. There is also a similar item of this one on Form B, and interestingly, that item

also had the passing rate of 100%. It is suggested that these two items should be deleted or revised since they were too easy for students and lacked item discrimination power.

Item 20 (NT2B)

…it is usually good to arrive __20__ than the appointment because there are usually forms that need to be filled out by the patient.

(A) later (B) earlier (C) quicker (D) slower

This item was easy for higher-proficiency students, all of whom answered the item correctly with justified reasons. Three of the lower-proficiency students (L04D, L05D, and L12D) also chose the correct option B, but only L05D provided a sensible reason. L12D used common sense to answer this item as he stated that one should usually arrive earlier when having an appointment with the doctor. L04D simply stated that since he chose "delay" for the previous item (i.e., item 19), he would choose "earlier" for item 20. He thought that the answers to these items should be contrary in meaning to each other.

Item 23 (ET1A)

According to the passage, what caused many of gold seekers to die?

(A) A lack of shoes. (B) Carrying heavy backpacks.

(C) Inadequate preparation. (D) Sudden cold weather.

This item had the second lowest passing rate (8.33%) on Form D. None of the lower-proficiency students got the correct option C, and only one higher-proficiency student (H05D) answered this item correctly. Nine students (H12D, H13D, H20D, H21D, L04D, L05D, L13D, L20D, and L21D) selected distractor D since the two words "cold weather" in option D appeared in the passage as well.

Item 24 (ET1A)

In the fifth line of the second paragraph, what does "**this**" refer to?

(A) The requirement for the gold seekers.

(B) The greatest adventures of the shoe owner.

(C) The Canadian government.

(D) Gold fever.

All higher-proficiency students chose the correct option A, but only three of them (H04D, H05D, and H20D) correctly recognized the reference of “this.” Two students (H12D and H21D) selected option A by using the strategy of guessing, and one student (H13D) thought “this” referred to the act of gold-seeking. Three of the lower-proficiency students (L12D, L13D, and L21D) also answered the item correctly, but all with the wrong reasons. L12 D thought “this” referred to gold lovers, and L13D thought “this” referred to the requirement. L21 D chose option A because it was mentioned just preceding the word “this” in the passage.

One issue worthy of note is that both of the two reviewers commented that there was no correct answer to this question because “this” in the passage referred to “one ton of supplies” carried by the gold seekers. Yet, the four higher-proficiency students who had the same interpretation of “this” as the two reviewers still chose option A. It was probably because option A was similar to their interpretation and the other three options were obviously incorrect. Concerning this issue, students, in general, seem to be very unlikely to challenge test-constructors’ authorities or the appropriateness of a correct answer. What students tend to do is select one answer among the four options, even none of which is considered acceptable or correct by the professionals.

Item 26 (NT2B)

Which of the following best describe David Smith?

(A) He cannot go out without a necktie on him.

(B) He thinks suits will make him look more handsome.

(C) He wore formal clothes to work every day five years ago.

(D) He now finds his work boring and feels tiresome to work.

This item was very easy for higher-proficiency students since all of them answered it correctly with the correct reason. Three of the lower-proficiency students (L12D, L13D, and L20D) also got the correct answer C, but only L12D provided the correct clue to this question.

General Discussion on Students' Performances on the Noteworthy Items

We have presented and discussed thirty-eight noteworthy items on the four forms of the mock tests. Strictly speaking, those items should be revised or replaced with other items because some of them had very high passing rates and the others had very low passing rates. Those items, in general, lacked discrimination power, and might be deemed "unacceptable" items. This result further strengthened our conclusions in Chapter Four that the mock tests constructed by the four teachers were not of a good quality and that many items should be revised.

Our close examinations of students' responses to the thirty-eight noteworthy items also revealed one interesting phenomenon. That is, although students answered an item correctly, they might have different interpretations of the correct option. In other words, students might get an item right for different reasons, some of which might be incorrect. As a matter of fact, it is found that students might have different responses to a certain option whether that option was the correct answer or a distractor. This interesting phenomenon was in line with the finding of previous research that students might get an item right for wrong reasons (e.g., Cohen, 1994; Cohen, 1998). Furthermore, it may also prove that the quantitative measure of raw scores on students' tests might not truthfully represent their ability or proficiency. It teachers want to figure out whether students truly understand the test items, some qualitative measures, like think-aloud, are needed.

Results of Students' Strategies to Answer Questions

Each of the forty-eight participants was required to do think-aloud while he or she was taking the SAET mock test, with me, the researcher, sitting beside him or her and doing the recording. Some of the participants did concurrent think-aloud pretty well, so I just asked them to illustrate some of the unclear points in their test-taking process at the end of our meeting. But some of the participants were not good at or

were not used to doing think-aloud, so when they lapsed into silence for some time, I would prompt them to talk by giving them some questions. Therefore, the students' protocols included not only their own talking, but also my interactions with them.

In this section, students' think-aloud protocols were analyzed for the purpose of figuring out students' strategies for answering vocabulary items, cloze items, and reading comprehension questions respectively. The results are presented as follows.

Results of Students' Strategies for Answering Vocabulary Items

To sort out students' strategies for answering vocabulary items, I examined the forty-eight students' think-aloud protocols over and over again until an exploratory coding scheme was formed. Since there is no existing coding scheme, to the best of my knowledge, for analyzing students' strategies for answering vocabulary items, I used my exploratory coding scheme to categorize students' strategies across four forms of tests. It is to note that I did not distinguish between behaviors and strategies in my categorization because I think both are part of students' test-taking processes. I thus used the term "strategies" to cover both of them in my coding scheme.

There were sixteen strategies in my coding scheme, and I further categorized them into six major types, namely, *selecting answers based on knowledge of word meanings, making an informed guess, making a random guess, eliminating, selecting answers based on other linguistic clues,* and *other non-linguistic strategies*. These strategies are presented in Table 26.

Among the sixteen strategies, two strategies need more explanation. They are strategies 1 and 10.

Strategy 1, "selecting an option based on perceived understanding of the word," means that students selected an option based on their understanding or interpretation of that option regardless of the correctness of their interpretations. In other words, if students chose an option based on their knowledge of that option, their strategy use

was coded as strategy 1. The following two examples were both coded as strategy 1.

Item 7 on Form D

They _____ planned to watch that film, but the tickets sold out, so they saw this film instead.

(A) gradually (B) originally (C) hardly (D) urgently

In answering this item, H05D first gave definitions to the four options, which were all correct definitions. Then, he selected option B, which was the correct answer to this item. H05D's strategy use for answering this item was coded as strategy 1 because he made his choice based on his understanding of the word "originally."

Item 2 on Form B

Instead of telling me _____, Judy wrote me that she didn't love me anymore.

(A) briefly (B) considerably (C) immediately (D) directly

In answering this item, H17A first gave definitions to the four options as well. However, her definitions of options B and D were incorrect. She defined option B as "think of" and option D as "definite." She then chose option D, which was the correct answer to this item. Even though H17A gave a wrong definition to the word "directly," her strategy use for answering this item was still coded as strategy 1 because she selected option D based on his knowledge of the word "directly."

Strategy 10, "selecting one option between two words similar in form," means that students selected an option between two words which were similar in spellings. The following example demonstrates the use of this strategy.

Item 10 on Form A

The English teacher _____ the English composition and then asked the students to completely understand the corrected part.

(A) revised (B) released (C) acclaimed (D) interpreted

In answering this item, L16A thought that the correct answers might either be option A or option B because the two words "revised" and "released" were similar in spellings, both of which began with the letters "re-." L16A then made a guess at option B. His strategy use was coded as strategy 10.

The sixteen strategies for answering vocabulary items are presented in Table 26.

Table 26.

Students' Strategies for Answering Vocabulary Items

Selecting answers based on knowledge of word meanings:
1. Selecting an option based on perceived understanding of the word
2. Selecting the only known word
Making an informed guess:
3. Selecting an option between two or among three known words
Making a random guess:
4. Selecting an option by intuition
5. Selecting an option between two or among three unknown words
6. Selecting an option by blind guessing because none of the four words are known
Eliminating:
7. Eliminating the unlikely option(s)
8. Selecting an unknown option because all the other three known words are not unlikely to be the correct answer
Selecting answers based on other linguistic clues:
9. Selecting an option by using grammatical clues
10. Selecting one option between two words similar in form
11. Selecting an option by recognizing the collocation that the option and the stem forms
Other non-linguistic strategies:
12. Selecting an option considering the position (A, B, C, or D) of that option
13. Putting an option back into the stem, and read the whole sentence
14. Stopping reading the other options when reaching the answer
15. Stopping reading the stem when reaching the answer
16. Changing an answer after having marked one

Based on the strategies listed in Table 26, I examined students' think-aloud protocols and sorted out the types of strategies students had used in reaching an answer, whether correct or incorrect, to a vocabulary item. It is to note that students might not use only one strategy to answer an item, and that they might combine two or more strategies in answering one item. The frequencies of each strategy that students used when doing the vocabulary items in the four forms of tests are presented

in Table 27.

Table 27.

Frequencies of Each Strategy Students Used in Answering Vocabulary Items

	Form A		Form B		Form C		Form D		Four forms		
Strategy	H	L	H	L	H	L	H	L	H	L	Total
Selecting answers based on knowledge of word meanings											(46%)
1.	40	31	41	17	45	19	43	24	169	91	260
2.	0	2	0	6	0	0	1	1	1	9	10
Making an informed guess											(1%)
3.	2	0	0	3	0	0	0	0	2	3	5
Making a random guess											(24%)
4.	1	1	1	4	0	1	0	1	2	7	9
5.	5	9	6	7	6	18	9	16	26	50	76
6.	2	6	0	12	2	16	3	16	7	50	57
Eliminating											(20%)
7.	13	8	15	15	11	12	11	10	50	45	95
8.	5	6	3	4	1	0	1	0	10	10	20
Selecting answers based on other linguistic clues											(4%)
9.	0	0	3	1	0	5	2	0	5	6	11
10.	0	2	0	0	0	0	0	0	0	2	2
11.	3	2	1	1	2	0	0	0	6	3	9
Other non-linguistic strategies											(5%)
12.	0	0	1	1	0	2	0	0	1	3	4
13.	0	1	1	7	0	2	0	2	1	12	13
14.	2	0	0	0	0	0	1	0	3	0	3
15.	2	1	0	0	0	0	0	0	2	1	3
16.	0	0	2	1	1	1	2	0	5	2	7
Total	75	69	74	79	68	76	73	70	290	294	584

Table 27 provides us with some information about students' frequently-used strategies and the differences of strategy use between higher-proficiency students and lower-proficiency students.

To begin with, the top three types of strategies students used most often across four forms are *selecting answers based on knowledge of word meanings* (46%),

making a random guess (24%), and *eliminating* (20%). This may suggest that students, most of the time, would tackle a vocabulary item directly by selecting a word they know. If they could not select one answer to an item based on a word's meaning, they would resort to other strategies, such as guessing, eliminating, or other strategies.

Concerning the strategy type "*selecting answers based on knowledge of word meanings*," it is found that higher-proficiency students used strategy 1 (selecting an option based on perceived understanding of the word) far more frequently than lower-proficiency students. On the other hand, lower-proficiency students used the strategy of selecting the only word they know in a vocabulary item (strategy 2) more frequently than high-proficiency students. It might be due to the fact that lower-proficiency students knew fewer words than their higher-proficiency counterparts, and thus they used strategy 2 more frequently.

An initial analysis of students' use of strategy 1 (selecting an option based on perceived understanding of the word) revealed two important facts about students' knowledge or usage of English words. First, students could not clearly distinguish words that are similar in forms. Take the following two words for example. Six students (H17A, L08A, L24A, H10B, H18B, and H23B) took the word "violate" for "violent," and one took (H16A) "violent" for "violate." Four students (L15B, H05D, H13D, and L21D) took the word "technique" for "technology." Second, students were often confused about the meaning of a word when it has a derivative in another part of speech. For instance, five students (H16A, L01A, L24A, L04D, and L21D) defined the word "considerably" as similar to its verb "consider," which means to think about or to regard. Four students (L10B, L23B, H11C, L20D) defined the word "hardly" as "one working diligently." A pedagogical implication in teaching concerning these findings is that students should be taught explicitly to distinguish words that are similar in forms or those whose derivatives may have different meanings.

In terms of the strategy type "*making a random guess*," lower-proficiency students used every strategy in this type (strategies 4, 5, and 6) more frequently than higher-proficiency students. In particular, lower-proficiency students adopted strategy 6 (selecting an option by blind guessing because none of the four words are known) far more frequently than their higher-proficiency counterparts. They would also "select an option between two or among three unknown words" or "select an option by intuition" more often than higher-proficiency students. In sum, although both higher- and lower-proficiency students would make a random guess when answering vocabulary items, lower-proficiency students did so more frequently. It is because high-proficiency students probably did not have the strong need to do so—*If they already knew the meaning of a word, why guess?*

The strategy type "*eliminating*" was also frequently used by both levels of students. In this type, higher-proficiency students used strategy 7 (eliminating the unlikely options) a little more frequently than lower-proficiency students, and both levels of students used strategy 8 (selecting an unknown option because all the other three known words are not unlikely to be the correct answer) equally frequently. Strategy 7 is also the second frequently-used strategy by higher-proficiency students, and it is the third frequently-used strategy by lower-proficiency students.

Concerning other strategy types, there seems to be a major difference in the use of strategy 13 between two levels of students. That is, lower-proficiency students would often "put an option back into the stem, and read the whole sentence." If they found the sentence was appropriate with the word in it, they would choose that word. In contrast, higher-proficiency students were less frequently to do so.

Results of Students' Strategies for Answering Cloze Items

To develop a coding scheme to analyze students' strategy use in answering the cloze items in this study, I used the categorization framework in Yamashita (2003) as

a starting point to formulate my own coding scheme. The framework in Yamashita (2003) was used to analyze students' think-aloud protocols of answering a gap-filling cloze test, and it included six major types: *clause level*, *sentence level*, *text level*, *extra-textual level*, *guessing*, and *missing*. These types referred to the sources of information students used to answer the items. The definitions of the six types in Yamashita (2003) are rephrased as follows. (1) *Clause level* information was the information provided by the clause in which an item appears. (2) *Sentence level* information was the information provided by a larger context than the clause in which an item appears, but within the sentence. (3) *Text level* information was the information provided by a larger context than the sentence in which an item appears, but a context from within the text. (4) *Extra-textual level* information was the information not provided by the text, which includes such mental resources as the student's background knowledge, beliefs, and images. (5) The type of *guessing* meant that students made a guess. (6) The type of *missing* meant that the student did not say anything about his/her cognitive processes or could not answer the item.

With Yamashita's (2003) framework as a starting point, I examined my participants' think-aloud protocols and formulated some concrete strategies for the framework. Then, a more suitable coding scheme was formed to analyze my data. There are seventeen strategies in my coding scheme, and they belong to five types, which are *clause-level strategies*, *sentence-level strategies*, *text-level strategies*, *extra-textual-level strategies*, and *other non-linguistic strategies*. My coding scheme for analyzing students' strategies for answering cloze items is shown in Table 28.

Table 28.

Students' Strategies for Answering Cloze Items

Clause-level strategies:
1. Using grammatical rules
2. Using collocation knowledge
3. Recognizing the idiomatic expression
4. Considering the meaning of a word or phrase in the option
Sentence-level strategies:
5. Recognizing a grammatical structure
6. Putting a word or phrase back into the sentence, and read it
Text-level strategies:
7. Considering the meaning of a transitional phrase in the option
8. Making use of the contextual clues
Extra-textual-level strategies:
9. Making judgment by personal belief
Other non-linguistic strategies:
10. Making a guess
11. Using elimination
12. Judging by intuition
13. Considering the position (A, B, C, or D) of an option
14. Changing an answer after having marked one
15. Leaving the sentence(s) without a blank unread
16. Guessing at test-constructors' considerations
17. Selecting an option which is different from the other three in form

There are two major differences between Yamashita's (2003) framework and my coding scheme. First, Yamashita's (2003) framework contained only six information sources while my coding scheme included seventeen strategies within five types. In essence, Yamashita (2003) and the present study investigated different issues: the former focused on the types of information students used to answer cloze items while the latter centered on the strategies students used to answer cloze items. I just adopted the names of Yamashita's framework and combined them with the strategies I formulated based on students' think-aloud protocols. In other words, I combined the strategies and the information sources used by students to answer cloze items. My

coding scheme is thus a pioneering one in literature. Second, the types of *guessing* and *missing* in Yamashita (2003) were not present in my coding scheme. The type of *guessing* in Yamashita (2003) was replaced with the strategy of "making a guess" in the type of *other non-linguistic strategies* in my coding scheme. As for the type of *missing*, I did not include it in my scheme because there was no missing data in the present study. It is mainly due to the different formats of cloze tests in the two studies. The cloze test in Yamashita (2003) was a gap-filling test; thus, students would leave a blank unfilled if they could not answer it. The cloze test in the present study was in the multiple-choice format, and there were no multiple-choice items unanswered because students could still make a guess at the options provided if they were not sure of the correct answers.

Among the seventeen strategies, my coding rationales of the first nine strategies need more explanation and are exemplified as follows.

Strategy 1: Using grammatical rules.

Item 19 on Form B

Any time later than that is considered impolite, because it __19__ the host and other guests waiting.

(A) makes (B) forces (C) lets (D) keeps

H07B used the word "waiting" as her clue to answer this item. She mentioned that both "makes" and "lets" were causative verbs, which should take the bare infinitive "wait," and that only the verb "keeps" could take the participle "waiting." Therefore, she chose option D as her answer. Since H07B used her grammatical knowledge of these verbs and since the clue she adopted (i.e., the word "waiting") was within the clause in which this item appeared, the way she used to answer this item was coded as strategy 1.

Strategy 2: Using collocation knowledge.

Item 13 on Form A

However, there are exceptions __13__ that generality.

(A) for (B) with (C) about (D) to

In answering this item, H16A used the word “exceptions” as her clue. She recognized that the word “exceptions” collocated with the preposition “to.” Thus, she chose option D as her answer without considering the other options. Since H16A used her collocation knowledge of the word “exceptions,” the word of which was in the same clause as the question item, H16A’s strategy use for answering this item was coded as strategy 2.

Strategy 3: Recognizing the idiomatic expression.

Item 18 on Form A

__18__ is called being “fashionably late.”

(A) It (B) There (C) This (D) What

H24A selected option D as her answer because she thought that “what is called” was an idiomatic expression she had learned before. Since the information she used was within the clause in which the item appeared, H24A’s strategy use for answering this item was coded as strategy 3.

Strategy 4: Considering the meaning of a word or phrase in the option.

Item 18 on Form D

Any time later than that is considered __18__, because it keeps the host and other guests waiting.

(A) friendly (B) hostile (C) agreeable (D) impolite

H20D answered this item based on the information in the first clause and the meanings of the four options. He mentioned that being late was being impolite. Therefore, he selected option D as his answer. Since H20D answered this item by giving a definition to each option, his strategy use for answering this item was coded as strategy 4.

Strategy 5: Recognizing a grammatical structure.

Item 11 on Form A

Because they hear more, good listeners tend to know more and to be more

sensitive to what is going on around them __11__ other people.

(A) as (B) with (C) than (D) about

In answering this item, H08A used the words “more sensitive” as her clue. She recognized that this item tested on the comparative structure “more…than,” so she chose option C as her answer. H08A’s strategy for answering this item was thus coded as strategy 5 since the clue was provided by a larger context within a sentence.

Strategy 6: Putting a word or phrase back into the sentence, and read it.

Item 17 on Form C

The only time __17__ is socially acceptable to be late is when going to a friend’s party.

(A) that (B) what (C) which (D) it

When answering this item, L06C put each of the options back in the blank and read the whole sentence. She finally chose option A as her answer because she found the sentence more appropriate with the word “that” in it. L06C’s strategy for answering this item was coded as strategy 6.

Strategy 7: Considering the meaning of a transitional phrase in the option.

Item 16 on Form B

These ideas also differ from time to time, and from country to country. __16__, in the United States, it is very important to be on time for almost all occasions.

(A) Therefore (B) For example (C) However (D) At first

In answering this item, H23B first gave definitions to each option, and then she chose option B as her answer. She thought that the author mentioned “in the United States” as an example to support his/her statement in the previous sentence. Since the clue H23B used to answer this item was beyond the sentence in which the item appeared, and she also took into consideration the meanings of each option, H23B’s strategy for answering this item was thus coded as strategy 7.

Strategy 8: Making use of the contextual clues.

Item 18 on Form B

A person usually tried to arrive about five minutes __18__ the invitation time, so that the host would have a little extra time to prepare for the guests. This is called

being “fashionably late.”

(A) after (B) before (C) on (D) by

When answering this item, H23B took two pieces of information into account. First, she thought that guests had to be late for five minutes so that the host could have time to prepare. Second, H23B regarded the phrase “fashionably late” as an important clue. Since it is called being fashionably “late,” the guests had to arrive five minutes “after” the invitation time. Thus, she chose option A as her answer. Because H23B took into account more information which covered two sentences, her strategy for answering this item was coded as strategy 8.

Strategy 9: Making judgment by personal belief.

Item 20 on Form D

However, when going to a doctor’s appointment, it is usually good to arrive __20__ than the appointment because there are usually forms that need to be filled out by the patient.

(A) later (B) earlier (C) quicker (D) slower

H20D answered this item by means of his personal belief instead of the clues in the context. He mentioned that when going to see a doctor, one should arrive “earlier,” which is common sense. He didn’t even read the clause following the blank (i.e., because there are usually forms that need to be filled out by the patient) before he selected option B as his answer. H20B’s strategy for answering this item was thus coded as strategy 9.

With the strategies listed in Table 28, I analyzed and calculated the types of strategies students used in answering the cloze items, and the types of strategies higher-proficiency and lower-proficiency students used respectively. Sometimes, students used one strategy to answer an item, and sometimes they used more than one strategy. I counted every strategy they verbalized as one token. The frequencies of each strategy that students used in answering cloze items in four forms of tests are presented in Table 29.

Table 29.

Frequencies of Each Strategy Students Used in Answering Cloze Items

	Form A		Form B		Form C		Form D		Four forms		
Strategy	H	L	H	L	H	L	H	L	H	L	Total
Clause level											(44%)
1.	6	7	11	12	20	16	5	9	42	44	86
2.	7	1	0	0	3	1	2	1	12	3	15
3.	3	2	0	0	1	1	1	0	5	3	8
4.	20	19	22	21	13	15	30	24	85	79	164
Sentence level											(5%)
5.	4	2	0	0	6	0	0	1	10	3	13
6.	0	2	2	2	3	5	1	0	6	9	15
Text level											(14%)
7.	5	4	7	13	3	3	5	6	20	26	46
8.	11	8	15	6	9	4	7	2	42	20	42
Extra-textual level											(2%)
9.	2	0	1	1	1	3	3	3	7	7	14
Non-linguistic											(36%)
10.	3	15	3	6	6	13	5	10	17	44	61
11.	18	9	13	16	13	23	11	7	55	55	110
12.	3	1	2	3	1	1	5	5	11	10	21
13.	0	0	1	0	0	1	0	0	1	1	2
14.	1	2	3	1	4	0	1	1	9	4	13
15.	0	4	0	1	0	2	4	2	4	9	14
16.	0	0	1	0	0	0	0	0	1	0	1
17.	0	0	0	0	0	0	0	1	0	1	1
Total	83	76	81	82	83	88	80	72	327	318	626

As shown in Table 29, the analyses of students' cloze test strategy use revealed several interesting trends. First, concerning the five strategy types, students tended to use *clause-level strategies* (44%) most frequently, followed by *non-linguistic strategies* (36%) and *text-level strategies* (14%). The finding that students used *clause-level strategies* most often might be due to the fact that many of the items constructed by the four teachers were local questions, so that students could answer those questions by using just clause-level strategies.

Second, among the seventeen strategies, the top three strategies that students used often frequently are strategy 4 (considering the meaning of a word or phrase in the option), strategy 11 (using elimination), and strategy 1 (using grammatical rules). This is not surprising that all students used strategy 4 most frequently because students needed to make sense of the options before they chose the answer.

Third, among the five strategy types, higher-proficiency students tended to use *clause-level* strategies more frequently than lower-proficiency students, especially strategy 2 (using collocation knowledge) and strategy 4 (considering the meaning of a word or phrase in the option). It is also found that higher-proficiency students tended to use strategy 8 (making use of the contextual clues), which belongs to the type of *text-level strategies*, more frequently than their lower-proficiency counterparts. Therefore, high-proficiency students performed better on items that tested on transitions, conjunctions, or other items that require knowledge of a global view of the passage. Moreover, higher-proficiency students also tended to use strategy 5 (recognizing a grammatical structure), which belongs to the type of *sentence-level strategies*, more frequently than their lower-proficiency counterparts.

Fourth, lower-proficiency students tended to use *non-linguistic strategies* more frequently than high-proficiency students. In particular, lower-proficiency students used strategy 10 (making a guess) more frequently than higher-proficiency students. This finding is no surprising since guessing is always a good strategy to answer a multiple-choice item if there is no better way of reaching an answer. In addition, lower-proficiency also used strategy 15 (leaving the sentences without a blank unread) more frequently than higher-proficiency students. However, among the type of *non-linguistic strategies*, higher-proficiency students tended to use strategy 14 (changing an answer after having marked one) more frequently than lower-proficiency students. This may suggest that higher-proficiency students were

more cautious about taking the test, and that they probably took more things into consideration, so that they would change their answer after they had marked one.

Finally, both levels of students used strategy 9 (making judgment by personal belief), strategy 11 (using elimination), and strategy 12 (judging by intuition) almost equally frequently. Among the three strategies, strategy 11 was the second frequently-used strategy by both levels of students.

Results of Students' Strategies for Answering Reading Comprehension Questions

To code students' think-aloud protocols for answering reading comprehension questions, I referred to Nevo's (1989) multiple-choice strategy checklist and Anderson et al.'s (1991) categorization of processing strategies to develop a coding scheme of my own. Nevo (1989) had developed a 16-item checklist for analyzing multiple-choice reading comprehension questions. I adopted four strategies in Nevo's (1989) checklist and modified them in my coding scheme. The four strategies in Nevo (1989) that I adopted are *guessing* (blind guessing not based on any particular rationale), *returning to the passage* (returning to the text to look for the correct answer, after reading the questions and the multiple-choice alternatives), *clues in the text* (locating the area in the text that the question referred to and then looking for clues to the answer in that context), and *length* (being drawn to an alternative because it was longer/shorter than the others). Anderson et al.'s (1991) had categorized students' forty-seven processing strategies into five categories, namely, *supervising strategies*, *support strategies*, *paraphrase strategies*, *strategies for establishing coherence in text*, and *test-taking strategies*. I adopted and modified five strategies in Anderson et al.'s (1991), all from the category of *test-taking strategies*, in my coding scheme. The five strategies are *selects an alternative through deductive reasoning*, *reads the questions and options after reading the passage*, *reads the questions and options before reading the passage*, *changes an answer after having marked one*, and *stops reading the*

options when they reach the answer.

By adopting and modifying the above strategies in Nevo (1989) and in Anderson et al. (1991), and by creating new strategies for coding data in the present study, I developed a sixteen-item coding scheme which is more appropriate for analyzing my participants' use of strategies for answering reading comprehension questions. The sixteen strategies are divided into four types, namely, *reading between the passage and the questions*, *making judgments on the options*, *eliminating*, and *guessing*). The coding scheme is presented in Table 30.

Table 30.
Students' Strategies for Answering Reading Comprehensions Questions

Reading between the passage and the questions:
1. Reading the whole passage to get a general idea before reading the questions
2. Reading the questions before reading the passage
3. Reading the first question before reading the passage
4. Reading the questions and looking for answers in the passage without reading the entire of it
Making judgments on the options:
5. Applying deductive reasoning
6. Considering the length of the option
7. Matching the options with the words or phrases in the passage
8. Giving a definition to the word(s) in each option
9. Returning to the passage to look for clues after reading the question
10. Returning to the passage to look for the tested word mentioned in the question
11. Stopping reading the other options when reaching an answer
12. Changing an answer after having marked one
Eliminating:
13. Eliminating an option not mentioned in the passage
14. Eliminating an option because it is a false statement according to the passage
Guessing:
15. Making a blind guess
16. Making an educated guess between two or among three options

There are five strategies in the coding scheme which need further explanation.

They are strategies 4, 5, 7, 9, and 10. Strategy 4 (reading the questions and looking for answers in the passage without reading the entire of it) belongs to the first type (*Reading between the passage and the questions*) in the coding scheme, and the first type describes how students dealt with the whole passage and the question items as a whole. Therefore, strategy 4 refers to the situation in which students read the questions first and then began to look for the answers in the passage without first finishing reading the whole passage. Strategies 5, 7, 9, and 10 belong to the second type (*making judgments on the options*) in the scheme, and the second type describes how students selected the options after reading the question items. The four strategies are exemplified as follows.

Strategy 5: Applying deductive reasoning.

Item 25 on Form A

Which if the following is the best title for the article?

(A) The Origin of Casual Friday
(B) How to Raise the Efficiency of Office Workers
(C) From Formal Wear to Casual Clothes
(D) The Dilemma between Suits and Jeans

In answering this item, H09A stated that the previous dress code in companies was to wear uniforms, and it was not until in the 1990s that employees were allowed to wear casual clothes. He thought that option C was a more complete description of the whole passage, so he chose it as his correct answer. Due to the above reasoning, H09A's strategy for answering this item was coded as strategy 5.

Strategy 7: Matching the options with the words or phrases in the passage.

Item 26 on Form A

Which of the following statements is true of "casual Friday?"

(A) It was also dubbed "dress-down Friday."
(B) In the early 1990, employees were allowed to wear casually in all the companies in America.
(C) This dress code became an immediate hit.
(D) It started out wearing casual clothes every day.

In answering this item, L16A chose option A immediately after he finished reading it because he found that the words “dress-down Friday” in option A also appeared in the passage. In other words, he matched the option with the words in the passage. L16A did not read the other options after he chose option A as his answer. Therefore, L16A used two strategies for answering this item: strategy 7 and strategy 11 (stopping reading the other options when reaching an answer).

Strategy 9: Returning to the passage to look for clues after reading the question.

Item 22 on Form C

According to the second paragraph, which of the following is **NOT TRUE** about the gold seeking journey in Alaska?

(A) People who took part in this journey were determined to be rich.
(B) The journey required its participants to carry necessities with them.
(C) The gold seekers knew the dangers before they embarked on the journey.
(D) There were challenges, such as lack of food and harsh weather, awaiting the gold seekers.

After reading the question stem and the four options, H06C returned to the second paragraph of the passage to look for clues for each option. He then found that the first line described option A, the third line described option B, and the second line described option D. After locating clues for the three options, H06C selected option C as his answer. Therefore, H06C’s strategy for answering this item was coded as strategy 9.

Strategy 10: Returning to the passage to look for the tested word mentioned in the question.

Item 28 on Form A

Which of the following words is synonymous with the word “**morale**?”

(A) integrity (B) productivity (C) happiness (D) enthusiasm

When H08A finished reading the question stem, she returned to the passage to locate the word “morale.” Then, she found that in the next line of the word “morale” appeared the word “productivity.” She thought that the words should be equal. Thus, she chose option B as her answer. H08A’s strategies for answering this item was

coded as strategy 10 and strategy 5 (applying deductive reasoning).

Based on the strategies listed in Table 30, I analyzed and calculated the types of strategies students used in answering reading comprehension questions. As in vocabulary and cloze items, students sometimes used one strategy to answer an item, and sometimes used more than one strategy. I counted every strategy they verbalized as one token. The frequencies of each strategy that students used in answering reading comprehension questions across four forms are presented in Table 31.

Table 31.

Frequencies of Each Strategy Students Used in Answering Reading Comprehension Questions

	Form A		Form B		Form C		Form D		Four forms		
Strategy	H	L	H	L	H	L	H	L	H	L	Total
Reading between the passage and the questions											(7%)
1.	2	5	5	5	3	2	4	2	14	14	28
2.	4	1	1	0	2	4	2	3	9	8	17
3.	0	0	0	0	1	0	0	0	1	0	1
4.	0	0	0	1	0	0	0	1	0	2	2
Making judgments on the options											(63%)
5.	21	16	31	14	32	20	21	11	105	61	166
6.	1	1	0	0	0	0	0	0	1	1	2
7.	30	30	16	25	10	12	15	17	71	84	155
8.	2	3	0	0	5	8	6	4	13	15	28
9.	6	3	3	11	13	2	11	5	33	21	54
10.	2	0	2	0	0	0	0	0	4	0	4
11.	0	1	0	3	1	1	3	1	4	6	10
12.	2	0	0	0	3	6	2	8	7	14	21
Eliminating											(24%)
13.	14	14	16	5	10	13	5	11	45	43	88
14.	17	5	21	6	5	4	14	6	57	21	78
Guessing											(6%)
15.	1	1	0	4	1	1	1	4	3	10	13
16.	0	5	2	2	1	11	2	5	5	23	28
Total	102	85	97	76	87	84	86	78	372	323	695

The results in Table 31 reveal several interesting findings about students' strategy use for answering reading comprehension questions. First, the top three strategies that students used most frequently are strategy 5 (applying deductive reasoning), strategy 7 (matching the options with the words or phrases in the passage), and strategy 13 (eliminating an option not mentioned in the passage). This may suggest that students, when answering reading comprehension questions, would resort to their reasoning frequently to choose an answer among four options. During the selecting process, they may also use matching and eliminating strategies as well since they usually used more than one strategy to answer the questions.

Second, higher-proficiency students used strategies 5, 7, and 14 (eliminating an option because it is a false statement according to the passage) more frequently than they used other strategies. On the other hand, lower-proficiency students used strategies 7, 5, and 13 more frequently than other strategies. The major difference between their frequently-used strategies is that higher-proficiency students used strategy 14 more often than strategy 13, while lower-proficiency students used strategy 13 much more frequently than strategy 14. The reason for lower-proficiency students' higher frequency use of strategy 13 than that of strategy 14 is probably due to the fact that judging whether an option is mentioned in the passage is easier than judging whether an option is a true statement because the former requires "matching" while the latter may require "reasoning." The results that lower-proficiency students used "matching" (strategy 7) more frequently than they used "deductive reasoning" (strategy 5) support my assumptions.

Third, higher-proficiency students tended to use strategies 5, 9 (returning to the passage to look for clues after reading the question), and 14 much more frequently than lower-proficiency students. On the other hand, lower-proficiency students tended to use strategies 7, 12 (changing an answer after having marked one), 15 (making a

blind guess), 16 (making an educated guess between two or among three options) more frequently than higher-proficiency students. Based on the differences between higher- and lower-proficiency students' use of strategies, it seemed that higher-proficiency students did more reasoning or higher-level thinking while answering reading comprehension questions, and lower-proficiency students tried textual matching or guessing more frequently. As for the strategy of "guessing," lower-proficiency students used "guessing" more frequently than higher-proficiency students not only on reading comprehension tests, but also on vocabulary items and cloze items. The limited linguistic knowledge or reading competence of the lower-proficiency learners might have accounted for their more frequent use of the strategy of guessing.

Finally, concerning the first type of strategies, twenty-eight students among forty-eight read the whole passage before answering the questions (strategy 1), and seventeen students read the passage after reading the questions (strategy 2). In addition, there seemed to be no major difference between higher- and lower-proficiency students' use of strategies 1 and 2. Those who adopted strategy 2 thought that if they read the questions before reading the passage, they would be able to better understand the main idea of the passage, and locate the key points in the passage more quickly. Adopting strategy 2 was their way to budge their time in answering the questions.

General Discussion on Students' Strategies for Answering Test Questions

So far, based on students' think-aloud protocols, I have examined their strategies for answering vocabulary, cloze, and reading comprehension items on four forms of SAET mock tests. In obtaining the results, I developed three coding schemes to analyze the data in the present study. Since the analyses were data-driven, the three coding schemes were innovative in some ways. For one, the vocabulary coding

scheme I developed is probably the first one of its kind in literature since there were no existing published coding schemes, to the best of my knowledge, to analyze students' strategies for answering vocabulary items. The one in the present study might be the first attempt. For another, my cloze coding scheme is also innovative in the way that it is a combination of Yamashita's (2003) framework of information sources and the strategies I formulated from students' think-aloud protocols. In other words, the strategies I developed not only described students' test-taking behavior but also indicated the sources of the information they used to answer the cloze items. As for the reading comprehension coding scheme, I selected some strategies from Nevo's (1989) checklist and Anderson et al.'s (1991) categorization of processing strategies to formulate a new coding scheme which is more appropriate for analyzing Taiwanese EFL students' strategies. In particular, I added two new strategies (i.e., strategies 3 and 4 in Table 30) concerning EFL students' responses to the reading passage and its questions that were not included in Nevo (1989) or in Anderson et al. (1991). In sum, although the three coding schemes in this study are primitive in nature and may need refinements, they still serve good starting points for further research on investigating Taiwanese EFL students' strategies for answering vocabulary, cloze, and reading comprehension items. The three coding schemes are also one of the contributions of the present study.

The major findings of students' strategies for answering vocabulary, cloze, and reading comprehension items are summarized respectively in the following. To begin with, concerning vocabulary items, the top three strategies that students used most frequently are *selecting an option based on perceived understanding of the word*, *eliminating the unlikely options*, and *selecting an option between two or among three unknown words*. These results suggest that students would often choose an option based on their understanding of a word's meaning. If they could not figure out the

meanings of the words, they would turn to use other strategies, such as eliminating or random guessing. It is also found that higher-proficiency students used the strategy of *selecting an option based on perceived understanding of the word* far more frequently than lower-proficiency students. This might be due to the fact that higher-proficiency students knew more words than lower-proficiency students. On the other hand, lower-proficiency students used the strategy type "*making a random guess"* far more frequently than higher-proficiency students. It is also found that both levels of students used the strategy type "*eliminating"* almost equally frequently. Since there seems to be little published research on students' strategies for answering vocabulary items, the results mentioned above shed some light on EFL students' strategy use on vocabulary items.

In regard to cloze items, students tended to use "*clause-level strategies*" most frequently than the other types of strategies, such as "*sentence-level strategies*" or "*text-level strategies*." In terms of the individual strategies, the top three strategies that students used most frequently are *considering the meaning of a word or phrase in the option*, *using elimination*, and *using grammatical rules*. It is found that higher-proficiency students tended to use three strategies much more frequently than lower-proficiency students. The three strategies are *making use of the contextual clues*, *using collocation knowledge*, and *recognizing a grammatical structure*. This suggests that higher-proficiency students were better than their lower-proficiency counterparts at using their grammar knowledge and the contextual information to answer cloze items. On the other hand, lower-proficiency students used the strategy of *making a guess* far more frequently than higher-proficiency students. It is also found that both groups of students used three strategies almost equally frequently. The three strategies are *making judgment by personal belief*, *using elimination*, and *judging by intuition*. This suggests that some non-linguistic strategies, such as *using elimination*, were used

frequently by students irrespective of their proficiency levels.

In terms of reading comprehension questions, the top three strategies that students used most frequently are *applying deductive reasoning*, *matching the options with the words or phrases in the passage*, and *eliminating an option not mentioned in the passage*. In addition, higher-proficiency students tended to use the strategies of *applying deductive reasoning* and *eliminating an option because it is a false statement according to the passage* far more frequently than lower-proficiency students. On the other hand, lower-proficiency students tended to use the strategies of *matching the options with the words or phrases in the passage* and *guessing* more often than their high-proficiency counterparts. This may suggest that higher-proficiency students would often answer reading comprehension questions based on the meanings of the options while lower-proficiency students would do so by using some non-linguistic strategies, such as *matching* or *guessing*. It is also found that both levels of students used the strategy of *eliminating an option not mentioned in the passage* almost equally frequently.

On the whole, students would use different strategies for answering different types of questions. However, it is found that students used the strategy of *elimination* quite frequently on all three types of questions in this study. Moreover, higher-proficiency students were found to use their vocabulary knowledge, grammar knowledge, and reasoning to answer the questions more frequently than lower-proficiency students, while lower-proficiency students used the strategy of *guessing* more frequently on three types of questions. Since there is little research on investigating EFL students' strategy use across three types of questions (i.e., vocabulary, cloze, and reading comprehension items) in one single study, the findings in the present study are of some significance in literature.

Although there seems to be little research on students' strategy use on vocabulary

items, there have been studies investigating EFL students' strategy use for answering cloze items (e.g., Cohen, 1984; Storey, 1997; Yamashita, 2003) and reading comprehension questions (e.g., Cohen, 1984; Nevo, 1989; Anderson et al., 1991). Some of the results in the present study corresponded with those in previous research, but some did not. In regard to research on cloze tests, both Cohen (1984) and this study found that students would use the strategy of *making guess* to answer cloze items. In addition, Storey (1997) and the present study also found that students would often use *elimination* when answering cloze items. However, some results in the present study did not correspond with the previous research. For one, Cohen (1984) found that when students did not know how to fill in a blank on the cloze test, poor students would leave it blank while better students would make guesses. However, in this study, lower-proficiency students made guesses more frequently than higher-proficiency students. This inconsistency might be due to the different cloze test formats in the two studies: Cohen (1984) used blank-filling cloze tests while the present study adopted multiple-choice cloze tests. Another difference between the result of this study and that of the previous research is that students used *text-level* information most frequently to answer the gap-filling cloze test in Yamashita (2003) while students in the present study used *clause-level* strategies most frequently to answer multiple-choice cloze items. One of the reasons resulting in the inconsistency might also be the different cloze test formats used in the two studies.

Concerning reading comprehension questions, there are also some similar findings between the present study and previous research. For example, both Cohen (1984) and this study found that students used the strategy of *matching* frequently, and that some students would read the questions first before reading the passages. In addition, both Anderson et al. (1991) and this study found that students used the strategies of *matching* and *eliminating* quite frequently when answering reading

comprehension questions. Storey's (1997) results also revealed students' frequent use of the strategy of *eliminating*. There is one difference between the results in the present study and those in Storey (1997). Students in this study used the strategy of *applying deductive reasoning* most frequently while participants in Storey (1997) used the strategy of *returning to the passage* most frequently. This inconsistency probably resulted from the fact that the coding schemes in the two studies were different in that the strategy of *applying deductive reasoning* in the present study was not included in Storey (1997).

Results of Students' Opinions about Think-aloud Method and This Study

When collecting students' think-aloud protocols, I collected not only their responses to each item on the mock test but also their opinions about this study. At the end of each meeting with students, I asked them whether the think-aloud method affected their test-taking process, and collected their opinions about this method. Then, I asked them to express their feelings about this study. Their responses are presented in Table 32. It is to note that not all students (forty-eight in total) responded to my questions. I just gave them a chance to speak up their mind about the experience of participating in this study.

As shown in Table 32, among the forty-eight participants in this study, twenty-eight of them (58%) stated that the method of think-aloud affected their test-taking process to a certain degree. It was interesting that among the twenty-eight students, nineteen of them were higher-proficiency students. That is, about 79% of higher-proficiency students (twenty-four in total) were affected by this method. On the other hand, thirteen participants (27%) claimed that their test-taking process was not affected by the method. It was surprising that twelve out of the thirteen participants were lower-proficiency students. That is, half of the lower-proficiency students were not affected by this method. Those who were not affected by

think-aloud method confessed that they had the habit of speaking up while reviewing lessons at home. Therefore, in doing think-aloud in this study, they were just speaking up their mind as they usually did at home. This finding suggests that lower-proficiency students tend to rely on "reading aloud" to comprehend a text.

Table 32.

Frequencies of Students' Opinions about Think-aloud and This Study

Item	HP	LP	Total
About the think-aloud method:			
1. It affected my test-taking process.	19	9	28
2. It did not affect my test-taking process.	1	12	13
3. It affected my accuracy rate in answering the questions.	2	0	2
4. It affected my speed of test-taking process.	5	0	5
5. It was very difficult.	0	4	4
6. It was tiring.	2	0	2
7. It was not a good method.	0	1	1
8. I was not used to it.	7	4	11
9. I talked less than I thought in mind.	0	3	3
10. I could not think when doing think-aloud.	1	0	1
11. I could not do the translation smoothly.	1	2	3
About this study:			
12. It was good.	7	8	15
13. It was fun.	1	6	7
14. It was interesting.	2	0	2
15. It was special.	1	0	1
16. It was challenging.	0	1	1
17. It was exciting.	0	1	1
18. It was no fun.	0	1	1
19. It was not interesting.	0	1	1
20. I felt nervous.	2	1	3
21. I felt frustrated.	1	0	1
22. I felt pressure.	0	1	1
23. It gave me a different experience.	1	0	1
24. It made me want to improve my English.	0	2	2
25. It helped me know more of my test-taking skills.	2	0	2

When asked about their opinions on the think-aloud method, eleven participants (23%) said that they were not used to it. Five students stated that it affected their speed of test-taking processes, and four mentioned that it was very difficult to do think-aloud. When asked about their opinions on the whole study, about half of the participants (56%) gave positive feedback, such as it was good, fun, interesting, special, challenging, or exciting. Only two participants gave negative feedback on this study. Thus, in general, the participants held a positive attitude toward this study although they had different performances in it.

I felt surprised at the result that more higher-proficiency students thought they were affected by the think-aloud method than lower-proficiency students did. I also observed in the data-collecting process that not every higher-proficiency student could talk about his or her test-taking process clearly. I found that some lower-proficiency students could do think-aloud pretty well, and they could think of reasonable explanations to justify their answers even though their answers were incorrect. Based on these results and observations, I think that with proper training, both higher- and lower-proficiency students could do think-aloud pretty well. Think-aloud could serve as a good method in collecting students' test-taking process. But it had to be done in a quiet place and on an individual basis, as one of my lower-proficiency students commented well, "I was not affected by the think-aloud method, but it would affect others!"

CHAPTER SIX

RESULTS AND DISCUSSION ON THE CONSISTENCY BETWEEN TEACHERS' TEST-CONSTRUCTING AND STUDENTS' TEST-TAKING CONSIDERATIONS

This chapter reports findings related to the third research question (RQ3) in the present study: Are students' considerations for answering the SAET mock tests consistent with teachers' test-constructing considerations? In order to answer this question, I compared the think-aloud protocols of students and teachers item by item. That is, I figured out what the teacher's testing point or consideration was in constructing an item, and then I examined whether students' considerations for solving that item were consistent with the teacher's considerations. The results are presented in the first section. The second section presents some items in which students' considerations for answering the mock tests were inconsistent with teachers' considerations for constructing the tests.

Results of Comparisons Between Teachers' and Students' Considerations

Before presenting the results of the third research question, I need to define three parts of the research question in the first place: (1) the teacher's consideration for constructing an item; (2) the student's consideration for answering that item; and (3) the consistency between the teacher's consideration and the student's consideration.

In this study, the teacher's consideration for constructing an item referred to the teacher's major concern about the testing point or the correct answer based on which he/she constructed a question stem and distractors. That is to say, although the teacher might take a lot of issues into account when constructing an item, only his/her consideration for the testing point or the clue he/she took in designing the correct answer would be taken as the comparison criterion. The following two

examples illustrate this point.

Item 2 on Form D

Human beings are not the only at risk of _____ the flu this season. The furry friends can fall ill as well.

(A) contracting (B) combating (C) separating (D) prolonging

When constructing this vocabulary item, ET 1 selected the word “contract” from the wordlist as her correct answer. She mentioned in her protocols that the word “contract” had two common meanings which belong to two different parts of speech. “Contract” can be a noun, referring to “an official agreement between two or more people;” it can also be a verb, meaning “getting an illness.” ET 1 wanted to test on students’ understanding of the verb meaning of “contract;” thus, she constructed an item using the word as her target word. In this case, ET 1’s main consideration for constructing this item was to examine whether students could correctly interpret the verb meaning of the word “contract.”

Item 20 on Form C

However, when going to a doctor’s appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be __20__ by the patient.

(A) figured out (B) set out (C) filled out (D) sent out

In constructing this item, ET 2 wanted to test on the collocation “fill out the form.” Therefore, she selected the phrase “filled out” in the cloze passage as her testing point, and came up with three verb phrases which also ended with the preposition “out.” In this case, ET 2’s consideration for constructing this item was to examine whether students could recognize the collocation of “filled out forms,” in which the verb phrase (filled out) and the noun (forms) did not occur adjacently.

The second part in the research question about “the student’s consideration for answering an item” is defined as follows. When students answered a question, they took many issues into account. But in this study, “the student’s consideration” refers to the main reason he/she selected an option as his/her answer to an item whether the

chosen answer was correct or not. The following example illustrate this point.

Item 4 on Form C

In order to _____ more information on the issue, the students decide to go to the library to find the books that will be helpful to them.

(A) resist (B) obtain (C) transfer (D) loosen

When answering this item, H06C first translated the stem into Chinese, and then he gave definitions to the words in each option. He correctly interpreted the word “obtain” in option B as “get,” and thought it fitted the stem properly. Thus, he chose option B as his answer to this item. Here, H06C’s consideration for answering the item was that the meaning of the word “obtain” was appropriate for the blank in the stem.

On the other hand, L14C answered this same item differently from H06C. Among the four options, L14C knew only the word “transfer” in option C. However, he thought that the word “transfer” did not fit the stem; thus, he made a random guess among three unknown words and selected option A (resist) as his answer to this item. In this case, L14C’s consideration for answering this item was categorized as “guessing.”

The third part of the research question that needs further explanation is “the consistency between the teacher’s consideration and the student’s consideration.” As mentioned earlier, the teacher’s consideration refers to his/her main reason for selecting a testing point, and the student’s consideration refers to his/her reason for choosing an option as the answer to an item. Therefore, based on considerations for the same item, a comparison can be made to examine whether there is any consistency between the teacher’s consideration and the student’s consideration. In this study, such comparisons were made item by item on the test for each of the forty-eight students, and thereby yielded three kinds of results: (1) the considerations between teachers and students were consistent; (2) the considerations between

teachers and students were inconsistent; and (3) the comparisons could not be made because students answered the item out of intuition, by guessing, or through elimination strategies without providing any specific reasons for selecting an option. The three kinds of results are exemplified as follows.

(1) The considerations between the teacher and the student were consistent.

Item 11 on Form A

Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them __11__ other people.

(A) as (B) with (C) than (D) about

NT 1 constructed this item to test on the comparative structure "more…than." Thus, her consideration was to examine whether students could recognize the comparative structure since the words "more" and "than" did not occur adjacently. When answering this item, H09A first located the two words "more sensitive" prior to the blank, and then selected option C as he recognized the comparative structure "more…than." In this case, H09A's consideration for answering this item was consistent with NT 1's consideration for constructing the item.

(2) The considerations between the teacher and the student were inconsistent.

Item 19 on Form B

Any time later than that is considered impolite, because it __19__ the host and other guests waiting.

(A) makes (B) forces (C) lets (D) keeps

ET 2 constructed this item to test on the usage of the verb "keep," especially the pattern of "keep + O + OC." She took the word "waiting" as a clue. Since only the verb "keep" could take the form of present participle (i.e., V-ing), ET 2 constructed three distractors which could not take the form of V-ing. However, when L23B answered this item, she did not make her choice based on the clue of "waiting," nor did she make a judgment on the usage of the verbs in the four options. Instead, she made a choice based on the meanings of the words in each option, and

selected option B, which she defined as “making someone do something.” In this case, L23B’s consideration for answering this item was inconsistent with ET 2’s consideration for constructing the item.

(3) The comparison between the teacher’s and the student’s considerations could not be made because the student answered the item either out of intuition, by guessing, or through elimination strategies.

Item 13 on Form A

However, there are exceptions __13__ that generality.

(A) for (B) with (C) about (D) to

NT 1 constructed this item to test on the collocation “exceptions to.” NT 1 wanted to examine whether students knew that the word “exceptions” takes the preposition “to.” However, in answering this item, L01A chose option B by guessing, and L24A chose option A out of intuition. In both cases, the comparisons between NT 1’s consideration and L01A’s or L24A’s considerations could not be made.

Based on the above definitions of the third research question, comparisons between teachers’ considerations for producing the items and students’ considerations for answering the items were made on an item basis for each student. The comparison results are shown in Table 33, in which the first kind of result was coded as “consistent,” the second “inconsistent,” and the third “others.”

Table 33.

Comparisons Between Teachers’ and Students’ Considerations

	Consistent	Inconsistent	Others
ET 1	91 (27%)	160 (48%)	85 (25%)
ET 2	94 (28%)	140 (42%)	102 (30%)
NT 1	129 (38%)	140 (42%)	67 (20%)
NT 2	128 (38%)	140 (42%)	68 (20%)
Total	442 (33%)	580 (43%)	322 (24%)

Note. The total frequency for the comparisons made for each teacher’s items are 336. (28 items x 12 students =336) Numbers in the parentheses are the percentages.

It is shown in Table 33 that the average consistency rate was only 33% in this study, which is not very high. Moreover, the consistency rates across the four teachers were generally lower than the inconsistency rates. These results suggest that what students took into account when answering the test items were not so congruent with teachers' considerations for constructing the items. In other words, students would often use clues different from what the teachers had used to answer the questions. One possible reason for such results is that many of the test items constructed by the four teachers were either problematic or inappropriate (see the results in Chapter Four). Students probably did not understand what the teachers wanted to test on; consequently, students' considerations for answering the items would not be consistent with teachers' considerations for producing the items.

In terms of the consistency rates, there was about 27% of consistency between students' and ET 1's considerations, 28% of consistency between students' and ET 2's considerations, 38% of consistency between students' and NT 1's considerations, and 38% of consistency between students' and NT 2's considerations. In other words, students' considerations were more in line with the considerations of the two novice teachers (NT 1 and NT 2) than with those of the two experienced teachers (ET 1 and ET 2). As for the inconsistency rates, there was about 48% of inconsistency between students' and ET 1's considerations, while there was about 42% of inconsistency between students and the other three teachers. The inconsistency rate appeared higher for ET 1 than the other teachers.

The lower consistency rates between students' considerations and the experienced teachers' test-constructing considerations might result from two factors. First, the two experienced teachers liked to use more difficult words in their items; therefore, students might not be able to understand the questions constructed by the experienced teachers at the outset, making them resort to "guessing" and

"eliminating" strategies more frequently, as shown in the percentages of Others for the two experienced teachers in Table 33. Second, the two experienced teachers tended to focus on "setting traps" for students to fall in, so they constructed several items testing on the points that students were often confused about or made mistakes on. In other words, the experienced teachers liked to make students "go astray" in answering their items; then it is little wonder that the consistency rates between students' considerations and experienced teachers' consideration were not very high.

Further analyses were conducted to explore which of the two proficiency groups of students thought in a way more congruent with the four teachers' test-constructing considerations. The results are shown in Table 34.

Table 34.

Comparisons Between Teachers' and Students' Considerations Across Two Proficiency Levels

	Higher-proficiency			Lower-proficiency		
	Consistent	Inconsistent	Others	Consistent	Inconsistent	Others
ET 1	66 (39%)	73 (43%)	29 (17%)	25 (15%)	87 (52%)	56 (33%)
ET 2	75 (45%)	64 (38%)	29 (17%)	19 (11%)	76 (45%)	73 (43%)
NT 1	96 (57%)	56 (33%)	16 (10%)	33 (20%)	84 (50%)	51 (30%)
NT 2	89 (53%)	58 (35%)	21 (13%)	39 (23%)	82 (49%)	47 (28%)
Total	326 (49%)	251(37%)	95 (14%)	116 (17%)	329 (49%)	227 (34%)

Note. The total frequency in each proficiency level for the comparisons made for each teacher's items are 168. (28 items x 6 students =168) Numbers in the parentheses are the percentages.

Table 34 shows that higher-proficiency students had higher consistency rates across the four teachers than lower-proficiency students, and that lower-proficiency students had higher inconsistency rates across the four teachers than higher-proficiency students. It is also found that the consistency rate between higher-proficiency students and NT 1 was the highest among the four rates in the higher-proficiency group, and that the consistency rate between lower-proficiency students and NT 2 was the highest among the four rates in the lower-proficiency

group. On the other hand, the consistency rate between higher-proficiency students and ET 1 was the lowest among the four rates, and the consistency rate between lower-proficiency students and ET 2 was the lowest among the four rates. In sum, students' considerations for answering test items were more congruent with those of novice teachers than with those of experienced teachers regardless of their proficiency levels.

Since each of the four teachers constructed three types of items (i.e., vocabulary, cloze, and reading comprehension) in this study, further analyses were also conducted to examine the consistency rates between students' considerations for answering different types of items and teachers' test-constructing considerations for them. The results are presented in Table 35.

Table 35.

Comparisons Between Teachers' and Students' Considerations on Three Types of Items

	Consistent	Inconsistent	Others
Vocabulary	154 (32%)	150 (31%)	176 (37%)
Cloze	161 (34%)	238 (50%)	81 (17%)
Reading	127 (33%)	192 (50%)	65 (17%)
Total	442 (33%)	580 (43%)	322 (24%)

Note. The total frequency for the comparisons made for each type of items is different. The total frequency for vocabulary items is 480, cloze items 480, and reading comprehension items 384.

The results in Table 35 suggest that there seemed to be little difference among the three consistency rates. But the inconsistency rate of vocabulary items was much lower than the other two rates, and the "others" rate of vocabulary items was much higher than the other two rates. It might suggest that in answering vocabulary items, students used guessing or elimination strategies more frequently than in answering cloze and reading comprehension items. As a result, they lowered the chances of clashing or coinciding with teachers' test-constructing considerations since they did

not provide their specific reasons for selecting an answer.

Further analyses were also conducted to examine the consistency of considerations between the two proficiency groups of students and the four teachers across the three types of items. The results are shown in Table 36.

Table 36.

Comparisons Between Teachers' and Students' Considerations on Three Types of Items Across Two Proficiency Levels

	Higher-proficiency			Lower-proficiency		
	Consistent	Inconsistent	Others	Consistent	Inconsistent	Others
Voc.	118 (49%)	77 (32%)	45 (19%)	36 (15%)	73 (30%)	131 (55%)
Cloze	115 (48%)	104 (43%)	21 (9%)	46 (19%)	134 (56%)	60 (25%)
Reading	93 (48%)	70 (36%)	29 (15%)	34 (18%)	122 (64%)	36 (19%)
Total	326 (49%)	251 (37%)	95 (14%)	116 (17%)	329 (49%)	227 (34%)

Note. The total frequency in each proficiency level for the comparisons made for each type of items is different. The total frequency for vocabulary items is 240, cloze items 240, and reading comprehension items 192.

Table 36 revealed some interesting facts. First, for the higher-proficiency students, the three consistency rates were quite similar. However, the inconsistency rate in cloze items was higher than the other two rates, and the "others" rate in cloze items was lower than the other two rates. This suggests that higher-proficiency students used fewer guessing or eliminating strategies when answering cloze items; instead, they would try to provide some reasons or clues for selecting an option. However, the considerations they offered were often inconsistent with the teachers' considerations.

Second, for the lower-proficiency students, the consistency rate for cloze items was a little bit higher than the other two rates while the inconsistency rate for reading comprehension items was much higher than the other two rates. Moreover, the "others" rate for vocabulary items was also much higher than the other two rates. These findings suggest that lower-proficiency students probably used the strategies

of guessing or eliminating more frequently in doing vocabulary items than in answering cloze or reading comprehension questions. This is understandable because vocabulary items contained fewer clues than cloze items or reading comprehension questions, both of which included a passage. If students could not figure out the meaning of a word, he/she could only answer the vocabulary item by guessing. On the other hand, it seems that lower-proficiency students would like to use information or clues in the passages instead of using guessing or intuition to answer reading comprehension questions. Nevertheless, they often adopted considerations different from what the teachers' had, thus resulting in the higher inconsistency rate for reading comprehension questions.

So far, we have examined the consistency rates of students' considerations for answering test items and teachers' considerations for constructing them from two perspectives: across four different teachers, and across three different item types. It is found that students' considerations, in general, were quite inconsistent with teachers' since the average consistency rate was only about 33% in this study. It is also found that students thought in a way more congruent with novice teachers than with experienced teachers. In addition, higher-proficiency students' considerations clashed more with teachers' considerations on cloze items while lower-proficiency students' considerations clashed more with teachers' considerations on reading comprehension questions. In the next section, the items that caused such inconsistency were examined and discussed.

Items That Caused Inconsistency Between Teachers' and Students' Considerations

In this section, I presented twelve items that caused inconsistency between teachers' test-constructing considerations and students' considerations for answering test items. My principles for selecting those items for discussion are as follows. First,

I counted the inconsistency frequencies of each item on each of the four forms. (The results are shown in Appendix L.) Then, from each type of questions on each form, I selected one item that had the highest frequency of inconsistency for discussion. After going through these procedures, I selected three items on each form, which are presented as follows. In discussing each item, I first described the teacher's considerations for constructing that item, and then I presented students' considerations for answering that item.

Items on Form A

The three items for discussion on Form A are items 3, 14, and 28.

Item 3 (NT1A)

As a good Taiwan citizen, we should _____ whatever is against the law.

(A) violate (B) frustrate (C) resist (D) decrease

This item was constructed by NT 1. In constructing this item, NT 1 said that she wanted to construct an item related to laws, and then she came up with this stem, meaning that "as a good Taiwanese citizen, we should *reject* anything that is against the law." She naturally used the word "resist" to mean "reject," according to her protocols. In addition, she thought the stem was an acceptable sentence, without considering that the word "resist" had other meanings, which would make her stem ambiguous or difficult to understand. One reviewer commented that the second part of this stem was quite unclear, and the other reviewer thought that the word "resist" did not seem to fit the logic of the stem.

Eight students' considerations for answering this item were inconsistent with NT 1's considerations. H08A and H09A translated the word "resist" as "insist," and chose option C, the correct answer. H09A interpreted the stem as "we should insist on not violating the law." L01A also selected option C, which, she thought, meant "assist," and she interpreted the stem as "we should assist the law." L08 A chose "resist" as well, but he mistook it for "rescue."

L16A and L17A selected option D, decrease, which they translated the word correctly and thought it fitted the stem. H01A and H16A selected option A, violate, and they thought of the stem as “we should not violate the law.”

These students’ responses to this item showed that none of them correctly interpreted the stem as NT 1 did. Even so, among the eight students, four (H08A, H09A, L01A, and L08A) answered the item correctly, and the other four (H01A, H16A, L16A, and L17A) incorrectly.

Item 14 (NT1A)

For example, John Steinbeck is said to __14__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) have been (B) be (C) has been (D) become

NT 1 constructed this item because she wanted to test on the structure “S + be said to + have + pp” when she saw the clause “yet he was hated by some of the people he wrote about.” She thought the word “hated” was a clue to answer this item since it was in past tense. Moreover, she thought that students would tend to put a verb root after the phrase “is said to” instead of putting a “have + pp.” structure after it. Therefore, NT 1 thought this was a good testing point.

Eleven students’ considerations for answering this item were inconsistent with NT 1’s considerations. Four students (H16A, H24A, L16A, and L24A) chose option B because they thought that “be a good listener,” a similar structure to “be a good guy,” was what John Steinbeck was going to do. Three students (H08A, L09A, and L17A) thought that the verb after the infinitive “to” should be a root, so they chose option B as well. H09A thought that the phrase “is said to be” was a common expression, and thus chose “be.” Both H17 A and H01A hesitated between option B and option D since both “be” and “become” had the meaning of “becoming.” H17A finally selected “be” because she thought that John Steinbeck had already been an excellent listener, not becoming one gradually. On the other hand, H01A selected

“become” because she thought “become” was better than “be” for this blank. L01A chose option C (has been), because she thought that the subject “John Steinbeck” was a third person singular, and thus “has been” was more appropriate for this blank.

These students’ considerations showed that none of them thought of the structure that NT 1 tested on; nor did they take the following clause “yet he was hated by some of the people he wrote about” as a clue. These eleven students’ considerations were all inconsistent with NT 1’s considerations, and their answers were all incorrect.

Item 28 (ET1B)

Which of the following words is synonymous with the word “**morale**?”

(A) integrity (B) productivity (C) happiness (D) enthusiasm

This item was constructed by ET 1. She stated that she wanted to construct a vocabulary item on a reading comprehension test, so she chose the word “morale” as her tested word. In the beginning, ET 1 was considering whether she wanted to test the synonym of the word “morale” or the definition of it. Later, she decided to test on the synonym of “morale,” and she referred to an online dictionary, and found one of its synonyms “enthusiasm,” which she interpreted as a strong feeling of interest to do something.

It seemed that ET 1 did not test on the meaning of the word “morale” in its context; instead, she just tested on the meaning of the word “morale” alone. This way, students might be able to answer this item without referring to the reading passage. If considering the word in its context, one reviewer commented that options B, C, D were all arguably possible answers.

Interestingly, almost every student who took Form A did not know the meaning of the phrase “is synonymous with.” However, they all guessed correctly that the item tested on the meaning of the word “morale” because the word “morale” was in bold type both in the question and in the passage.

Concerning this item, eight students' considerations were inconsistent with ET 1's considerations. Two of them (L08A and L24A) chose the correct option D (enthusiasm), but with different interpretations of it. L08A regarded "enthusiasm" as "leisure" while L24A translated it as "emotions" or "moods." Both of their interpretations were different from ET 1's. This showed that students may answer an item correctly even though their considerations for answering that item were inconsistent with teachers' considerations.

Six of the eight students (H01A, H08A, H09A, H24A, L01A, L17A) chose option B (productivity). H08A and H09A found the word "productivity" near the word "morale," and thought that the two might be similar in meaning. H01A and H24A thought that the passage was mainly about how to improve employees' productivity, so they chose option B, without relating it to the word "morale." L01A interpreted "productivity" as "efficiency," and she referred to the sentence that contained "morale." In other words, L01A thought that "productivity" had a similar meaning to "morale." L17A chose "productivity" without knowing its meaning, but she thought it was related to "product," a word she knew. Like H01A and H24A, L17A did not relate the word "productivity" to "morale."

Items on Form B

The three items for discussion on Form B are items 6, 15, and 22.

Item 6 (NT1B)

The father held a _____ attitude on his daughter's marriage. He let her decide who would be the one she could rely on for life.

(A) comfortable (B) promising (C) liberal (D) sincere

This item was constructed by NT 1. She first decided to use the word "liberal" among four pre-selected words as her correct option, and began to write a stem on it. After she wrote the stem "*The father held a _____ attitude on his daughter's marriage*," she thought the other three options would be acceptable answers for this

stem as well. Thus, she added a second sentence "*He let her decide who would be the one she could rely on for life*," thinking that this would make the other three options unacceptable. In other words, NT 1 viewed the second sentence as an important clue to this item.

Concerning this item, eleven students' considerations for answering this item were inconsistent with NT 1's. H23B chose the correct answer C (liberal), but she failed to access the correct meaning of "liberal." She interpreted "liberal" as "free," and thought it fitted the stem well. L18B chose option A (comfortable), because he thought that his father would make his daughter's life "comfortable."

Four students (H02B, H07B, L15B, and L23B) chose option B (promising) due to different considerations. H02B thought the stem meant that the father supported his daughter's marriage, and gave her a "promise." L15B also interpreted "promising" as "promise," and thus chose option B because she thought option A (comfortable) was an unlikely answer and she did not know the meanings of options C and D. H07B interpreted "promising" as "allowing." She selected option B because she eliminated options A and D, and she did not know the meaning of option C. L23B also viewed "promising" as "allowing" and selected it as her answer although she did not understand the meaning of the first stem sentence, "The father held a _____ attitude on his daughter's marriage."

Five students (H15B, H10B, H18B, L07B, and L10B) chose option D (sincere), also for different considerations. H15B mistook "sincere" for "serious," and he thought that the father held a "serious" attitude toward his daughter's marriage. Most importantly, he skipped the second stem sentence, and just made his choice based on the first stem sentence. L10B also interpreted "sincere" as "serious." She thought that since the marriage was about his daughter's life, the father should hold a "serious" attitude toward it. H18B gave the correct definition of "sincere," but she

misinterpreted the first stem sentence. She viewed “held a _____ attitude” as “held some kind of event for audience.” H10B interpreted “sincere” as “considerate,” and she interpreted the stem as “his father respected her daughter’s choice of selecting a spouse.” L07B did not know the meaning of “sincere,” and she just recognized that it was one of the letter closings. She also had a wrong interpretation of the stem because she thought that “the daughter could rely on her father in her life.”

It is noteworthy that all of these eleven students did not know the correct meaning of “liberal,” the correct answer to this stem. Except for H23B, who gave a wrong definition of “liberal,” and H18B, who thought the word was related to “library,” the remaining nine students all expressed that they did not know this word.

Item 15 (ET2A)

Thus, depending on __15__ a good listener does, he may become either popular or disliked in his lifetime.

(A) that (B) what (C) which (D) how

ET 2 constructed this item to test on the compound relative pronoun. ET 2 stated in her protocols that she liked to test on compound relative pronouns in constructing cloze items. If a cloze passage contained the compound relative pronoun “what,” she would make an item on it. Moreover, ET 2 had a set of distractors to compete with “what,” namely, “that, which, and how.” The reason for using “how” as a distractor was that ET 2 wanted to test students’ ability in distinguishing the two phrases, “what a good listener does,” and “how a good listener does it.” ET 2 learned from her teaching experiences that many students were often confused about the two phrases; thus, she constructed this item on the cloze passage in Material A.

There were nine students whose considerations for answering this item were inconsistent with ET 2’s considerations. Two of the nine students (H18B and H23B) got the correct answer, but for different reasons. H18B was hesitant between option

B (what) and option D (how). At first, she said that teachers liked to test on “what,” not “how” on cloze tests. Yet, she chose “how” (option D) because she was afraid that the teacher would test on the less frequently-tested “how” this time. However, she changed her answer to “what” (option B) in the end, because she found that she had written three Ds in this section; therefore, she thought that the last item in this section would be option B (what). On the other hand, H23B was hesitant between option A (that) and option B (what). Then, she reasoned that the sentence following “that” should be very long, but here in item 15 (depending on “that” a good listener does), the clause following “that” was short. Therefore, H23B selected option B (what) in the end.

Two students (L23B and L07B) chose option A (that) for different considerations. L23B thought about the meaning of each option, and finally chose option A. However, she reviewed “that” as “however,” and thought that it would be the probable answer for the blank because all the other three options were unacceptable in terms of their meanings. L07B chose option A (that) as well because she thought that the blank of item 15 was not essential and could be omitted without threatening the grammaticality of the sentence. Therefore, she selected “that” since “that” could be omitted and was often omitted in a sentence. It is obvious that both L23 B and L07B failed to recognize the need of a compound relative pronoun in item 15. On the other hand, L18B thought that item 15 asked for a relative pronoun, but he chose a simple relative pronoun, “which” (option C) instead of a compound relative pronoun, “what” (option B).

Four students (H07B, H10B, H15B, and L10B) selected option D (how) for this item, and among them, three (H07B, H10B, and L10B) hesitated between option B (what) and option D (how). All of them interpreted the item as “how to be a good listener,” and they seemed to make their final choice based on the meaning of the

phrase rather than on the grammar that item 15 required.

Judging from these students' considerations, it seemed that many of the students were not familiar with the grammar of compound relative pronouns. In other words, their considerations for answering this item did not match ET 2's considerations for constructing it.

Item 22 (ET2A)

Where was the old leather shoe first located?

(A) In a national museum. (B) On an icy mountain trail.

(C) On a display shelf in a store. (D) In an underground gold mine.

ET 2 constructed this item, which was a local question on details. ET 2 stated that she used the word "located" in the question to replace "found" or "discovered" mentioned in the passage. She constructed the correct option B in a way that she wanted students to integrate the ideas mentioned in the first paragraph; thus, the words she used in option B were a little different from those used in the passage. ET 2 also mentioned that this item would be easy for students to find the correct answer.

Nine students did not answer this item the way ET 2 had expected. L15 B chose the correct option B because she found that the passage mentioned a "trail," and the "cold weather." She matched the "icy" mountain with the cold weather, and thus chose option B. H10B also selected option B by using the sentence "*Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice?*" as a clue. She said that this sentence mentioned ice, and that the woman must have climbed the mountain of ice, so she chose option B. H10B's considerations seemed to resemble those of ET 2; yet on second thought, she used just the strategy of "matching" the word "icy" in the question with "ice" in the passage without synthesizing the ideas in the first paragraph. As a result, her considerations for answering this item were still deemed inconsistent with ET 2's.

Three students (H07B, L02B, and L23B) chose option D for this item. H07B

hesitated between option B and option D because she thought the woman had been to a mountain of ice and the shoe was related to gold mine. H07B later decided to choose option D because the passage mentioned about “Alaska,” a place to seek gold. L02B mentioned that the woman’s shoe was found in a place people sought gold. Since option D contained the word “gold;” thus, he chose it. L23B had similar considerations as L02B. She said that the passage was about gold; therefore, she chose option D, which contained the word “gold.”

Four students (H15B, L07B, L10B, and L18B) chose option A for similar reasons. They all used the sentence “*Such is the case with an old leather shoe in a museum in Alaska*” in the first paragraph as a clue for this item. When they saw the word “museum” in option A, they though it would be the correct answer.

Items on Form C

The three items for discussion on Form C are items 6, 19, and 23.

Item 6 (ET2B)

With the release of its new smart phones, the manufacturer Nokia _____160% more app downloads than Apple.

(A) boasts (B) revises (C) maintains (D) approaches

ET 2 constructed this item on current events, but she did not mention in her protocols how she came up with such a stem and the correct option A (boasts). Yet she explained clearly how she designed the three distractors, namely, options B, C, and D. She said that she put option C here on purpose because she wanted students to eliminate it first since it was the most unlikely option in this item. Option B was used to “lure” students who thought they could “*revise* or *rewrite* the history,” in ET 2’s words. Option A (approach) was “a big trap” ET 2 had set for students. She said that the word “approach” had the meaning of “reaching” or “moving toward something;” thus, students would regard the phrase “approach 160% more app downloads” acceptable in English as it was in Chinese. However, ET 2 said that

“approaching a certain number” was grammatical while “approaching a certain percentage more than the other” was ungrammatical. Therefore, she wanted to test students on their knowledge of the usage of the verbs “approach” and “boast.”

Seven students’ considerations for answering this item were inconsistent with ET 2’s considerations. Contrary to ET 2’s assumption that students would eliminate option C first, three of students (H19C, H22C, and L19C) chose option C (maintains), but with different interpretations of it. H19C thought the word “maintain” had the meaning of “addition,” H22C translated “maintain” as “include,” and L19C interpreted it as “machine.”

Four students (H03C, H06C, H11C, and L22C) chose option D (approaches). One of them (L22C) translated “approach” as “apply” while the other three interpreted the word as “reaching,” and thought that the meaning of “approaching 160%” was acceptable. Obviously, the three students “fell into the trap ET 2 had set” since they did not know the correct usage of the verb “approach.”

All the seven students did not get the correct answer A (boasts), and four of them (H19C, H11C, L19C, and L22C) stated that they did not know the meaning of the word “boast.” Three students (H03C, H06C, and H22C) knew the correct meaning of the word, but did not choose it as their answer. One reviewer commented that the stem of this item did not provide sufficient clues, and that it should be revised.

Item 19 (ET2B)

Being on time __19__. One should also not arrive early for a friend’s party, because it would rush the host. However, when going to a doctor’s appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be filled out by the patient.

(A) goes neither way (B) goes both ways (C) goes one way (D) goes either way

This item was a global question constructed by ET 2. She said that students had to understand the whole passage so as to answer this question correctly. The correct

answer to this item, according to ET 2, was option B, but one reviewer thought that options B and D were both acceptable answers. ET 2 had her considerations for designing those options. She thought that options A and C should be easily eliminated by students if they understood the whole passage, because the two options were incorrect. The most difficult part for students was to decide whether option B (goes both ways) or option D (goes either way) was correct. ET 2 thought that only option B was correct because one had to do both things: to be on time for some occasions, *and* to be late for others.

Nine students' considerations for answering this item clashed with ET 2's considerations. Three of the nine students (H06C, H14C, and H22C) chose option D. H22C mentioned that one of the options was "goes one way," so there might be the other way. Thus, she reasoned that one should choose between the two ways, and she selected "goes either way." Both H06C and H14C reasoned that since the passage mentioned about "going to a party later" and "not to go to a party early," there were two ways to choose from. Thus, they considered "goes either way" the correct answer. It was clear that both of them had misinterpreted the main ideas of this passage.

Six students (H03C, H19C, , L14C, L06C, L11C, and L22C) selected option C (goes one way). H03C, L11C, and L22C thought that "being on time" was one way, so they chose option C. L06C hesitated between option C and option D, and she selected option C in the end because she said that there was only "one" example in the passage. L14 C selected option C because he thought "goes one way" was used to modify the following sentence "<u>One</u> should also not arrive early…." H19C said that the passage told us "not to arrive early" and "not to arrive late;" that is, there was only <u>one</u> way for us: "being late." Thus, she chose option C (goes one way.)

Judging from these students' considerations, it is clear that none of the students

made good use of the clues in the second paragraph, because none of them mentioned the clue that one should arrive earlier for a doctor's appointment in answering this item. Students' considerations showed that their strategies for solving global questions need to be improved.

Item 23 (NT2A)

What can we infer from this passage?

(A) The woman brought with her the supplies which weighed over 40 pounds.

(B) The woman who joined the gold-seeking trip followed the crowd in clothing.

(C) The woman knew the journey was full of dangers when she decided to make it.

(D) The woman could not stand the cold weather because she lost the leather shoe.

NT 2 constructed this inferential item based on the sentence "*It is a woman's shoe of a style popular in the 1890s*" in the first paragraph of the passage. He reasoned that since that shoe was popular at that time, the woman might be following the fashion of wearing the shoes as the crowd did. Based on such reasoning, he wrote the correct option (B) by changing some key words in it, such as replacing "shoes" with "clothing." As for the other distractors, NT 2 used some ideas in the passage and modified them to construct the distractors. Thus, he thought that the four options met the requirements of inference. However, both of the two reviewers commented that none of the options could be inferred from the passage. In other words, they thought that there was no correct answer to this item.

Eleven students had considerations different from NT 2's for answering this item. Among them, L19C chose the correct answer B through elimination. Although he did not know what "the gold-seeking trip" and "the crowd in clothing" meant, he still selected it because he found the other three options wrong. Three students (H06C, H19C, and L22C) chose option A. They said that options B, C, and D were not mentioned in the passage, and option A was in the second paragraph; thus, they

selected option A based on the clue "*They would carry their supplies in backpacks each weighing up to fifty pounds*."

Seven students (H03C, H11C, H22C, L03C, L06C, L11C, and L14C) chose option C, yet, for different considerations. L06C and L14C thought that option C was a correct statement, and thus chose it. H03C and L11C used the sentence "*Whoever dropped the shoe must have been a brave and determined woman*" as a clue to select option C. They thought that since the woman was brave, the statement in option C would be correct. H11C thought that the main idea of the passage was a dangerous journey. He thus selected option C, which contained the word "dangers" in it. L03C chose option C because she felt that the statement was quite positive, so it might be the correct answer. Finally, H22C reasoned that there were many dangers in the trip, such as lack of food and harsh weather; thus, she selected option C.

Items on Form D

The three items for discussion on Form D are items 1, 14, and 23.

Item 1 (ET1A)

English is a(n) _____ language, serving as a necessary tool to communicate with people with diverse nationalities.

(A) current (B) dominant (C) accurate (D) fashionable

This item was constructed by ET 1, who chose the word "dominant" as her target word and wrote a stem by herself. In designing the stem, the first thought came to her mind was "English is a dominant language," in which the word "dominant" meant "more powerful and important than others." Then, she made a stem explaining why English is a dominant language. ET 1 used words that collocate with "language" in Chinese (but not necessarily in English) as her distractors, such as "current, accurate, and fashionable."

Six students' considerations for answering this item were inconsistent with those of ET 1. Two of them (L04D and L12D) chose option D (fashionable) while

four students (H04D, H05D, H13D, and H21D) chose option A (current). Those who chose “current” all gave a correct definition of it, and they all made the selection right away after they finished reading the stem and the four options. It seemed that the expression “current language” flashed into their mind quickly just as “dominant language” came to ET 1’s mind immediately.

Among the six students, only H04D knew the correct meaning of “dominant.” Therefore, “a dominant language” might not be so familiar to students as “a current language” or “a fashionable language” was. Moreover, the second part of the stem (*serving as a necessary tool to communicate with people with diverse nationalities*) did not seem to function well as ET 1 had expected because all of the students did not use it as a clue to answer this item. They simply used their collocation knowledge of the word “language” to answer it.

Item 14 (ET1A)

For example, John Steinbeck is said __14__ an excellent listener, yet he was hated by some of the people he wrote about.

(A) being (B) to be (C) to have been (D) to have

Like NT 1, ET 1 constructed this item to test students’ knowledge of the structure “S + be said to + have + pp.” In constructing the distractors, ET 1 included “to be” and “to have” in them because she learned from her teaching experiences that many students could not distinguish between the two phrases. Thus, the two options were “the traps” she had set for students.

Eleven students’ considerations for answering this item were inconsistent with those of ET 1. One student (L20D) hesitated between option B and option D, and then he selected option D (to have) because he found it very strange to have two “be” verbs in one sentence. Three students (H04D, L12D, and L13D) chose option A, but for different considerations. L13D stated that since options B, C, and D contained the infinitive “to,” he wanted to choose an option which did not include

the infinitive “to” in it. Therefore, he selected option A (being). H04D chose option A because she thought “being” was more appropriate than the other options for this blank. Ll2D used his grammatical knowledge to answer this item. He said that because “is said” was a passive voice and there was already an “is” there, thus he chose “being” to fit the grammatical structure there.

Seven students selected option B (to be). Five of them (H05D, H12D, H13D, H20D, and H21D) mentioned that “is said to be” was a common expression which they encountered quite frequently. L05D chose option B because he thought the word “said” should be followed with “to.” Yet, he considered options C and D unacceptable even thought they contained “to” as well; thus, he selected option B. L21D chose option B due to the meaning of the phrase “to be,” which she interpreted as “to be a good one.”

These students’ considerations showed that none of the students regarded this item as a test on the structure “S + be said to + have + pp.” Many of them just thought of it as a grammatical item on verb tense or a test on a fixed expression. Based on students’ performances on and considerations for this item, it is obvious that this structure was not easy for students to recognize, and was probably difficult for them to learn it well.

Item 23 (ET1A)

According to the passage, what caused many of gold seekers to die?

(A) A lack of shoes. (B) Carrying heavy backpacks.

(C) Inadequate preparation. (D) Sudden cold weather.

ET 1 constructed this inferential item based on the clues in the second paragraph of the reading passage in Material A. In ET 1’s interpretation of the passage, she thought that many gold seekers died during the trip due to insufficient equipment. Thus, she wrote “inadequate preparation” as her correct option. The other distractors were designed by using some key words in the passage, such as

"shoes" and "cold weather." According to ET 1, option D was designed on purpose. Since the passage mentioned "*many died of starvation and exposure to cold weather*," she constructed an option "sudden cold weather" to "confuse" students. ET 1 reasoned that "cold weather" would be an acceptable answer, but "*sudden* cold weather" was incorrect. Therefore, she added the word "sudden" to the phrase "cold weather" to "lure students into her trap."

Eleven students had fallen into her trap since nine of them (H12D, H13D, H20D, H21D, L04D, L05D, L13D, L20D, and L21D) chose option D (Sudden cold weather.). for the same reason of using the sentence "*many died of starvation and exposure to cold weather*" as their consideration for answering this item. None of them noticed that the word "sudden" in option D was not mentioned in the passage, and they chose option D because of the phrase "cold weather" in it.

Two students selected option A (A lack of shoes.). H04D thought that since the people lost shoes, they might die. L12D reasoned that the second paragraph mainly talked about the fact that shoes caused gold lovers to die.

These students' considerations for answering this item suggested that they did not use their understanding of the second paragraph to answer this item as ET 1 had expected. Instead, most of them used "matching" strategies of looking for similar words in the question, options, and the passage. Thus, it was very easy for them to "fall into the trap" ET 1 had set for them if they were not careful enough.

General Discussion on the Inconsistency Between Teachers' and Students' Considerations on the Four Forms

After examining the twelve items that caused inconsistency between teachers' and students' considerations, we had the following four observations.

First, students' insufficient vocabulary knowledge and misinterpretations of the vocabulary stems might have caused the inconsistency between teachers' and

students' considerations for vocabulary items. The finding that students had different interpretations of the vocabulary stems from those of the teachers who constructed them might be due to the fact that some stems were ambiguous and difficult to understand. Since students could not figure out what the teacher meant in a stem, they were very likely to use considerations different from the teacher's. On the other hand, it is also often the case that students' insufficient vocabulary knowledge resulted in the inconsistency on vocabulary items. For example, some students knew too little vocabulary, and some could not distinguish words that are similar in forms, and often mistook one for the other (e.g., *resist* vs. *insist*). The ignorance or misinterpretation of words might have accounted for a large percentage of the inconsistency on vocabulary items.

Second, the clues that students used to answer a cloze item were often different from those that teachers used to construct that item. The adoption of different clues by teachers and students might account for part of the inconsistency between their considerations on cloze items. In addition, students sometimes would not be able to recognize the testing points on which teachers constructed items; therefore, they would use considerations different from what teachers had expected to answer the items. Students' low proficiency in English might result in their inability to recognize teachers' testing points on cloze tests.

Third, students' frequent reliance on the strategy of "matching" might have caused the inconsistency between students' and teachers' considerations for reading comprehension items. It is found that when answering reading comprehension questions, students often matched the options with words or phrases in the passage without synthesizing the clues in the context. Moreover, students would use the strategy of "matching" to answer almost every type of reading comprehension questions (i.e., global, local, referential and inferential types). It was students' heavy

reliance on “matching” without critical thinking that the inconsistency between teachers’ and students’ considerations arose.

Fourth, close examinations of the items which caused inconsistency between teachers’ and students’ considerations showed that there were quite a few cases in which students answered items correctly for reasons that were inconsistent with teachers’ considerations. This finding is in line with Cohen (1994), which showed that students may get an item wrong for the right reasons or right for the wrong reasons. But since this issue is not the major concern in the present study, we did not do further analyses of it, and left this issue for future research.

Our third research question was motivated by Cohen (1984) and Nevo (1989), both of which have indicated the phenomenon that students might answer test items in ways different from what teachers have expected. The results and findings in this chapter serve as empirical evidence for their statement because there were many cases in which students’ and teachers’ considerations were different. Furthermore, the findings of the third research question in this study are significant in the EFL context because the present study was the first one to examine whether students’ considerations matched teachers’ test-constructing considerations.

In literature, Gierl (2001) has conducted a similar study, which compared cognitive representations of test developers and those of elementary school students on a mathematics test. Gierl (2001) found that the overall match between the test developers’ expected responses and the students’ observed responses was 53.7%. However, in the present study, the average consistency rate between teachers’ test-constructing considerations and students’ considerations for answering the tests was only 33%. These two results suggest that test-takers indeed would think differently from test-writers on the same test items. Moreover, since the consistency rates in the two studies were not very high, the validity of the test items were

threatened since the items constructed by the test-writers did not measure exactly what they wanted to measure. To shed more light on the phenomenon of “the inconsistency between test-writers’ test-constructing assumptions or considerations and test-takers’ responses to or considerations for answering the items” in the testing field, more research is needed to examine this “closeness-of-fit” issue, just as Gierl (2001) and the present study have conducted.

CHAPTER SEVEN

CONCLUSION

This chapter consists of four sections. The first section presents a summary of the major findings of this study. The second section reports several pedagogical implications. The third section addresses the limitations of the present study. The fourth section suggests directions for future research.

Summary of the Major Findings

The purposes of this study were to investigate how experienced and novice teachers constructed SAET mock tests, to examine how higher- and lower-proficiency students took the mock tests, and to explore whether there was any match or mismatch between teachers' test-constructing considerations and students' considerations for answering the test items constructed by teachers. To achieve these purposes, think-aloud method was adopted to collect data, which included four teachers' think-aloud protocols of constructing twenty-eight items of multiple-choice questions on vocabulary, cloze, and reading comprehension, and forty-eight students' think-aloud protocols of answering those items. Analyses of the participants' protocols yielded the following major results.

Concerning teachers' test-constructing considerations, it is found that the two experienced teachers and the two novice teachers seemed to make different types of considerations in their test-constructing processes. The experienced teachers' considerations were more student-oriented because they often designed their test items based on students' frequently-made mistakes or on the vocabulary or grammar that students were confused about. In addition, the experienced teachers sometimes liked to use difficult words or expressions to construct their items. By contrast, the two novice teachers tended to take test-construction principles into consideration while producing their test items. It seems that the more experienced the teachers are,

the less likely they are to follow the test-construction principles they have learned from teacher education courses. Nevertheless, whether the teachers' test-constructing considerations were student-oriented or principle-oriented, the four teachers' considerations, in general, did not correspond with the authority's criteria for test appropriateness. As a result, many of the items the teachers constructed were deemed inappropriate in terms of the authority's criteria. With regard to test qualities, the mock tests that the four teachers constructed could not be considered of good qualities since there were some flaws or problems with the test items. It is also found that the two experienced teachers did not seem to produce better test items than the two novice teachers did. As a matter of fact, the experienced teachers constructed more problematic vocabulary and reading comprehension items than the novice teachers.

In terms of students' strategy use, it is found that students adopted different strategies when answering different types of questions. In answering vocabulary items, students often selected an option based on their perceived understanding of that word, eliminated the unlikely options, or made a random guess. Moreover, higher-proficiency students tended to consider the meaning of each option more frequently than lower-proficiency students, while lower-proficiency students resorted to random guessing more frequently than higher-proficiency students. In answering cloze items, students would often consider the meaning of the options, use elimination, or applying grammatical rules to help them select the correct answer. Higher-proficiency students were found to use contextual clues far more frequently than lower-proficiency students on cloze items, while lower-proficiency students used guessing more often than higher-proficiency students. When answering reading comprehension questions, students often used the strategies of deductive reasoning, matching, and eliminating most frequently. In addition, higher-proficiency students were found to use deductive reasoning and eliminating far more frequently than

lower-proficiency students, while lower-proficiency students adopted matching and guessing more often than higher-proficiency students. On the whole, both levels of students used the strategy of elimination quite frequently across three types of questions. Moreover, higher-proficiency students used their vocabulary knowledge, grammar knowledge, or reasoning to answer the questions more frequently than lower-proficiency students, while lower-proficiency students used the strategy of guessing more frequently than higher-proficiency students on three types of questions.

The comparisons of teachers' test-constructing considerations and students' considerations for answering the test items showed that the overall consistency rate between teachers' and students' considerations was only about 33% in this study. It is also found that students thought in a way more congruent with novice teachers than with experienced teachers. In terms of the proficiency levels, higher-proficiency students' considerations clashed more with teachers' considerations on cloze items while lower-proficiency students' considerations clashed more with teachers' considerations on reading comprehension questions. In sum, the overall low consistency rate in this study suggests that there was a great mismatch between the considerations of teachers and of students. The flaws or problems of the question items and students' limited vocabulary knowledge and low proficiency levels might be possible reasons that led to such inconsistency.

Pedagogical Implications

This study has explored teachers' test-constructing processes, examined students' strategy use in answering the test questions, and compared teachers' test-constructing considerations and students' considerations for answering the test items produced by teachers. Based on the findings of this study, we draw the following pedagogical implications.

To begin with, teachers or test-constructors are advised to regularly attend

seminars or workshops on test construction to polish their test-construction skills. It is found in this study that both experienced teachers and novice teachers constructed many poor and inappropriate questions items. Therefore, a regular reminder from the test-construction training programs would be beneficial to improve teachers' ability in test-construction. Moreover, it is also necessary for teachers to have their own items reviewed or modified by their colleagues or some professionals because teachers are often unaware of the flaws in the items they have constructed. When constructing a large-scale test, like a midterm or a final in school, teachers are strongly advised to do so. In a context where the task of constructing a midterm or a final is one teacher's solitary work, the reviewing of the test items is of vital importance. If possible, test-construction had better be done in a team which includes both experienced and novice teachers. In this way, the experienced teachers can provide materials on which students frequently make mistakes, and the novice teachers can apply their still-fresh knowledge of testing principles to the test-construction task. It is hoped that by incorporating both groups of teachers' considerations into the test-construction process, the items designed by the team would contain fewer problems or flaws, and would be more appropriate for the targeted test-takers.

In classroom teaching, teachers can turn their testing points on a test into useful teaching points. It is important that teachers explain the difficult items or the "tricky" items they construct on a test for those students who take that test. For example, it is found in this study that many test-takers mistook "violate" for "violent," and "resist" for "insist, " which suggested that the test-takers might be confused about some similarly-looking words. Thus, in a follow-up teaching session, teachers can highlight or introduce words that are similar in spellings to help students distinguish those words well. This kind of follow-up teaching sessions on test items would be beneficial to students since they can have a second chance to learn the testing points they are not

familiar with prior to taking the test.

Finally, since test scores alone could not reflect students' real ability on a test, teachers or test-constructors may need to adopt some qualitative measures (e.g., think-aloud), along with test scores, to help them gain a better understanding of students' true ability on a test. It is found in this study that there were many cases in which students answered an item correctly for wrong reasons or for considerations different from the test-constructors' considerations. Moreover, the strategies of guessing and matching were also involved frequently in students' test-taking processes. Therefore, teachers who want to explore whether their students truly understand the test items are advised to use think-aloud method to examine their students' test-taking processes. By analyzing students' think-aloud protocols, teachers can accurately locate their students' errors or mistakes and thus better understand students' true knowledge of the test items. In addition, the results of students' think-aloud protocols will serve as valuable materials on which the lessons in a remedial class can focus.

Limitations of the Study

This study examined how teachers constructed SAET mock tests and how students answered those tests. Although it has yielded several interesting findings which provide practical pedagogical implications for teachers and test constructors, this study has three major limitations.

First, this study adopted a convenient sampling method to recruit the student participants, who were all students in the same senior high school in northern Taiwan. Therefore, generalizations of the results of this study should be made with caution.

Second, the participants were asked to do think-aloud while performing their tasks. Despite receiving instructions and practice sessions prior to the study, some participants (including ET 2 and some students) were still unable to do verbal report

well due to their lack of familiarity with this method or not being used to it. Even though the participants had a fluent verbal report behavior, we still could not ensure that they had verbalized the strategies or considerations on their minds clearly and completely as some participants had mentioned on their feedback sheets or in their interview sessions that what they said was less than what they thought. Since the participants might have something left unsaid on their minds, their think-aloud protocols might not represent a whole picture of their test-constructing or test-taking processes.

Third, since this is a qualitative, data-driven study, which involves a lot of coding work, the design of the coding scheme and the reliability of coding were important. Although we made every effort to design sound coding schemes, there still might be some flaws in them. Furthermore, despite our effort to ensure the consistency of coding, there might be strategies or considerations coded under different categories in the coding process. The absence of the second coder was thus one limitation in this study.

Directions for Future Research

Based on the limitations and findings of this study, future research in this area can adopt the following directions.

First, in terms of the student participants, future research can recruit students from a larger scale of more diverse backgrounds, so that the results could be generalized to a wider population.

Second, due to the limitations of think-aloud method mentioned above, researchers who would like to adopt think-aloud method can pay attention to two issues. For one, select participants prudentially. It is suggested that articulate students, rather than reticent students, be better candidates as participants because the former could talk more in their verbal report than the latter. For the other, give the

participants more training sessions to practice the think-aloud method until the participants are used to it.

Third, the results of the third research questions in the study suggested that students might answer an item correctly for reasons that are inconsistent with teachers' considerations. Since this issue was not further examined in the present study, future research can explore this issue more thoroughly. It would be an interesting line of validity research in testing field.

REFERENCES

Afflerbach, P., & Johnston, P. (1984). On the use of verbal reports in reading research. *Journal of Reading Behavior*, *16*(4), 307-322.

Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, *14*(2), 115-129.

Anderson, N.J., Bachman, L., Perkins, K., & Cohen, A. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing*, *8*(1), 41-66.

Arndt, V. (1987). Six writers in search of texts: A protocol-based study of L1 and L2 writing. *ELT Journal*, *41*(4), 257-267.

Bachman, L. F. (1997). Generalizability theory. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (pp. 255-262). Dordrecht: Kluwer Academic.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language Testing*, *17*(1), 1-42.

Bailey, K. M., & Brown, J. D. (1996). Language testing courses: What are they? In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 236-256). Philadelphia, PA: Multilingual Matters.

Banerjee, J. & Luoma, S. (1997). Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (pp. 275-287). Dordrecht: Kluwer Academic.

Block, E. (1986). The comprehension strategies of second language readers. *TESOL Quarterly*, *20*(3), 463-494.

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing*, *12*(1), 1-15.

Brown, J. D. (Ed.). (1998). *New ways of classroom assessment*. Alexandria, VA: TESOL.

Brown, J. D., & Bailey, K. M. (2008). Language testing courses: What are they in 2007? *Language Testing*, *25*(3), 349-383.

Buck, G. (1991). The test of listening comprehension: An introspective study. *Language Testing*, *11*(2), 145-170.

Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education*, *35*(6), 57-60.

Chapelle, C. (1988). Field independence: A source of language test variation? *Language Testing*, *5*(1), 62-82.

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing*, *1*(1), 70-81.

Cohen, A. D. (1987). Studying learner strategies: How we get the information. In A. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 31-40). Englewood Cliffs, NJ: Prentice Hall.

Cohen, A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Heinle and Heinle.

Cohen, A. D. (1998). Strategies and processes in test taking and SLA. In L.F. Bachman, & A.D. Cohen (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 90-111). Cambridge: Cambridge University Press.

Cohen, A. D. (2006). The coming of age of research on test-taking strategies. *Language Assessment Quarterly*, *3*(4), 307-331.

Cohen, A. D., & Olshtain, C. (1993). The production of speech acts by EFL learners. *TESOL Quarterly*, *27*(1), 33-56.

Coniam, D. (2009). Investigating the quality of teacher-produced tests for EFL

students and the effects of training in test development principles and practices on improving test quality. *System, 37*(2), 226-242.

Davidson, F., & Lynch, B.K. (2002). *Testcraft—A teacher's guide to writing and using language test specifications.* New Haven and London: Yale University Press.

Davies, A. (1997). Demands of being professional in language testing. *Language Testing, 14*(3), 328-339.

Douglas, D. (2000). *Assessing language for specific purposes: Theory and practice*. Cambridge: Cambridge University Press.

Douglas, D., & Seliner, L. (1985). Principles for language tests within the 'discourse domains' theory of interlanguage: Research, test construction and interpretation. *Language Testing*, *2*(2), 205-226.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis: Verbal reports as data*. Cambridge, MA: MIT Press.

Ericsson, K. A., & Simon, H. A. (1993). P*rotocol analysis: Verbal reports as data*. (Rev. ed.). Cambridge, MA: MIT Press.

Faerch, C., & Kasper, G. (1987). From product to process—introspective methods in second language research. In C. Faerch & G.. Kasper (Eds.), *Introspection in second language research* (pp. 5-23). Clevedon: Multilingual Matters Ltd.

Fulcher, G.. (1996). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

Gass, S., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Gierl, M. J. (1997). Comparing cognitive representations of test developers and students on a mathematics test with Bloom's taxonomy. *Journal of Educational Research, 91*(1), 26-32.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

Gruba, P. & Corbel, C. (1997). Computer-based testing. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education*. *Volume 7: Language testing and assessment* (pp. 141-149). Dordrecht: Kluwer Academic.

Hale, G. A. (1988). Student major field and text context: interactive effects on reading comprehension in the Test of English as a Foreign Language. *Language Testing, 5*(1), 49-61.

Haney, W., & Scott, L. (1987). Talking with children about tests: An exploratory study of test item ambiguity. In R. O. Freedle & R. P. Duran (Eds.), *Cognitive and linguistic analyses of test performance* (pp. 298-368). Norwood, NJ: Ablex.

Heaton, J. B. (1988). *Writing English language tests* (New ed.). New York: Longman.

Herman, J. L., Aschbacher, P. R., & Winters, L. (1992). *A practical guide to alternative assessment*. Alexandria, VA: Association for Supervision and Curriculum Development.

Hill, K. (1993). The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 209-230). Mahwah, NJ: Lawrence Erlbaum.

Hudson, T., Detmer, E., Brown, J. D. (1992). *A framework for testing cross-cultural pragmatics*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Hudson, T., Detmer, E., Brown, J. D. (1995). *Developing prototypic measures of cross-cultural pragmatics*. Honolulu: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). New York: Cambridge

University Press.

Jafarpur, A. (2003). Is the test constructor a facet? *Language Testing, 20*(1), 57-87.

Johnson, R., Becker, P., & Olive, F. (1999). Teaching the second-language testing course through test development by teachers-in-training. *Teacher Education Quarterly*, *26*(3), 71-82.

Kirschner, M., Spector-Cohen, E., & Wexler, C. (1996). A teacher education workshop on the construction of EFL tests and materials. *TESOL Quarterly*, *30*(1), 85-111.

Kunnan, A. J. (Ed.). (1998). Special issue: Structural equation modeling. *Language Testing*, *15*(3).

Laufer, B. & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing*, *16*(1), 33-51.

Lay, N. D. S. (1982). Composing processes of adult ESL learners: A case study. *TESOL Quarterly, 16*(3), 406.

Leighton, J. P., Gokiert, R. J., Cor, M. K., & Heffernan, C. (2010). Teachers beliefs about the cognitive diagnostic information of classroom- versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 7-21.

Lim, G.. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28*(4), 543-560.

Lynch, B. K. (1997). In search of the ethical test. *Language Testing*, *14*(3), 315-327.

Lynch, B. K. & Davidson, F. (1997). Criterion referenced testing. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education. Volume 7: Language testing and assessment* (pp. 263-273). Dordrecht: Kluwer Academic.

MacKay, R. (1974). Standardized tests: Objectives/objectified measures of

"competence." In A. V. Cicourel et al. (Eds.), *Language use and school performance* (pp. 218-247). New York: Academic.

McNamara, T. F. (1997). Performance testing. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education*. *Volume 7: Language testing and assessment* (pp. 131-139). Dordrecht: Kluwer Academic.

Moghaddam, S. (2010). Cultural schemata: Iranian students' test-taking processes for cloze tests. *Education, Business and Society: Contemporary Middle Eastern Issues, 3*(3), 188.

Nation, P. (1990). *Teaching and learning vocabulary*. Boston, MA: Heinle and Heinle.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199-215.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*(3), 231-259.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments*. Honolulu: University of Hawaii at Manoa.

Norris, S. P. (1991). Informal reasoning assessment: Using verbal reports of thinking to improve multiple-choice test validity. In J. F. Voss, D. N. Perkins, & J. W. Segal (Eds.), *Informal Reasoning and Education* (pp. 451-472). Hillsdale, NJ: Lawrence Erlbaum Associates.

Oller, J. W. (1979). *Language tests at school*. London: Longman.

Orr, M. (2002). The FCE speaking test: Using rater reports to help interpret test scores. *System, 30*(2), 143-154.

Park, S. (2009). Verbal report in language testing. *The Journal of Kanda University of International Studies, 21*, 287-307.

Pienemann, M., Johnson, J., & Brindley, G. (1988). Constructing an acquisition-based procedure for language assessment. *Studies in Second Language Acquisition,*

10(2), 217-243.

Pollitt, A. (1997). Rasch measurement in latent trait models. In C. Clapham, & D. Corson (Eds.), *Encyclopedia of language and education, Volume 7: Language testing and assessment* (pp. 243-254). Dordrecht: Kluwer Academic.

Pritchard, R. (1990). The effects of cultural schemata on reading processing strategies. *Reading Research Quarterly*, *25*(4), 273-295.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly*, *19*(2), 229-258.

Read, J. (2000). *Assessing vocabulary knowledge and use*. Cambridge: Cambridge University Press.

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing, 13*(2), 173-190.

Ross, S. (1997). An introspective analysis of listener inferencing on a second language listening test. In G.. Kasper & E. Kellerman (Eds.), *Communication strategies: Psycholinguistic and sociolinguistic perspectives*. (pp. 216-237). London: Longman.

Rupp, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing, 23*(4), 441-474.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. *Language Testing*, *17*(1), 85-114.

Shohamy, E. (1997). Testing methods, testing consequences: Are they ethical? *Language Testing*, *14*(3), 340-349.

Stansfield, C. W. (1993). Ethics, standards and professionalism in language testing.

Issues in Applied Linguistics, *4*(2), 15-30.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, March*, 534-539.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing*, *14*(2), 214-231.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing*, *18*(3), 275-302.

Tsai, P. C. (2008). *The effects of types of rhetorical tasks, English proficiency, and writing anxiety on senior high school students' English writing performance*. Unpublished master's thesis, National Taiwan Normal University, Taipei, Taiwan.

Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, *10*(1), 41-69.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2), 197-223.

Weir, C. J. (1990). *Communicative language testing*. New York: Prentice Hall.

Wiggins, G.. (1993). Assessment: Authenticity, context and validity. *Phi Delta Kappan, November*, 200-214.

Yamashita, J. (2003). Processes of taking a gap-filling test: Comparison of skilled and less skilled EFL readers. *Language Testing*, *20*(3), 267-293.

Appendix A: Research Consent Form for Teachers

研究參與同意書（教師版）

背景介紹：本研究為國立臺灣師範大學英語研究所教學組博士生曾繁萍的博士論文計劃。研究旨在瞭解高中英文教師命題過程、高中學生答題過程、以及兩者間的關係。透過高中英文教師的參與和協助，將有助於學界瞭解高中英文教師命題時的考慮要點為何，研究結果將有益於高中英文師資培育及高中英文的課堂教學。

研究過程：本研究將採用質性研究法。若您同意參與本研究，您將要進行三項任務：首先，您需要填寫一份基本資料問卷表。接著，您必須根據研究者提供的文本，在一週之內擬定一份英文學測模擬試題（內含 28 題選擇題）；在出題的過程中，您必須同時進行有聲思考法 (think aloud)，並用錄音筆錄下您的發言（中、英文皆可）。最後，您必須接受研究者對您的訪談。為了略為彌補您參與本研究所花費的時間與體力，您將獲得新台幣 6,000 元的研究參與者費用。

可能導致的副作用或威脅：本研究對參與者無顯著的身心威脅。唯實驗過程為期略長（約一週），可能會使參與者產生若干心理壓力；又或者，參與者對研究方法或錄音器材不甚熟悉，而產生一些焦慮。若您在參與實驗的過程中產生疲憊或負面的情緒，可隨時考慮暫停或終止參與本研究。您暫停或中止參與研究的決定，將不會對您產生任何負面的影響。

機密性：您的問卷填答結果、錄音筆錄下的內容、及訪談的內容將會以保密的態度處理，您的姓名也將由一個研究代號取代。除了有關機構依法調查外，您的隱私將會被小心維護。研究結果即使發表，您的身份仍將保密。因此，您可以放心參與此研究。

您決定參與此項研究計劃是完全自願的，您有不參加的權利。若您對參與此研究的相關權益仍有疑問，可和研究者曾繁萍同學聯絡。

參與者聲明：我已完全瞭解此同意書的內容，並同意自願參與這個研究。

參與者簽名（正楷）：_____

同意書簽署日期：_____　研究預定開始日期：_____

聯絡方式：Cell phone: _____　E-mail：_____

研究者聲明：我保證我本人已經對上述人士解釋過本研究，包括本研究的目的、程序與方法、及參加本研究可能的相關副作用和權益。所有被提出之疑問，均已獲得滿意的答覆。

研究者：國立臺灣師範大學英語研究所教學組　博士生曾繁萍

指導教授：國立臺灣師範大學英語系　程玉秀博士

Appendix B: Background Questionnaire

親愛的老師：您好！

感謝您願意參與本研究。請您仔細閱讀以下問題並回答之。您所提供的資料會謹慎保密，僅供學術研究之用，請放心填寫。謝謝您的協助。

臺師大英研所教學組博士生　曾繁萍敬上

1. 填表日期：_____
2. 姓名：_____
3. 性別：□男　□女
4. 最高學歷：_____ 大學 _____ 系/所
5. 是否修過語言測驗 (Language Testing) 相關課程：□是　□否
6. 高中英文教學年資： _____年
7. 目前任教學校：_____
8. 您是否有學測英文模擬考出題經驗：
   □是 (有_____次，替_____出題)　　□否
9. 您對歷屆的學測英文考題熟悉度為：
   □非常熟悉　　□熟悉　　□不熟悉　　□非常不熟悉
10. 您對學測英文考題的出題方針熟悉度為：
    □非常熟悉　　□熟悉　　□不熟悉　　□非常不熟悉
11. 您自行出題給任教班級學生考試的頻率為：
    □常常，每週都有　□偶爾，有需要才出　□很少，因使用現成的測驗卷
12. 您替校內英文段考出題頻率為：□每次段考均要出題　□每學期至少一次
    □每學年至少一次　□其它 (_____)
13. 貴校一份英文段考試題由幾位老師命題：□一位　□二位　□三位以上
14. 貴校的英文段考試題在給學生施測前，是否有經過審題：□是　　□否
15. 貴校在段考後，是否有作英文段考試題分析：□是　　□否
16. 您出一份英文段考試題，需要多久時間：□一至三天　□一週　□二週
17. 您覺得很難命題的英文選擇題考題有哪些：(可複選)
    □文意字彙　□克漏字　□閱讀測驗　□無
18. 您覺得很容易命題的英文選擇題考題有哪些：(可複選)
    □文意字彙　□克漏字　□閱讀測驗　□無
19. 您覺得自己的英文試題命題技巧：
    □非常好　　□良好　　□尚可　　□有待加強　　□非常不好
20. 請說明：當您被指派要出一份英文段考試題時的感想。

    _____

21. 請說明：您曾經面臨過的英文試題命題困難。

    _____

Appendix C: Feedback Sheet

研究參與回饋單

親愛的_____老師：您好！

感謝您參與我的博士論文研究。為了使我更了解您對參與此次研究的想法及意見，請您撥冗回答以下的問題。您所提供的資料僅供學術研究之用，並會匿名處理，請放心作答，謝謝您的協助。

臺師大英研所教學組博士生　曾繁萍敬上

1. 關於此次命題，整體來看，您覺得較難出題的文本是：
   □Material A　□Material B

2. 關於此次命題，整體來看，您覺得較難出題的題型是：
   □詞彙題　□綜合測驗　□閱讀測驗

3. 關於此次命題，整體來看，您覺得較容易出題的題型是：
   □詞彙題　□綜合測驗　□閱讀測驗

4. 關於此次命題，您覺得最難出題的部份是：
   □Material A 的詞彙題 □Material A 的綜合測驗 □Material A 的閱讀測驗
   □Material B 的詞彙題 □Material B 的綜合測驗 □Material B 的閱讀測驗

5. 關於此次命題，您覺得最容易出題的部份是：
   □Material A 的詞彙題 □Material A 的綜合測驗 □Material A 的閱讀測驗
   □Material B 的詞彙題 □Material B 的綜合測驗 □Material B 的閱讀測驗

6. 此次命題時，您同步使用了有聲思考法 (think aloud)。請敘述該方法對您命題時所產生的影響，以及您對該方法運用於命題時的可行性或看法。

   _____

   _____

   _____

   _____

   _____

7. 請敘述您對參與此研究的一些心得或看法。

   _____

   _____

   _____

   _____

Appendix D: Shortened Version of FLPT

本測驗共有三大題，40 小題，答錯不倒扣，總分 40 分。作答時間為 30 分鐘。請同學看清楚題目說明之後，直接於題目卷上作答。

第一大題　字彙測驗

(I) 以下 1－5 題每題含一空格，請由四個備選答案中，選一個填入空格後可使該句意思完整的字彙作答。

1. Many tourists like to buy _____ to remember places they have visited.
   (A) souvenirs (B) rewards (C) champions (D) monuments
2. Speech, however courteous, may _____ an attitude of hostility to a sensitive ear.
   (A) conspire (B) convene (C) convey (D) contrive
3. They should _____ the committee members to see how much support they have.
   (A) deduce (B) nourish (C) breed (D) poll
4. Some companies pay their employees a _____ after they retire.
   (A) commission (B) deposit (C) pension (D) fee
5. The antique stand had been _____ carved, showing a scene of rising phoenixes and dragons.
   (A) elaborately (B) literally (C) scholarly (D) principally

(II) 以下 6－10 題每題含一劃線部份，請由四個備選答案中，選一個最符合該畫線字或詞之意義者來作答。

6. These days it is not necessary to <u>put up with</u> such long hours and low pay.
   (A) bear (B) arrange (C) pursue (D) connect
7. The child has been <u>brought up</u> very properly.
   (A) lifted (B) treated (C) raised (D) dressed
8. <u>Cross out</u> the last two names on the list, will you?
   (A) Delete (B) Print (C) Join (D) Circle
9. Jessie <u>declined</u> Bob's offer to give her a ride home.
   (A) forgot (B) refused (C) accepted (D) overlooked
10. In clothes, Steve's <u>preferences</u> run to bright colors and simple patterns.
    (A) positions (B) tastes (C) routines (D) adjustments

第二大題　文法測驗

第 11－30 題：請從四個備選答案中，選一個最適合一般情況下的字或詞作答。

11. Let’s get something to eat. It’s almost _____.
(A) eating dinner time (B) dinner time
(C) time of dinner (D) dinner’s time
12. We were rudely _____ by our upstairs neighbor at two o’clock in the morning.
(A) waked (B) awake
(C) wake up (D) awakened
13. If I got a promotion, I _____ treat you all to a fancy dinner.
(A) shall (B) will
(C) ought (D) would
14. Under the pressure from his colleagues, he was made _____ that he agreed with the abandonment of the plan.
(A) say (B) saying
(C) to say (D) by saying
15. You need to fill _____ these forms to obtain a visa.
(A) on (B) for
(C) out (D) with
16. Look! The kitchen door was forced open. The burglars _____ have gotten in from here.
(A) must (B) are supposed to
(C) ought (D) were able to
17. According to a recent report, living conditions for _____ have not improved yet.
(A) poors (B) a poor
(C) the poor (D) the poors
18. Jack _____ play tennis every weekend before the injured his arm.
(A) was used to (B) got used to
(C) became used to (D) used to
19. Many firms try to survive by lowering costs and _____ productivity.
(A) rising (B) raising
(C) arising (D) arousing
20. We _____ a seminar this afternoon. The presenter is sick.
(A) maybe don’t have (B) might not have
(C) might have no (D) maybe not have
21. This dress _____ me NT$3,000.
(A) cost (B) is cost
(C) spent (D) is spent
22. It’ll be _____ if you use a calculator to add these numbers up.
(A) much easier (B) more easy
(C) much more easy (D) much more easier

23. ____ for my parents' support, I could never have graduated from medical school.
(A) It had not been (B) Had it not been
(C) Would it have been (D) It would have been

24. Let's begin with _____.
(A) lesson tenth (B) the lesson tenth
(C) the lesson ten (D) lesson ten

25. A recession like ____ of 1974 would put many small businesses into bankruptcy.
(A) what (B) that
(C) which (D) one

26. _____ so long?
(A) Does your hair always have been (B) Has your hair always been
(C) Have your hair always been (D) Has been your hair always

27. At the moment, the workshop _____ in the conference room.
(A) held (B) is hold
(C) is holding (D) is being held

28. Excuse me, I have to _____ to my client.
(A) make a call (B) make a phone
(C) call a phone (D) telephone a call

29. When the hostess introduced Jack to his ex-wife, Jack acted as though he _____ her before.
(A) never meets (B) was never met
(C) had never met (D) has never met

30. The suspension bridge shook violently ____ the heavy storm swept over the area.
(A) meantime (B) during
(C) as (D) through

第三大題　閱讀測驗

本大題共有十題，含三篇短文，每篇文章之後有幾個相關的問題，每題均有四個備選答案，請選一個最適合者作答。

Questions 31－32

A new way has been found to reduce the damage fruit flies cause to starfruit grown in Malaysia. A chemical has been developed that attracts fruit flies. This chemical is mixed with insecticide and then put on some of the leaves of the starfruit plants. The flies, when they smell the chemical, go to the leaves and are then killed by the insecticide. As a result, there are fewer fruit flies and less damage is caused to the starfruit crop.

31. Why do farmers in Malaysia need this new chemical?
    (A) Fruit flies are harming their starfruit.
    (B) The starfruit is a danger to the fruit flies.
    (C) They cannot use the old insecticides.
    (D) They need the chemical to protect the leaves.
32. How does the chemical help to protect the starfruit?
    (A) It is a kind of poison that causes less damage to fruit flies.
    (B) It keeps the flies away from the starfruit plants.
    (C) It helps the starfruit to grow faster.
    (D) It attracts the flies to the insecticide, which kills them.

Questions 33－36

Despite the fact that malaria has been eradicated in many parts of the world, 270 million cases of this disease appear annually. In the past, chloroquine has been successfully used to treat malaria patients while DDT has been used to control the mosquitoes that carry the disease, but scientists are now faced with malaria that are resistant to chloroquine and mosquitoes undeterred by DDT.

The challenge now is to develop a malaria vaccine. This vaccine will probably be a combination of several antigens, each of which is capable of identifying the malaria intruder and stimulating the body's immune system to produce protective antibodies in response. It is unlikely, however, that this new vaccine will be ready for use in the near future.

33. Why do doctors have difficulty treating malaria today?
    (A) Chloroquine is no longer as effective as in the past in the treatment of malaria.
    (B) There are too many new cases of malaria to handle.
    (C) Few doctors are very experienced in treating this disease.
    (D) Malaria is a new disease that has been researched very little.
34. How do doctors hope to be able to treat malaria in the future?
    (A) With a more effective type of chloroquine.
    (B) With a new vaccine.
    (C) With parasites less resistant to chloroquine.
    (D) With more widespread use of DDT.
35. According to this passage, what is true of malaria?
    (A) It will probably be brought under control soon.
    (B) 270 million people have died from this disease.
    (C) In much of the world, it has disappeared.
    (D) It is a disease which infects mosquitoes but is of no danger to man.

36. It a vaccine is developed, how will it work?
    (A) It will stimulate the mosquitoes' immune system.
    (B) It will help the body protect itself against malaria.
    (C) It will use antibodies to fight malaria.
    (D) It will produce protective antigens in the body's immune system.

Questions 37－40

The "balance of nature" is not an empty phrase. Nature provides a population to occupy a suitable environment and cuts down surplus population to fit the available food supply. One means of reducing surplus population is predators; other are parasites and diseases. Also, high population density produces nervous disorders and even drives animals to mass migrations, like the lemmings of Norway who plunge into the sea.

That predator populations mount to control other animals has long been known. Many years ago, it was observed that the fox population living around Hudson Bay went up and down about a year after the rabbit population had gone up and down.

37. What is a "predator"?
    (A) Lemmings that are found in Norway.
    (B) A disease used to control animal populations.
    (C) A parasite that infects animals.
    (D) An animal that kills other animals.
38. How does nature bring itself into balance?
    (A) By decreasing the animal population.
    (B) By increasing the food supply.
    (C) By decreasing the food supply.
    (D) By increasing the animal population.
39. What does the phrase "balance of nature" mean?
    (A) The relationship of wildlife to man.
    (B) The adequacy of the food supply to support its animal population.
    (C) The ratio of small game to predators.
    (D) The destruction of predators.
40. According to this passage, what is the effect of high population density among some animals?
    (A) It automatically guides their rate of reproduction.
    (B) Population density does not really affect them.
    (C) It produced nervous disorders.
    (D) Population density among some animals never occurs.

Appendix E: Research Consent Form for Students

研究參與同意書（學生版）

背景介紹：本研究為國立臺灣師範大學英語研究所教學組博士生曾繁萍的博士論文計劃。研究旨在瞭解高中英文教師命題過程、高中學生答題過程、以及兩者間的關係。透過高中學生的參與和協助，將有助於學界瞭解高中學生在回答英文試題時所考慮的要點及所用的策略為何，研究結果將有益於高中英文師資培育及高中英文的課堂教學。

研究過程：本研究將採用質性研究法。若您同意參與本研究，您將要進行兩項任務：首先，您必須要在一個小時內作答完一份英文學測模擬試題（內含 28 題選擇題）；在答題過程中，您必須同時進行有聲思考法 (think aloud)，並用錄音筆錄下您的發言（中、英文皆可）。緊接著，您必須回答研究者針對您的答題過程所問的問題。為了略為彌補您參與本研究所花費的時間與體力，您將獲得新台幣 100 元的研究參與者費用。

可能導致的副作用或威脅：本研究對參與者無顯著的身心威脅。唯參與者將作答一份英文學測模擬試題，並同步進行有聲思考法 (think aloud)，參與者可能會知覺到任務的困難，而產生若干心理壓力或焦慮。若您在參與實驗的過程中產生疲憊或負面的情緒，可隨時考慮暫停或終止參與本研究。您暫停或中止參與研究的決定，將不會對您產生任何負面的影響。

機密性：您的英文學測模擬試題作答結果、錄音筆錄下的內容、及訪談的內容將會以保密的態度處理，您的姓名也將由一個研究代號取代。除了有關機構依法調查外，您的隱私將會被小心維護。研究結果即使發表，您的身份仍將保密。因此，您可以放心參與此研究。

您決定參與此項研究計劃是完全自願的，您有不參加的權利。若您對參與此研究的相關權益仍有疑問，可和研究者曾繁萍同學聯絡。

參與者聲明：我已完全瞭解此同意書的內容，並同意自願參與這個研究。

參與者簽名（正楷）：_____

同意書簽署日期：_____　研究預定開始日期：_____

聯絡方式：Cell phone: _____　E-mail：_____

研究者聲明：我保證我本人已經對上述人士解釋過本研究，包括本研究的目的、程序與方法、及參加本研究可能的相關副作用和權益。所有被提出之疑問，均已獲得滿意的答覆。

研究者：國立臺灣師範大學英語研究所教學組　博士生曾繁萍

指導教授：國立臺灣師範大學英語系　程玉秀博士

Appendix F: Materials for Test Construction

Material A

◎請根據下列的單字表，出 5 題「詞彙」題 (vocabulary items)。

1. absolute	11. continuously	21. figure	31. obviously
2. accurate	12. contract	22. frustrate	32. optimistic
3. addition	13. current	23. hesitation	33. practical
4. anxiety	14. decrease	24. immediately	34. recite
5. briefly	15. directly	25. influence	35. resist
6. candidate	16. disaster	26. interpret	36. routine
7. capacity	17. dominant	27. loosen	37. scarcely
8. client	18. errand	28. measure	38. squeeze
9. competitive	19. famously	29. miracle	39. transfer
10. considerably	20. fearful	30. obtain	40. violate

◎請根據以下文章，出 5 題「綜合測驗」題 (cloze items)。

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. In addition, good listeners are inclined to accept or tolerate rather than to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are exceptions to that generality. For example, John Steinbeck is said to have been an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen contributed to his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, depending on what a good listener does, he may become either popular or disliked in his lifetime.

◎請根據以下文章，出 4 題「閱讀測驗」題 (reading comprehension items)。

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look much. It is a woman’s shoe of a style popular

in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100, 000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

Material B

◎請根據下列的單字表，出 5 題「詞彙」題 (vocabulary items)。

1. amaze	11. division	21. narrowly	31. revise
2. approach	12. exception	22. occupation	32. rumor
3. average	13. expressively	23. originally	33. sincere
4. barrier	14. freeze	24. overthrow	34. solution
5. boast	15. gossip	25. preserve	35. technique
6. collapse	16. gradually	26. promising	36. temporary
7. comfortable	17. hardly	27. propose	37. totally
8. confess	18. legend	28. rapidly	38. urgently
9. contest	19. liberal	29. relative	39. vision
10. dealer	20. maintain	30. reluctant	40. voluntary

◎請根據以下文章，出 5 題「綜合測驗」題 (cloze items)。

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. For example,

in the United States, it is very important to be on time for almost all occasions. The only time it is socially acceptable to be late is when going to a friend's party. A person usually tried to arrive about five minutes after the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being "fashionably late." Any time later than that is considered impolite, because it keeps the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would rush the host. However, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be filled out by the patient.

◎請根據以下文章，出 4 題「閱讀測驗」題 (reading comprehension items)。

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a conservative dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee morale. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

Appendix G: Four Forms of the SAET Mock Tests

學測英文模擬試題 (Form A)

一、詞 彙 題

說明：第 1 題至第 10 題，每題有 4 個選項，其中只有一個是正確或最適當的選項。

1. Despite continual misfortunes happening to me, I still feel _____ about my future.
   (A) dominant (B) competitive (C) optimistic (D) practical
2. Instead of telling me _____, Judy wrote me that she didn't love me anymore.
   (A) briefly (B) considerably (C) immediately (D) directly
3. As a good Taiwan citizen, we should _____ whatever is against the law.
   (A) violate (B) frustrate (C) resist (D) decrease
4. Most people enjoy watching famous paintings even though they can hardly _____ the painter's thoughts.
   (A) transfer (B) interpret (C) measure (D) frustrate
5. Doing all the housework for Mom is my weekly _____. I do so every Sunday morning.
   (A) errand (B) capacity (C) routine (D) contract
6. Without warning, the bridge _____, causing the cars to fall into the river. The tragedy was attributed to the erosion of the foundation and a lack of maintenance.
   (A) suspended (B) collapsed (C) constructed (D) extended
7. According to the investigation, the language _____ caused the air crash. The misunderstanding of the pilot's spoken English was to blame.
   (A) barrier (B) acquisition (C) requirement (D) resistance
8. The demanding teacher always gives students a lot of homework and tests. _____, students complain of that. In the end, they adapt to the way he teaches.
   (A) Tentatively (B) Alternatively (C) Unfortunately (D) Originally
9. School on a tight budget had difficulty continuing hiring guards for the campus safety and some generous parents responded to that by donating money on a _____ basis.
   (A) violent (B) hostile (C) voluntary (D) reluctant
10. The English teacher _____ the English composition and then asked the students to completely understand the corrected part.
    (A) revised (B) released (C) acclaimed (D) interpreted

二、綜 合 測 驗

說明：第 11 題至第 20 題，每題一個空格，請依文意選出最適當的一個選項。

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them __11__ other people. __12__, good listeners are inclined to accept or tolerate rather than to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are exceptions __13__ that generality. For example, John Steinbeck is said to __14__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen contributed to his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, depending on __15__ a good listener does, he may become either popular or disliked in his lifetime.

11. (A) as (B) with (C) than (D) about
12. (A) In fact (B) In addition (C) In short (D) In other words
13. (A) for (B) with (C) about (D) to
14. (A) have been (B) be (C) has been (D) become
15. (A) that (B) what (C) which (D) how

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. For example, in the United States, it is very important to be on time for almost all __16__. The only time it is socially acceptable to be late is when going to a friend's party. A person usually tried to arrive about five minutes __17__ the invitation time, so that the host would have a little extra time to prepare for the guests. __18__ is called being "fashionably late." Any time later than that is considered impolite, because it keeps the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would rush the host. __19__, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be __20__ by the patient.

16. (A) phenomena (B) excursions (C) factors (D) occasions
17. (A) before (B) as soon as (C) after (D) as early as
18. (A) It (B) There (C) This (D) What
19. (A) Instead (B) Likewise (C) Therefore (D) However
20. (A) filled out (B) applied for (C) filled with (D) taken over

三、閱 讀 測 驗

說明：第 21 題至第 28 題，每題請分別根據各篇文章之文意選出最適當的一個選項。

第 21 至 24 題爲題組

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

21. Why does the author say the shoe is special?
    (A) The shoe belongs to a woman.
    (B) The shoe is popular in the 1890s.
    (C) It was found on the trail which people seeking gold would pass.
    (D) It's made of leather and stored in a museum in Alaska.
22. What do we know about the shoe?
    (A) It looks ordinary just like any other shoe.
    (B) It was thrown away because its owner no longer needed it.
    (C) It was accidentally dropped by a woman climbing up stairs.
    (D) Its owner died of starvation.
23. According to the article, what is true about "gold fever?"
    (A) Many gold seekers died of hunger and dry weather.
    (B) Whoever wanted to seek gold had to carry one ton of necessities with them.
    (C) A successful gold seeker must have a pair of expensive shoes.

(D) The gold seekers' backpacks weighed no more than forty pounds.

24. What can we infer about the owner of the shoe?

(A) She took no more than thirty trips in order to carry her supplies.

(B) She had to endure the humid temperature for one year.

(C) She was a brave gold seeker.

(D) She hurt herself and dropped her shoe on her journey.

第 25 至 28 題爲題組

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a conservative dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee **morale**. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

25. Which of the following is the best title for the article?

(A) The Origin of Casual Friday

(B) How to Raise the Efficiency of Office Workers

(C) From Formal Wear to Casual Clothes

(D) The Dilemma between Suits and Jeans

26. Which of the following statements is true of "casual Friday?"

(A) It was also dubbed “dress-down Friday.”
(B) In the early 1990, employees were allowed to wear casually in all the companies in America.
(C) This dress code became an immediate hit.
(D) It started out wearing casual clothes every day.

27. According to the passage, why have a lot of companies adopted a casual dress code?
(A) Young hires enjoy the casual working environment.
(B) The companies cannot afford to buy suits for their employees.
(C) Adopting less traditional dress code is extremely crucial to the recruitment of new employees.
(D) The employees work more efficiently in a good mood.

28. Which of the following words is synonymous with the word “**morale**?”
(A) integrity (B) productivity (C) happiness (D) enthusiasm

學測英文模擬試題 (Form B)

一、詞 彙 題

說明：第 1 題至第 10 題，每題有 4 個選項，其中只有一個是正確或最適當的選項。

1. Mr. Wang is one of my father’s _____, who constantly seeks legal advice at Father’s law firm.
(A) miracles (B) clients (C) routines (D) errands

2. Taiwan Railway Administration is working hard to make their trains’ arrival time _____ so as to win people’s trust back.
(A) accurate (B) current (C) practical (D) dominant

3. Informed of his admission to his ideal university, Patrick could _____ control his joy and let out a cry.
(A) obviously (B) continuously (C) briefly (D) scarcely

4. President Ma Ying-jeou calls for reforms, and thus takes immediate _____ to put his policy into practice.
(A) figures (B) influences (C) contracts (D) measures

5. Sandy’s father was assigned to run a branch company in Mainland China; thus, she had no choice but to _____ to another school there.
(A) violate (B) recite (C) transfer (D) frustrate

6. The father held a _____ attitude on his daughter’s marriage. He let her decide who would be the one she could rely on for life.

(A) comfortable (B) promising (C) liberal (D) sincere

7. The student who had overslept _____ caught the school bus with her long hair clipped by the bus door.
(A) originally (B) expressively (C) totally (D) narrowly

8. With his parents' patience and company for years, the retarded child _____ catch up with his classmates.
(A) rapidly (B) gradually (C) hardly (D) urgently

9. The _____ to the regulation on speed limit is allowed when the ambulance is on its duty.
(A) exception (B) approach (C) solution (D) technique

10. If good friends misunderstand each other, their friendship will be difficult to ____.
(A) preserve (B) freeze (C) maintain (D) confess

二、綜 合 測 驗

說明：第 11 題至第 20 題，每題一個空格，請依文意選出最適當的一個選項。

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. __11__ they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. __12__ , good listeners are inclined to accept or tolerate rather than to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are exceptions to that generality. For example, John Steinbeck __13__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen __14__ his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, depending on __15__ a good listener does, he may become either popular or disliked in his lifetime.

11. (A) Because (B) Unless (C) If (D) Despite
12. (A) In consequence (B) In the end (C) In contrast (D) In addition
13. (A) was portrayed as (B) was told to pretend
(C) is prone to be (D) is said to have been
14. (A) contributed to (B) originated from (C) involved in (D) substituted for
15. (A) that (B) what (C) which (D) how

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. __16__, in the

United States, it is very important to be on time for almost all occasions. The only time it is __17__ acceptable to be late is when going to a friend's party. A person usually tried to arrive about five minutes __18__ the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being "fashionably late." Any time later than that is considered impolite, because it __19__ the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would rush the host. __20__, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be filled out by the patient.

16. (A) Therefore (B) For example (C) However (D) At first
17. (A) hardly (B) wisely (C) socially (D) directly
18. (A) after (B) before (C) on (D) by
19. (A) makes (B) forces (C) lets (D) keeps
20. (A) Actually (B) However (C) In conclusion (D) Unfortunately

三、閱 讀 測 驗

說明：第 21 題至第 28 題，每題請分別根據各篇文章之文意選出最適當的一個選項。

第 21 至 24 題爲題組

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska.

Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

21. What is the thesis statement of this article?
    (A) Alaska museum's talking object drew many people.
    (B) Bring enough supplies when going on adventures.
    (C) Magnificent stories might lay behind silent objects.
    (D) Gold fever took away people's lives due to the weather.
22. Where was the old leather shoe first located?
    (A) In a national museum. (B) On an icy mountain trail.
    (C) On a display shelf in a store. (D) In an underground gold mine.
23. What is **false** about people with "gold fever?"
    (A) Many failed because their supplies were not sufficient.
    (B) They went with a dream to become rich.
    (C) Quite a few were unaware of what they would come across.
    (D) They had to pack light in order to travel fast.
24. What can we infer about the shoe's owner from this article?
    (A) She was a person with great courage.
    (B) She was in the business of trading gold.
    (C) She was a fashion queen in the 19th century.
    (D) She lost her shoe while carrying supplies back and forth.

第 25 至 28 題爲題組

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a conservative

dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee morale. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

25. What is the best title for this article?
    (A) The One-day-a-week Casual Friday
    (B) The Reasons for Wearing Casual Clothes to Work
    (C) The Comparison Between Suits and Casual Clothes
    (D) All Employers Love Casual Clothes
26. Which statement is correct?
    (A) It's difficult to recruit new workers if there's a dress code in a company.
    (B) In the early 1990s, employers can wear whatever they like to work every day.
    (C) David Smith works for a software company.
    (D) David Smith works harder when he works in a suit.
27. According to the research by Levi Strauss, how do most employers think about dress code?
    (A) It makes their employees less productive.
    (B) It's practicable only on Friday.
    (C) It helps their employees work efficiently in a good mood.
    (D) It has a negative impact on their productivity.
28. Which of the following statement is **NOT** the reason most employees love casual dress?
    (A) It's comfortable to wear casual dress.
    (B) Their bosses prefer casual dress rather than suits.
    (C) Casual clothes are much cheaper than suits.
    (D) It makes them work more happily and productively.

學測英文模擬試題 (Form C)

一、詞 彙 題

說明：第 1 題至第 10 題，每題有 4 個選項，其中只有一個是正確或最適當的選項。

1. To prevent people from drunk driving, the government should take the necessary

_____ to punish those who drive cars after they have drinks.
(A) contract (B) errand (C) influence (D) measure

2. Because of the _____ of birth rate, the country is now facing the problem of having fewer and fewer young people to support it.
(A) decrease (B) addition (C) capacity (D) routine

3. In nature, the _____ animal is usually the largest and most powerful creature that other animals would not fight with.
(A) fearful (B) optimistic (C) dominant (D) practical

4. In order to _____ more information on the issue, the students decide to go to the library to find the books that will be helpful to them.
(A) resist (B) obtain (C) transfer (D) loosen

5. Although the rain is light, if it falls _____, there will be a flood hitting the town.
(A) briefly (B) scarcely (C) immediately (D) continuously

6. With the release of its new smart phones, the manufacturer Nokia _____ 160% more app downloads than Apple.
(A) boasts (B) revises (C) maintains (D) approaches

7. Well goes the saying “A distant _____ is not as good as a near neighbor.” That is, good neighbors are a lot more helpful when we are in need.
(A) vision (B) legend (C) dealer (D) relative

8. While _____ and situational loneliness can be a normal, healthy part of life, chronic loneliness can be a very sad and sometimes very dangerous condition.
(A) reluctant (B) liberal (C) temporary (D) sincere

9. Faced with a grave danger of our drinking water, we all _____ need to apply a solution to this problem lest our health be affected day after day by industrial waste in the river.
(A) narrowly (B) rapidly (C) hardly (D) urgently

10. The suspect didn’t _____ her crime to the murder until she saw the videotape.
(A) propose (B) confess (C) collapse (D) gossip

二、綜 合 測 驗

說明：第 11 題至第 20 題，每題一個空格，請依文意選出最適當的一個選項。

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. __11__, good listeners are inclined to accept or tolerate than to judge and criticize. Therefore, they have __12__ enemies than other people. In fact, they are probably the most loved of people. However, there are __13__ to that generality. For

example, John Steinbeck is said to __14__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen __15__ his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, depending on what a good listener does, he may become either popular or disliked in his lifetime.

11. (A) In contrast (B) As a result (C) In addition (D) For that reason
12. (A) few (B) fewer (C) little (D) less
13. (A) possibilities (B) reasons (C) inclusions (D) exceptions
14. (A) have been (B) has been (C) had been (D) having been
15. (A) resulted from (B) contributed to (C) consisted of (D) persisted in

People have different ideas about what exactly is being on time and being late. 16__ ideas also differ from time to time, and from country to country. For example, in the United States, it is very important to be on time for almost all occasions. The only time __17__ is socially acceptable to be late is when going to a friend's party. A person usually tried to arrive about five minutes after the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being "fashionably late." Any time later than that is considered impolite, because it __18__ the host and other guests waiting.

Being on time __19__. One should also not arrive early for a friend's party, because it would rush the host. However, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be __20__ by the patient.

16. (A) These (B) Other (C) Theirs (D) Our
17. (A) that (B) what (C) which (D) it
18. (A) keeps (B) makes (C) has (D) gets
19. (A) goes neither way (B) goes both ways (C) goes one way (D) goes either way
20. (A) figured out (B) set out (C) filled out (D) sent out

三、閱 讀 測 驗

說明：第 21 題至第 28 題，每題請分別根據各篇文章之文意選出最適當的一個選項。

第 21 至 24 題爲題組

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was

discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

21. What is the purpose of this article?
    (A) To inform readers of the stories behind an ordinary object.
    (B) To introduce a famous collection of objects in a museum of Alaska.
    (C) To discourage readers from seeking gold because it is very dangerous.
    (D) To convince readers that the woman who dropped the shoes must be a millionaire.
22. According to the second paragraph, which of the following is **NOT TRUE** about the gold-seeking journey in Alaska?
    (A) People who took part in this journey were determined to be rich.
    (B) The journey required its participants to carry necessities with them.
    (C) The gold seekers knew the dangers before they embarked on the journey.
    (D) There were challenges, such as lack of food and harsh weather, awaiting the gold seekers.
23. What can we infer from this passage?
    (A) The woman brought with her the supplies which weighed over 40 pounds.
    (B) The woman who joined the gold-seeking trip followed the crowd in clothing.
    (C) The woman knew the journey was full of dangers when she decided to make it.
    (D) The woman could not stand the cold weather because she lost the leather shoe.
24. Why did the author mention the old leather shoe?
    (A) To discuss the trend of shoe wearing in the 1890s.
    (B) To prove that men were not the only participants to seek gold.
    (C) To introduce the Chilkoot Pass, the most dangerous site in Alaska.

(D) To show that a common thing like this may have some tales to tell.

第 25 至 28 題爲題組

Five years ago, David Smith wore an expensive suit to work every day. “I was a clothes addict,” he jokes. “I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled.” Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. “I’m working harder than ever,” David says, “and I need to feel comfortable.”

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as “dress-down Friday” or “casual Friday.” “What started out as an extra one-day-a-week benefit for employees has really become an everyday thing,” said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it’s easier for a company to attract new employees if it has a casual dress code. “A lot of young people don’t want to dress up for work,” says the owner of a software company, “so it’s hard to hire people if you have a conservative dress code.” Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that **casual dress improves employee morale**. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. “Suits are expensive, if you have to wear one every day,” one person said. “For the same amount of money, you can buy a lot more casual clothes.”

25. David Smith is a(n) _____.
    (A) tailor (B) consultant (C) employee (D) employer
26. Those who believe that **casual dress improves employee morale** have faith in _____ if they let their employees dress comfortably to work.
    (A) family harmony (B) productivity and creativity
    (C) luck in the jackpot (D) earning money by saving
27. According to the article, what is **NOT** an advantage of wearing casually to office?
    (A) Staying conservative. (B) Being economical.
    (C) Feeling at ease. (D) Boosting innovation.
28. Which of the following is the author’s opinion toward office dress code?

(A) It helps companies to recruit more new blood to work for them.
(B) The author supports that companies should allow a casual dress code.
(C) The author considers wearing formally is good to a company's image.
(D) The author did not state his or her own opinion on this issue.

學測英文模擬試題 (Form D)

一、詞彙題

說明：第 1 題至第 10 題，每題有 4 個選項，其中只有一個是正確或最適當的選項。

1. English is a(n) _____ language, serving as a necessary tool to communicate with people with diverse nationalities.
(A) current (B) dominant (C) accurate (D) fashionable
2. Human beings are not the only at risk of _____ the flu this season. The furry friends can fall ill as well.
(A) contracting (B) combating (C) separating (D) prolonging
3. Despite its _____ of no more than 12 passengers, the van had 16 students in it.
(A) capability (B) evaluation (C) measurement (D) capacity
4. Recently, with oil and electricity prices going up, the commodity prices increased _____. Thus, more and more people barely make both ends meet.
(A) slightly (B) considerably (C) roughly (D) tentatively
5. It is of significant importance for drivers to have the _____ maintenance of their cars.
(A) hesitant (B) temporary (C) routine (D) reckless
6. Since it is Chinese Lunar New Year, the kids persuade their _____ parents to buy the toys for them as gifts.
(A) liberal (B) sincere (C) reluctant (D) voluntary
7. They _____ planned to watch that film, but the tickets sold out, so they saw this film instead.
(A) gradually (B) originally (C) hardly (D) urgently
8. Miranda filled in the blank of _____ with the word "nurse," which means she worked as a nurse.
(A) occupation (B) relative (C) technique (D) division
9. Feeling guilty, the naughty boy who broke the window finally _____ that he did it and apologized for his wrongdoing.
(A) confessed (B) boasted (C) rumored (D) proposed
10. Since there are fewer and fewer people speaking the language, people from the

tribe are trying to _____ it in case it will disappear in the future.

(A) maintain (B) overthrow (C) revise (D) preserve

二、綜 合 測 驗

說明：第 11 題至第 20 題，每題一個空格，請依文意選出最適當的一個選項。

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should __11__. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. In addition, good listeners are inclined to accept or tolerate __12__ to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are __13__ to that generality. For example, John Steinbeck is said __14__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen contributed to his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, __15__ what a good listener does, he may become either popular or disliked in his lifetime.

11. (A) despise (B) treasure (C) ignore (D) command
12. (A) as well as (B) instead of (C) regardless of (D) rather than
13. (A) attachments (B) addiction (C) exceptions (D) expectations
14. (A) being (B) to be (C) to have been (D) to have
15. (A) depending on (B) speaking of (C) compared with (D) different from

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. __16__, in the United States, it is very important to be on time for almost all occasions. The only time it is socially acceptable to be late is __17__ going to a friend's party. A person usually tried to arrive about five minutes after the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being "fashionably late." Any time later than that is considered __18__, because it keeps the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would __19__ the host. However, when going to a doctor's appointment, it is usually good to arrive __20__ than the appointment because there are usually forms that need to be filled out by the patient.

16. (A) In addition (B) For example (C) As a result (D) Even so
17. (A) how (B) why (C) when (D) where

18. (A) friendly (B) hostile (C) agreeable (D) impolite
19. (A) rush (B) delay (C) grip (D) free
20. (A) later (B) earlier (C) quicker (D) slower

三、閱 讀 測 驗

說明：第 21 題至第 28 題，每題請分別根據各篇文章之文意選出最適當的一個選項。

第 21 至 24 題爲題組

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. **This** was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

21. What is the main idea of the passage?
    (A) The history of gold fever.
    (B) Traveling light is crucial to seeking gold.
    (C) Shoes are the perfect option on display.
    (D) Daily stuff, like a shoe, can tell us an amazing story.
22. Which of the following is true about Chilkoot Pass?
    (A) It was the shortcut to the gold mine.
    (B) It was the celebrated path for gold seekers.
    (C) It was carved out of ice.
    (D) It, located on the field, emerged from gold fever.

23. According to the passage, what caused many of gold seekers to die?
(A) A lack of shoes. (B) Carrying heavy backpacks.
(C) Inadequate preparation. (D) Sudden cold weather.

24. In the fifth line of the second paragraph, what does "**this**" refer to?
(A) The requirement for the gold seekers.
(B) The greatest adventures of the shoe owner.
(C) The Canadian government.
(D) Gold fever.

第 25 至 28 題爲題組

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a **conservative** dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee morale. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

25. Which of the following statements about the study by Levi Strauss and Company is **NOT TRUE**?
(A) 15% of employers think casual wear will make employees unwilling to work.
(B) 85% of employees believe that they are in a bad mood for working in casual

wear.

(C) 4% of employers argue that employees will produce less when they wear casual clothes.

(D) Those who welcome the policy of casual wear think they will save money by buying casual clothes instead of formal ones.

26. Which of the following best describe David Smith?

(A) He cannot go out without a necktie on him.

(B) He thinks suits will make him look more handsome.

(C) He wore formal clothes to work every day five years ago.

(D) He now finds his work boring and feels tiresome to work.

27. What does "**conservative**" in the fourth line of the third paragraph refer to?

(A) Resistant to accept new ideas.

(B) Addicted to extraordinary things.

(C) Eager to preserve cultural heritage.

(D) Sensitive to the right time to do the right thing.

28. What is this article mainly about?

(A) The most attractive office wearing style.

(B) The naming of "casual Friday" and its dress code.

(C) Formal suits and their beneficial effects on the firm.

(D) Casual wear and its positive influence on employees.

Appendix H: Dates of Data Collection

Dates for Collecting Teachers' Verbal Report Data

	Dates of Data Collection	
Participant	Beginning of the task	End of the task
ET 1	Jan, 16, 2013	Feb. 3, 2013
ET 2	Jan. 18, 2013	Feb. 7, 2013
NT 1	Jan. 16, 2013	Feb. 6, 2013
NT 2	Jan. 17, 2013	Feb. 13, 2013

Dates for Collecting Students' Verbal Report Data

Participant	Date & Starting time		Participant	Date & Starting time	
H01A	3/07	12:00	L01A	5/07	13:50
H02B	3/14	16:40	L02B	5/02	08:10
H03C	3/19	16:40	L03C	4/25	10:10
H04D	3/12	16:40	L04D	5/30	09:10
H05D	3/13	12:00	L05D	6/06	09:10
H06C	3/15	12:00	L06C	5/30	16:40
H07B	3/20	12:00	L07B	6/11	16:40
H08A	3/14	12:00	L08A	6/10	16:40
H09A	3/21	12:00	L09A	4/30	12:50
H10B	3/21	16:40	L10B	6/04	16:40
H11C	4/25	08:10	L11C	6/06	16:40
H12D	3/11	16:40	L12D	6/20	12:00
H13D	5/22	09:10	L13D	6/18	16:40
H14C	4/09	12:00	L14C	5/28	16:40
H15B	4/25	12:00	L15B	6/19	09:10
H16A	5/29	09:10	L16A	4/11	09:10
H17A	5/23	09:10	L17A	5/02	11:05
H18B	6/05	09:10	L18B	6/17	16:40
H19C	6/06	12:00	L19C	6/03	16:40
H20D	5/22	12:00	L20D	6/13	16:40
H21D	5/29	12:00	L21D	5/01	12:00
H22C	5/09	08:10	L22C	4/10	09:10
H23B	5/23	12:00	L23B	5/08	09:10
H24A	5/30	12:00	L24A	6/20	16:40

Note. The data procedures all carried out between March through June in the year 2013.

Appendix I: Teacher-constructed SAET Mock Tests

SAET Mock Test Constructed by ET 1

Material A

I. Vocabulary

1. English is a(n) _____ language, serving as a necessary tool to communicate with people with diverse nationalities.
   (A) current (B) dominant (C) accurate (D) fashionable
2. Human beings are not the only at risk of _____ the flu this season. The furry friends can fall ill as well.
   (A) contracting (B) combating (C) separating (D) prolonging
3. Despite its _____ of no more than 12 passengers, the van had 16 students in it.
   (A) capability (B) evaluation (C) measurement (D) capacity
4. Recently, with oil and electricity prices going up, the commodity prices increased _____. Thus, more and more people barely make both ends meet.
   (A) slightly (B) considerably (C) roughly (D) tentatively
5. It is of significant importance for drivers to have the _____ maintenance of their cars.
   (A) hesitant (B) temporary (C) routine (D) reckless

II. Cloze

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should __1__. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. In addition, good listeners are inclined to accept or tolerate __2__ to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are __3__ to that generality. For example, John Steinbeck is said __4__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen contributed to his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, __5__ what a good listener does, he may become either popular or disliked in his lifetime.

1. (A) despise (B) treasure (C) ignore (D) command
2. (A) as well as (B) instead of (C) regardless of (D) rather than
3. (A) attachments (B) addiction (C) exceptions (D) expectations
4. (A) being (B) to be (C) to have been (D) to have
5. (A) depending on (B) speaking of (C) compared with (D) different from

III. Reading comprehension

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. **This** was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

1. What is the main idea of the passage?
   (A) The history of gold fever.
   (B) Traveling light is crucial to seeking gold.
   (C) Shoes are the perfect option on display.
   (D) Daily stuff, like a shoe, can tell us an amazing story.
2. Which of the following is true about Chilkoot Pass?
   (A) The shortcut to the gold mine.
   (B) The celebrated path for gold seekers.
   (C) It was carved out of ice.
   (D) It, located on the field, emerged from gold fever.
3. According to the passage, what caused many of gold seekers to die?
   (A) A lack of shoes. (B) Carrying heavy backpacks.
   (C) Inadequate preparation. (D) Sudden cold weather.
4. In the fifth line of the second paragraph, what does "**this**" refer to?
   (A) The requirement for the gold seekers.
   (B) The greatest adventures of the shoe owner.
   (C) The Canadian government. (D) Gold fever.

Material B

I. Vocabulary

1. Without warning, the bridge _____, causing the cars to fall into the river. The tragedy was attributed to the erosion of the foundation and a lack of maintenance.
   (A) suspended (B) collapsed (C) constructed (D) extended
2. According to the investigation, the language _____ caused the air crash. The misunderstanding of the pilot's spoken English was to blame.
   (A) barrier (B) acquisition (C) requirement (D) resistance
3. The demanding teacher always gives students a lot of homework and tests. _____, students complain of that. In the end, they adapt to the way he teaches.
   (A) Tentatively (B) Alternatively (C) Unfortunately (D) Originally
4. School on a tight budget had difficulty continuing hiring guards for the campus safety and some generous parents responded to that by donating money on a _____ basis.
   (A) violent (B) hostile (C) voluntary (D) reluctant
5. The English teacher _____ the English composition and then asked the students to completely understand the corrected part.
   (A) revised (B) released (C) acclaimed (D) interpreted

II. Cloze

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. For example, in the United States, it is very important to be on time for almost all __1__. The only time it is socially acceptable to be late is when going to a friend's party. A person usually tried to arrive about five minutes __2__ the invitation time, so that the host would have a little extra time to prepare for the guests. __3__ is called being "fashionably late." Any time later than that is considered impolite, because it keeps the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would rush the host. __4__, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be __5__ by the patient.

1. (A) phenomena (B) excursions (C) factors (D) occasions
2. (A) before (B) as soon as (C) after (D) as early as
3. (A) It (B) There (C) This (D) What
4. (A) Instead (B) Likewise (C) Therefore (D) However
5. (A) filled out (B) applied for (C) filled with (D) taken over

III. Reading comprehension

Five years ago, David Smith wore an expensive suit to work every day. “I was a clothes addict,” he jokes. “I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled.” Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. “I’m working harder than ever,” David says, “and I need to feel comfortable.”

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as “dress-down Friday” or “casual Friday.” “What started out as an extra one-day-a-week benefit for employees has really become an everyday thing,” said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it’s easier for a company to attract new employees if it has a casual dress code. “A lot of young people don’t want to dress up for work,” says the owner of a software company, “so it’s hard to hire people if you have a conservative dress code.” Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee **morale**. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. “Suits are expensive, if you have to wear one every day,” one person said. “For the same amount of money, you can buy a lot more casual clothes.”

1. Which of the following is the best title for the article?
   (A) The Origin of Casual Friday
   (B) How to Raise the Efficiency of Office Workers
   (C) From Formal Wear to Casual Clothes
   (D) The Dilemma between Suits and Jeans
2. Which of the following statements is true of “casual Friday?”
   (A) It was also dubbed “dress-down Friday.”
   (B) In the early 1990, employees were allowed to wear casually in all the companies in America.
   (C) This dress code became an immediate hit.
   (D) It started out wearing casual clothes every day.
3. According to the passage, why have a lot of companies adopted a casual dress code?

(A) Young hires enjoy the casual working environment.
(B) The companies cannot afford to buy suits for their employees.
(C) Adopting less traditional dress code is extremely crucial to the recruitment of new employees.
(D) The employees work more efficiently in a good mood.

4. Which of the following words is synonymous with the word "**morale**?"
(A) integrity (B) productivity (C) happiness (D) enthusiasm

SAET Mock Test Constructed by ET 2

Material A

I. Vocabulary

1. Mr. Wang is one of my father's _____, who constantly seeks legal advice at Father's law firm.
(A) miracles (B) clients (C) routines (D) errands
2. Taiwan Railway Administration is working hard to make their trains' arrival time _____ so as to win people's trust back.
(A) accurate (B) current (C) practical (D) dominant
3. Informed of his admission to his ideal university, Patrick could _____ control his joy and let out a cry.
(A) obviously (B) continuously (C) briefly (D) scarcely
4. President Ma Ying-jeou calls for reforms, and thus takes immediate _____ to put his policy into practice.
(A) figures (B) influences (C) contracts (D) measures
5. Sandy's father was assigned to run a branch company in Mainland China; thus, she had no choice but to _____ to another school there.
(A) violate (B) recite (C) transfer (D) frustrate

II. Cloze

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. __1__ they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. __2__ , good listeners are inclined to accept or tolerate rather than to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are exceptions to that generality. For example, John Steinbeck __3__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen __4__ his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, depending on __5__ a good listener does, he may become either popular or disliked in his

lifetime.

1. (A) Because (B) Unless (C) If (D) Despite
2. (A) In consequence (B) In the end (C) In contrast (D) In addition
3. (A) was portrayed as (B) was told to pretend
   (C) is prone to be (D) is said to have been
4. (A) contributed to (B) originated from (C) involved in (D) substituted for
5. (A) that (B) what (C) which (D) how

III. Reading comprehension

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

1. What is the thesis statement of this article?
   (A) Alaska museum's talking object drew many people.
   (B) Bring enough supplies when going on adventures.
   (C) Magnificent stories might lay behind silent objects.
   (D) Gold fever took away people's lives due to the weather.
2. Where was the old leather shoe first located?
   (A) In a national museum. (B) On an icy mountain trail.
   (C) On a display shelf in a store. (D) In an underground gold mine.
3. What is **false** about people with "gold fever?"

(A) Many failed because their supplies were not sufficient.
(B) They went with a dream to become rich.
(C) Quite a few were unaware of what they would come across.
(D) They had to pack light in order to travel fast.

4. What can we infer about the shoe's owner from this article?
(A) She was a person with great courage.
(B) She was in the business of trading gold.
(C) She was a fashion queen in the 19th century.
(D) She lost her shoe while carrying supplies back and forth.

Material B

I. Vocabulary

1. With the release of its new smart phones, the manufacturer Nokia _____ 160% more app downloads than Apple.
(A) boasts (B) revises (C) maintains (D) approaches
2. Well goes the saying "A distant _____ is not as good as a near neighbor." That is, good neighbors are a lot more helpful when we are in need.
(A) vision (B) legend (C) dealer (D) relative
3. While _____ and situational loneliness can be a normal, healthy part of life, chronic loneliness can be a very sad and sometimes very dangerous condition.
(A) reluctant (B) liberal (C) temporary (D) sincere
4. Faced with a grave danger of our drinking water, we all _____ need to apply a solution to this problem lest our health be affected day after day by industrial waste in the river.
(A) narrowly (B) rapidly (C) hardly (D) urgently
5. The suspect didn't _____ her crime to the murder until she saw the videotape.
(A) propose (B) confess (C) collapse (D) gossip

II. Cloze

People have different ideas about what exactly is being on time and being late. __1__ ideas also differ from time to time, and from country to country. For example, in the United States, it is very important to be on time for almost all occasions. The only time __2__ is socially acceptable to be late is when going to a friend's party. A person usually tried to arrive about five minutes after the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being "fashionably late." Any time later than that is considered impolite, because it __3__ the host and other guests waiting.

Being on time __4__. One should also not arrive early for a friend's party, because it would rush the host. However, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be __5__ by the patient.

1. (A) These (B) Other (C) Theirs (D) Our
2. (A) that (B) what (C) which (D) it
3. (A) keeps (B) makes (C) has (D) gets
4. (A) goes neither way (B) goes both ways (C) goes one way (D) goes either way
5. (A) figured out (B) set out (C) filled out (D) sent out

III. Reading comprehension

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a conservative dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that **casual dress improves employee morale**. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

1. David Smith is a(n) _____.
   (A) tailor (B) consultant (C) employee (D) employer
2. Those who believe that **casual dress improves employee morale** have faith in

_____ if they let their employees dress comfortably to work.

(A) family harmony (B) productivity and creativity

(C) luck in the jackpot (D) earning money by saving

3. According to the article, what is **NOT** an advantage of wearing casually to office?

(A) Staying conservative. (B) Being economical.

(C) Feeling at ease. (D) Boosting innovation.

4. Which of the following is the author's opinion toward office dress code?

(A) It helps companies to recruit more new blood to work for them.

(B) The author supports that companies should allow a casual dress code.

(C) The author considers wearing formally is good to a company's image.

(D) The author did not state his or her own opinion on this issue.

SAET Mock Test Constructed by NT 1

Material A

I. Vocabulary

1. Despite continual misfortunes happening to me, I still feel _____ about my future.

(A) dominant (B) competitive (C) optimistic (D) practical

2. Instead of telling me _____, Judy wrote me that she didn't love me anymore.

(A) briefly (B) considerably (C) immediately (D) directly

3. As a good Taiwan citizen, we should _____ whatever is against the law.

(A) violate (B) frustrate (C) resist (D) decrease

4. Most people enjoy watching famous paintings even though they can hardly _____ the painter's thoughts.

(A) transfer (B) interpret (C) measure (D) frustrate

5. Doing all the housework for Mom is my weekly _____. I do so every Sunday morning.

(A) errand (B) capacity (C) routine (D) contract

II. Cloze

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them __1__ other people. __2__, good listeners are inclined to accept or tolerate rather than to judge and criticize. Therefore, they have fewer enemies than other people. In fact, they are probably the most loved of people. However, there are exceptions __3__ that generality. For example, John Steinbeck is said to __4__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen contributed to his capacity to write. Nevertheless, the results of his listening did not

make him popular. Thus, depending on __5__ a good listener does, he may become either popular or disliked in his lifetime.

1.	(A) as	(B) with	(C) than	(D) about
2.	(A) In fact	(B) In addition	(C) In short	(D) In other words
3.	(A) for	(B) with	(C) about	(D) to
4.	(A) have been	(B) be	(C) has been	(D) become
5.	(A) that	(B) what	(C) which	(D) how

III. Reading comprehension

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman's shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn't need in order to travel lighter?

Over 100,000 people with "gold fever" made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we do know is that she took part in one of the greatest adventures in the 19th century.

1. Why does the author say the shoe is special?
   (A) The shoe belongs to a woman.
   (B) The shoe is popular in the 1890s.
   (C) It was found on the trail which people seeking gold would pass.
   (D) It's made of leather and stored in a museum in Alaska.
2. What do we know about the shoe?
   (A) It looks ordinary just like any other shoe.
   (B) It was thrown away because its owner no longer needed it.
   (C) It was accidentally dropped by a woman climbing up stairs.

(D) Its owner died of starvation.

3. According to the article, what is true about “gold fever?”
   (A) Many gold seekers died of hunger and dry weather.
   (B) Whoever wanted to seek gold had to carry one ton of necessities with them.
   (C) A successful gold seeker must have a pair of expensive shoes.
   (D) The gold seekers’ backpacks weighed no more than forty pounds.
4. What can we infer about the owner of the shoe?
   (A) She took no more than thirty trips in order to carry her supplies.
   (B) She had to endure the humid temperature for one year.
   (C) She was a brave gold seeker.
   (D) She hurt herself and dropped her shoe on her journey.

Material B

I. Vocabulary

1. The father held a _____ attitude on his daughter’s marriage. He let her decide who would be the one she could rely on for life.
   (A) comfortable (B) promising (C) liberal (D) sincere
2. The student who had overslept _____ caught the school bus with her long hair clipped by the bus door.
   (A) originally (B) expressively (C) totally (D) narrowly
3. With his parents’ patience and company for years, the retarded child _____ catch up with his classmates.
   (A) rapidly (B) gradually (C) hardly (D) urgently
4. The _____ to the regulation on speed limit is allowed when the ambulance is on its duty.
   (A) exception (B) approach (C) solution (D) technique
5. If good friends misunderstand each other, their friendship will be difficult to ____.
   (A) preserve (B) freeze (C) maintain (D) confess

II. Cloze

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. __1__, in the United States, it is very important to be on time for almost all occasions. The only time it is __2__ acceptable to be late is when going to a friend’s party. A person usually tried to arrive about five minutes __3__ the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being “fashionably late.” Any time later than that is considered impolite, because it __4__

the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would rush the host. __5__, when going to a doctor's appointment, it is usually good to arrive earlier than the appointment because there are usually forms that need to be filled out by the patient.

1.	(A) Therefore	(B) For example	(C) However	(D) At first
2.	(A) hardly	(B) wisely	(C) socially	(D) directly
3.	(A) after	(B) before	(C) on	(D) by
4.	(A) makes	(B) forces	(C) lets	(D) keeps
5.	(A) Actually	(B) However	(C) In conclusion	(D) Unfortunately

III. Reading comprehension

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a conservative dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee morale. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

1. What is the best title for this article?

(A) The One-day-a-week Casual Friday

(B) The Reasons for Wearing Casual Clothes to Work
(C) The Comparison between Suits and Casual Clothes
(D) All Employers Love Casual Clothes

2. Which statement is correct?
(A) It's difficult to recruit new workers if there's a dress code in a company.
(B) In the early 1990s, employers can wear whatever they like to work every day.
(C) David Smith works for a software company.
(D) David Smith works harder when he works in a suit.

3. According to a research by Levi Strauss, how do most employers think about dress code?
(A) It makes their employees less productive.
(B) It's practicable only on Friday.
(C) It helps their employees work efficiently in a good mood.
(D) It has a negative impact on their productivity.

4. Which of the following statement is NOT the reason most employees love casual dress?
(A) It's comfortable to wear casual dress.
(B) Their bosses prefer casual dress rather than suits.
(C) Casual clothes are much cheaper than suits.
(D) It makes them work more happily and productively.

SAET Mock Test Constructed by NT 2

Material A

I. Vocabulary

1. To prevent people from drunk driving, the government should take the necessary _____ to punish those who drive cars after they have drinks.
(A) contract (B) errand (C) influence (D) measure

2. Because of the _____ of birth rate, the country is now facing the problem of having fewer and fewer young people to support it.
(A) decrease (B) addition (C) capacity (D) routine

3. In nature, the _____ animal is usually the largest and most powerful creature that other animals would not fight with.
(A) fearful (B) optimistic (C) dominant (D) practical

4. In order to _____ more information on the issue, the students decide to go to the library to find the books that will be helpful to them.
(A) resist (B) obtain (C) transfer (D) loosen

5. Although the rain is light, if it falls _____, there will be a flood hitting the town.
(A) briefly (B) scarcely (C) immediately (D) continuously

II. Cloze

Most people like to talk, but few people like to listen. Yet listening well is a rare talent that everyone should treasure. Because they hear more, good listeners tend to know more and to be more sensitive to what is going on around them than other people. __1__, good listeners are inclined to accept or tolerate than to judge and criticize. Therefore, they have __2__ enemies than other people. In fact, they are probably the most loved of people. However, there are __3__ to that generality. For example, John Steinbeck is said to __4__ an excellent listener, yet he was hated by some of the people he wrote about. No doubt his ability to listen __5__ his capacity to write. Nevertheless, the results of his listening did not make him popular. Thus, depending on what a good listener does, he may become either popular or disliked in his lifetime.

1.	(A) In contrast	(B) As a result	(C) In addition	(D) For that reason
2.	(A) few	(B) fewer	(C) little	(D) less
3.	(A) possibilities	(B) reasons	(C) inclusions	(D) exceptions
4.	(A) have been	(B) has been	(C) had been	(D) having been
5.	(A) resulted from	(B) contributed to	(C) consisted of	(D) persisted in

III. Reading comprehension

Every object tells a story. Even the most ordinary objects can present to us powerful images. Sometimes it is the ordinary nature of these objects that actually makes them so extraordinary. Such is the case with an old leather shoe in a museum in Alaska. At first glance it does not look like much. It is a woman’s shoe of a style popular in the 1890s. But what is unique about this shoe is where it was found. It was discovered on the Chilkoot Pass, the famous trail used by the people seeking gold in Alaska. Who it belonged to or why it was left there is not known. Was it perhaps dropped by accident as the woman climbed up the 1,500 stairs carved out of ice? Or did she throw away goods that she didn’t need in order to travel lighter?

Over 100,000 people with “gold fever” made this trip hoping to become millionaires. Few of them understood that on their way they would have to cross a harsh wilderness. Unprepared for such a dangerous journey, many died of starvation and exposure to the cold weather. The Canadian government finally started requiring the gold seekers to bring one ton of supplies with them. This was thought to be enough for a person to survive for one year. They would carry their supplies in backpacks each weighing up to fifty pounds; it usually took at least 40 trips to get everything to the top and over the pass. Whoever dropped the shoe must have been a brave and determined woman. Perhaps she was successful and made it to Alaska. Perhaps she had to turn back in defeat. No one will ever know for sure, but what we

do know is that she took part in one of the greatest adventures in the 19th century.

1. What is the purpose of this article?
   (A) To inform readers of the stories behind an ordinary object.
   (B) To introduce a famous collection of objects in a museum of Alaska.
   (C) To discourage readers from seeking gold because it is very dangerous.
   (D) To convince readers that the woman who dropped the shoes must be a millionaire.
2. According to the second paragraph, which of the following is **NOT TRUE** about the gold-seeking journey in Alaska?
   (A) People who took part in this journey were determined to be rich.
   (B) The journey required its participants to carry necessities with them.
   (C) The gold seekers knew the dangers before they embarked on the journey.
   (D) There were challenges, such as lack of food and harsh weather, awaiting the gold seekers.
3. What can we infer from this passage?
   (A) The woman brought with her the supplies which weighed over 40 pounds.
   (B) The woman who joined the gold-seeking trip followed the crowd in clothing.
   (C) The woman knew the journey was full of dangers when she decided to make it.
   (D) The woman could not stand the cold weather because she lost the leather shoe.
4. Why did the author mention the old leather shoe?
   (A) To discuss the trend of shoe wearing in the 1890s.
   (B) To prove that men were not the only participants to seek gold.
   (C) To introduce the Chilkoot Pass, the most dangerous site in Alaska.
   (D) To show that a common thing like this may have some tales to tell.

Material B

I. Vocabulary

1. Since it is Chinese Lunar New Year, the kids persuade their _____ parents to buy the toys for them as gifts.
   (A) liberal (B) sincere (C) reluctant (D) voluntary
2. They _____ planned to watch that film, but the tickets sold out, so they saw this film instead.
   (A) gradually (B) originally (C) hardly (D) urgently
3. Miranda filled in the blank of _____ with the word “nurse,” which means she worked as a nurse.
   (A) occupation (B) relative (C) technique (D) division
4. Feeling guilty, the naughty boy who broke the window finally _____ that he did

it and apologized for his wrongdoing.

(A) confessed (B) boasted (C) rumored (D) proposed

5. Since there are fewer and fewer people speaking the language, people from the tribe are trying to _____ it in case it will disappear in the future.

(A) maintain (B) overthrow (C) revise (D) preserve

II. Cloze

People have different ideas about what exactly is being on time and being late. These ideas also differ from time to time, and from country to country. __1__, in the United States, it is very important to be on time for almost all occasions. The only time it is socially acceptable to be late is __2__ going to a friend's party. A person usually tried to arrive about five minutes after the invitation time, so that the host would have a little extra time to prepare for the guests. This is called being "fashionably late." Any time later than that is considered __3__, because it keeps the host and other guests waiting.

Being on time goes both ways. One should also not arrive early for a friend's party, because it would __4__ the host. However, when going to a doctor's appointment, it is usually good to arrive __5__ than the appointment because there are usually forms that need to be filled out by the patient.

1. (A) In addition (B) For example (C) As a result (D) Even so
2. (A) how (B) why (C) when (D) where
3. (A) friendly (B) hostile (C) agreeable (D) impolite
4. (A) rush (B) delay (C) grip (D) free
5. (A) later (B) earlier (C) quicker (D) slower

III. Reading comprehension

Five years ago, David Smith wore an expensive suit to work every day. "I was a clothes addict," he jokes. "I used to carry a fresh suit to work with me so I could change if my clothes got wrinkled." Today David wears casual clothes—khaki pants and a sports shirt—to the office. He hardly ever wears a necktie. "I'm working harder than ever," David says, "and I need to feel comfortable."

More and more companies are allowing their office workers to wear casual clothes to work. In the United States, the change from formal to casual office wear has been gradual. In the early 1990s, many companies allowed their employees to wear casual clothes on Friday (but only on Friday). This became known as "dress-down Friday" or "casual Friday." "What started out as an extra one-day-a-week benefit for employees has really become an everyday thing," said business consultant Maisly Jones.

Why have so many companies started allowing their employees to wear casual clothes? One reason is that it's easier for a company to attract new employees if it has a casual dress code. "A lot of young people don't want to dress up for work," says the owner of a software company, "so it's hard to hire people if you have a **conservative** dress code." Another reason is that people seem happier and more productive when they are wearing comfortable clothes. In a study conducted by Levi Strauss and Company, 85 percent of employers said that they believe that casual dress improves employee morale. Only 4 percent of employers said that casual dress has a negative impact on productivity. Supporters of casual office wear also argue that a casual dress code helps them save money. "Suits are expensive, if you have to wear one every day," one person said. "For the same amount of money, you can buy a lot more casual clothes."

1. Which of the following statements about the study by Levi Strauss and Company is **NOT TRUE**?
   (A) 15% of employers think casual wear will make employees unwilling to work.
   (B) 85% of employers believe that they have bad mood for working in casual wear.
   (C) 4% of employers argue that employees will produce less when they wear casual clothes.
   (D) Those who welcome the policy of casual wear think they will save money by buying casual clothes instead of formal ones.
2. Which of the following best describe David Smith?
   (A) He cannot go out without a necktie on him.
   (B) He thinks suits will make him look more handsome.
   (C) He wore formal clothes to work every day five years ago.
   (D) He now finds his work boring and feels tiresome to work.
3. What does "**conservative**" in the fourth line of the third paragraph refer to?
   (A) Resistant to accept new ideas.
   (B) Addicted to extraordinary things.
   (C) Eager to preserve cultural heritage.
   (D) Sensitive to the right time to do the right thing.
4. What is this article mainly about?
   (A) The most attractive office wearing style.
   (B) The naming of "casual Friday" and its dress code.
   (C) Formal suits and their beneficial effects on the firm.
   (D) Casual wear and its positive influence on employees.

Appendix J: Words Chosen by Different Teachers in Their Tests

Material A		
Word chosen by four teachers	dominant (adj.)	ET 1, ET 2, NT 1, NT 2
Word chosen by three teachers	transfer (v.)	ET 2, NT 1, NT 2
	capacity (n.)	ET 1, NT 1, NT 2
	contract (n.)	ET 2, NT 1, NT 2
	errand (n.)	ET 2, NT 1, NT 2
	routine (n.)	ET 2, NT 1, NT 2
	practical (adj.)	ET 2, NT 1, NT 2
	briefly (adv.)	ET 2, NT 1, NT 2
Word chosen by two teachers	frustrate (v.)	ET 2, NT 1
	resist (v.)	NT 1, NT 2
	violate (v.)	ET 2, NT 1
	influence (n.)	ET 2, NT 2
	measure (n.)	ET 2, NT 2
	accurate (adj.)	ET 1, ET 2
	current (adj.)	ET 1, ET 2
	optimistic (adj.)	NT 1, NT 2
	considerably (adv.)	ET 1, NT 1
	continuously (adv.)	ET 2, NT 2
	immediately (adv.)	NT 1, NT 2
	scarcely (adv.)	ET 2, NT 2
Material B		
Word chosen by three teachers	confess (v.)	ET 2, NT 1, NT 2
	liberal (adj.)	ET 2, NT 1, NT 2
	reluctant (adj.)	ET 1, ET 2, NT 2
	sincere (adj.)	ET 2, NT 1, NT 2
	hardly (adv.)	ET 2, NT 1, NT 2
	originally (adv.)	ET 1, NT 1, NT 2
	urgently (adv.)	ET 2, NT 1, NT 2
Word chosen by two teachers	boast (v.)	ET 2, NT 2
	collapse (v.)	ET 1, ET 2
	maintain (v.)	ET 2, NT 1
	propose (v.)	ET 2, NT 2
	revise (v.)	ET 1, ET 2
	relative (n.)	ET 2, NT 2
	technique (n.)	NT 1, NT 2
	voluntary (adj.)	ET 1, NT 2
	gradually (adv.)	NT 1, NT 2
	narrowly (adv.)	ET 2, NT 1
	rapidly (adv.)	ET 2, NT 1

Appendix K: Students' Answers to the Items on Each Form

Form A: Higher-proficiency students

	Vocabulary											Cloze										Reading								
	NT1A					ET1B						NT1A					ET1B					NT1A					ET1B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	C	D	C	B	C	B	A	D	C	A	C	B	D	A	B	D	C	C	D	A	C	A	B	C	C	A	D	D	score
01	C	D	A	A	C	B	A	D	C	D	B	A	D	D	B	D	A	A	C	B	C	C	B	C	A	A	D	B	**15**
08	B	D	C	C	C	B	A	D	C	B	C	A	D	B	B	D	A	A	D	B	C	A	C	C	C	A	C	B	**17**
09	C	D	C	A	C	B	A	C	C	A	C	A	D	B	D	D	D	A	A	C	B	A	B	C	C	C	D	B	**16**
16	C	B	A	C	A	C	A	B	C	D	B	B	D	B	D	C	A	A	C	A	C	C	A	C	C	A	C	D	**11**
17	C	D	C	C	C	B	A	A	C	A	C	D	D	B	D	D	C	A	D	B	C	A	B	C	C	A	A	C	**19**
24	C	D	C	B	C	B	A	B	D	D	C	A	C	B	D	A	C	D	D	C	C	C	B	C	C	C	D	B	**15**

Form A: Lower-proficiency students

	Vocabulary											Cloze										Reading								
	NT1A					ET1B						NT1A					ET1B					NT1A					ET1B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	C	D	C	B	C	B	A	D	C	A	C	B	D	A	B	D	C	C	D	A	C	A	B	C	C	A	D	D	score
01	A	D	C	B	C	B	A	D	C	A	C	B	B	C	B	C	D	A	A	D	C	C	B	D	C	B	D	B	**16**
08	D	A	C	B	C	B	A	C	B	C	D	A	C	B	D	D	C	A	A	D	D	C	B	C	C	A	D	D	**13**
09	D	D	A	C	C	C	D	C	A	D	B	B	D	B	A	C	B	C	B	D	B	D	B	C	B	A	D	D	**10**
16	A	C	D	C	D	B	D	B	C	B	B	A	C	B	A	C	C	A	B	D	C	C	B	A	C	A	B	B	**7**
17	C	B	D	A	C	C	A	D	A	D	D	A	A	B	B	C	A	A	A	B	C	D	A	C	A	B	C	B	**7**
24	C	C	D	B	B	B	A	A	C	A	C	A	A	B	C	A	C	A	D	D	C	A	B	C	B	A	D	D	**16**

Form B: Higher-proficiency students

	Vocabulary										Cloze										Reading								
	ET2A						NT1B				ET2A				NT1B						ET2A				NT1B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	B	A	D	D	C	C	D	B	A	C	A	D	D	A	B	B	C	A	D	B	C	B	D	A	B	A	C	B	score
02	B	A	C	D	B	B	C	B	A	C	A	D	D	D	B	B	C	A	D	B	C	B	D	A	B	D	C	B	**22**
07	B	A	D	D	C	B	B	B	A	C	A	D	A	C	D	B	C	B	D	B	C	D	C	D	C	A	C	B	**18**
10	B	C	D	C	D	D	C	A	C	C	A	D	B	A	D	B	C	A	D	B	C	B	D	A	B	A	C	B	**19**
15	B	A	D	A	C	D	B	B	C	C	A	C	B	A	D	B	D	A	D	A	D	A	B	A	B	A	C	B	**16**
18	A	B	A	B	C	D	C	B	A	C	D	D	D	A	B	B	C	A	D	B	C	B	D	A	B	A	C	B	**21**
23	B	A	A	A	C	C	D	B	A	C	A	B	A	B	B	B	C	A	D	B	C	B	D	A	B	A	C	B	**23**

Form B: Lower-proficiency students

	Vocabulary										Cloze										Reading								
	ET2A						NT1B				ET2A				NT1B						ET2A				NT1B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	B	A	D	D	C	C	D	B	A	C	A	D	D	A	B	B	C	A	D	B	C	B	D	A	B	A	C	B	score
02	D	B	B	C	B	B	D	C	B	C	A	D	A	A	B	B	C	A	C	B	D	D	D	A	C	A	D	D	**13**
07	C	B	B	C	C	D	D	B	A	D	C	A	C	D	A	B	C	B	C	A	B	A	B	D	C	C	D	B	**7**
10	B	A	A	C	B	D	D	C	A	C	C	D	A	C	D	B	D	A	C	C	D	A	B	A	B	D	C	B	**12**
15	C	B	C	B	C	B	D	B	D	A	C	C	A	A	B	B	A	B	A	B	B	B	A	D	C	D	C	B	**10**
18	C	D	C	B	B	A	D	A	A	D	C	A	A	C	C	B	A	C	D	A	A	A	B	D	A	B	D	C	**4**
23	D	B	A	B	D	B	D	C	C	C	A	D	B	B	A	B	A	A	C	A	C	D	A	D	C	D	C	B	**9**

Form C: Higher-proficiency students

	Vocabulary											Cloze										Reading								
	NT2A									ET2B		NT2A						ET2B				NT2A				ET2B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	D	A	C	B	D	A	D	C	D	B	C	B	D	A	B	A	D	A	B	C	A	C	B	D	C	B	A	D	score
03	A	A	C	B	D	D	D	C	D	B	B	B	A	A	A	A	A	A	C	C	A	A	C	D	C	B	A	D	**19**
06	B	A	C	B	D	D	D	A	B	D	A	B	D	B	B	A	A	B	D	C	A	C	A	D	C	B	D	C	**15**
11	C	A	C	B	C	D	D	A	D	A	B	B	A	A	B	A	A	B	D	C	C	C	C	A	C	B	B	D	**14**
14	A	A	A	B	D	A	D	C	D	B	C	B	D	A	A	A	D	A	D	C	A	C	B	D	C	D	A	D	**23**
19	C	A	C	B	D	C	D	C	B	B	C	B	D	A	B	A	C	A	C	C	B	A	A	A	C	B	D	D	**18**
22	C	A	A	B	C	C	D	C	B	A	B	B	D	B	B	A	D	B	D	C	B	A	C	A	C	B	B	A	**12**

Form C: Lower-proficiency students

	Vocabulary											Cloze										Reading								
	NT2A									ET2B		NT2A						ET2B				NT2A				ET2B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	D	A	C	B	D	A	D	C	D	B	C	B	D	A	B	A	D	A	B	C	A	C	B	D	C	B	A	D	score
03	C	B	B	A	C	C	D	B	B	C	D	D	A	C	C	C	A	B	C	C	D	A	C	C	D	B	C	B	**3**
06	D	A	C	B	C	D	D	B	B	A	B	A	C	B	D	A	A	B	C	D	B	C	C	A	D	B	C	B	**8**
11	B	A	C	B	B	D	D	C	C	A	C	B	A	D	B	A	A	D	C	B	A	B	C	A	D	D	D	B	**10**
14	B	B	D	A	D	D	D	C	B	B	C	B	D	B	A	A	B	A	C	C	A	C	C	C	C	B	C	B	**14**
19	C	A	B	B	A	C	A	D	C	D	B	A	C	A	B	C	A	B	D	D	B	A	B	C	C	B	A	D	**9**
22	D	A	B	B	D	D	A	A	C	B	B	A	B	C	A	C	C	D	C	C	A	B	A	A	D	D	C	A	**7**

Form D: Higher-proficiency students

	Vocabulary														Cloze								Reading							
	ET1A						NT2B				ET1A				NT2B								ET1A				NT2B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	score
Key	B	A	D	B	C	C	B	A	A	D	B	D	C	C	A	B	C	D	A	B	D	B	C	A	B	C	A	D	
04	A	A	D	B	C	B	B	A	A	A	B	D	D	A	C	B	C	D	A	B	D	B	A	A	B	C	C	D	**20**
05	A	A	D	B	C	A	B	A	A	D	B	C	B	B	C	B	D	C	C	B	D	D	C	A	B	C	C	D	**17**
12	A	A	A	B	A	C	B	A	A	D	B	D	C	B	B	B	C	D	A	B	D	C	D	A	A	C	C	D	**19**
13	A	A	B	A	C	A	B	B	A	D	B	D	A	B	A	B	C	D	A	B	D	D	D	A	B	C	A	A	**18**
20	C	A	D	C	C	A	B	B	A	D	D	D	C	B	C	B	C	D	A	B	A	C	D	A	B	C	C	D	**17**
21	A	C	A	B	D	C	B	B	A	D	B	B	A	B	C	B	C	D	A	B	D	C	D	A	B	C	A	D	**17**

Form D: Lower-proficiency students

	Vocabulary														Cloze								Reading							
	ET1A						NT2B				ET1A				NT2B								ET1A				NT2B				
No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	
Key	B	A	D	B	C	C	B	A	A	D	B	D	C	C	A	B	C	D	A	B	D	B	C	A	B	C	A	D	score
04	D	C	B	C	C	D	B	A	A	B	A	B	D	A	C	B	C	D	B	B	A	A	D	D	B	A	B	D	**10**
05	B	A	A	D	C	C	B	A	D	A	C	B	A	B	A	B	C	D	B	B	B	B	D	C	A	B	B	A	**12**
12	D	C	B	C	C	B	B	B	D	B	A	A	A	A	A	B	C	D	B	B	A	D	A	A	B	C	B	D	**11**
13	B	C	A	A	C	A	B	D	A	D	D	D	A	A	C	B	C	C	B	A	D	D	D	A	B	C	A	C	**13**
20	D	C	D	A	B	C	C	B	A	C	D	A	B	D	A	B	A	B	B	C	A	C	D	D	A	C	B	A	**6**
21	C	B	C	A	C	C	B	A	A	D	B	D	A	B	C	B	C	C	A	A	A	D	D	A	C	A	D	C	**12**

Appendix L: Frequencies of the Comparisons Between Students' Test-taking Strategies and Teachers' Test-constructing Considerations

Form A

		Higher-proficiency			Lower-proficiency		
Item	No.	Consistent	Inconsistent	Others	Consistent	Inconsistent	Others
Vocabulary	1	5	1	0	0	0	6
	2	4	2	0	1	1	4
	3	1	4	1	0	4	2
	4	0	4	2	1	3	2
	5	5	1	0	2	1	3
	6	5	1	0	3	1	2
	7	6	0	0	4	0	2
	8	3	1	2	2	2	2
	9	5	0	1	2	2	2
	10	0	2	4	0	1	5
Cloze	11	4	2	0	2	4	0
	12	1	4	1	0	5	1
	13	4	0	2	0	2	4
	14	0	6	0	0	5	1
	15	3	3	0	1	4	1
	16	3	3	0	0	3	3
	17	2	4	0	2	3	1
	18	2	4	0	0	5	1
	19	3	3	0	2	4	0
	20	2	4	0	0	5	1
Reading	21	4	2	0	4	2	0
	22	2	2	2	1	5	0
	23	3	2	1	1	2	3
	24	5	0	1	2	2	2
	25	4	1	1	2	3	1
	26	1	2	3	1	3	2
	27	3	1	2	3	1	2
	28	0	4	2	0	4	2

Form B

Item	No.	Higher-proficiency			Lower-proficiency		
		Consistent	Inconsistent	Others	Consistent	Inconsistent	Others
Vocabulary	1	4	0	2	0	0	6
	2	4	0	2	1	1	4
	3	1	2	3	0	3	3
	4	2	4	0	0	3	3
	5	4	2	0	3	0	3
	6	0	6	0	0	5	1
	7	0	5	1	0	0	6
	8	4	2	0	1	3	2
	9	4	2	0	0	3	3
	10	5	1	0	1	2	3
Cloze	11	5	1	0	1	4	1
	12	4	2	0	3	1	2
	13	0	4	2	0	3	3
	14	2	1	3	0	3	3
	15	1	5	0	1	4	1
	16	6	0	0	6	0	0
	17	5	1	0	1	3	2
	18	4	2	0	2	3	0
	19	6	0	0	0	6	0
	20	5	1	0	1	3	2
Reading	21	4	2	0	1	3	2
	22	3	3	0	0	6	0
	23	2	4	0	0	3	3
	24	5	1	0	2	3	1
	25	3	1	2	1	5	0
	26	4	2	0	1	5	0
	27	5	0	1	3	3	0
	28	5	0	1	1	2	3

Form C

Item	No.	Higher-proficiency			Lower-proficiency		
		Consistent	Inconsistent	Others	Consistent	Inconsistent	Others
Vocabulary	1	0	4	2	0	1	5
	2	6	0	0	2	1	3
	3	3	2	1	1	2	3
	4	6	0	0	1	0	5
	5	4	2	0	1	4	1
	6	1	5	0	0	2	4
	7	4	2	0	2	3	1
	8	1	0	5	0	0	6
	9	2	3	1	0	2	4
	10	2	0	4	0	1	5
Cloze	11	0	4	2	0	4	2
	12	6	0	0	1	3	2
	13	3	2	1	1	3	2
	14	3	2	1	1	4	1
	15	1	2	3	0	3	3
	16	4	1	1	3	2	1
	17	1	3	2	0	5	1
	18	2	4	0	0	4	2
	19	0	5	1	0	4	2
	20	4	2	0	1	2	3
Reading	21	3	3	0	2	3	1
	22	1	3	2	1	5	0
	23	0	5	1	0	6	0
	24	1	4	1	0	5	1
	25	6	0	0	1	1	4
	26	2	2	2	0	5	1
	27	1	4	1	0	4	2
	28	4	2	0	0	4	2

Form D

Item	No.	Higher-proficiency			Lower-proficiency		
		Consistent	Inconsistent	Others	Consistent	Inconsistent	Others
Vocabulary	1	0	4	2	0	2	4
	2	1	2	3	0	2	3
	3	2	2	2	0	2	4
	4	3	2	1	0	2	4
	5	3	1	2	0	5	1
	6	0	4	2	0	1	5
	7	6	0	0	5	1	0
	8	3	3	0	0	2	4
	9	5	0	1	1	3	2
	10	5	1	0	2	3	1
Cloze	11	4	2	0	1	2	3
	12	4	2	0	2	3	1
	13	1	5	0	0	2	4
	14	0	6	0	0	5	1
	15	0	6	0	0	4	2
	16	6	0	0	6	0	0
	17	5	1	0	5	1	0
	18	5	1	0	2	2	2
	19	2	2	2	0	5	1
	20	2	4	0	1	5	0
Reading	21	5	0	1	0	6	0
	22	0	5	1	0	5	1
	23	1	5	0	0	6	0
	24	3	1	2	1	5	0
	25	2	3	1	2	4	0
	26	4	1	1	1	4	1
	27	2	4	0	1	3	2
	28	5	1	0	2	4	0