ISSN 1930-9317

TOEFL

TOEFL iBT™ Research Report

***TOEFLiBT*-10**
September ***2009***

The Speaking Section of the TOEFL iBT™ (SSTiBT): Test-Takers' Reported Strategic Behaviors

Merrill Swain

Li-Shih Huang

Khaled Barkaoui

Lindsay Brooks

Sharon Lapkin

The Speaking Section of the TOEFL iBT™ (SSTiBT): Test-Takers' Reported Strategic Behaviors

Merrill Swain
The Ontario Institute for Studies in Education of the University of Toronto, Canada

Li-Shih Huang
University of Victoria, British Columbia, Canada

Khaled Barkaoui
York University, Toronto, Ontario, Canada

Lindsay Brooks and Sharon Lapkin
The Ontario Institute for Studies in Education of the University of Toronto, Canada

RR-09-30

ETS is an Equal Opportunity/Affirmative Action Employer.

As part of its educational and social mission and in fulfilling the organization's non-profit Charter and Bylaws, ETS has and continues to learn from and also to lead research that furthers educational and measurement research to advance quality and equity in education and assessment for all users of the organization's products and services.

Abstract

This study responds to the *Test of English as a Foreign Language*™ (TOEFL®) research agenda concerning the need to understand the processes and knowledge that test-takers utilize. Specifically, it investigates the strategic behaviors test-takers reported using when taking the Speaking section of the TOEFL iBT™ (SSTiBT). It also investigates how the reported strategic behaviors differed across integrated and independent tasks in the SSTiBT, as well as the relationship between test-takers' reported strategic behaviors and their performance on the tasks as determined by their test scores. The participating students were 14 graduate and 16 undergraduate engineering students whose first language was Chinese.

The results indicate that test-takers reported using 49 separate strategies when completing the SSTiBT tasks. Of the five strategy categories, the metacognitive, communication, and cognitive strategies were proportionally the most frequently reported. The interrelationships among these three categories were negative. Undergraduates reported using significantly more communication strategies, whereas graduates reported using significantly more cognitive and affective strategies. No statistically significant differences were found in reported strategy use across proficiency levels. The integrated tasks were more alike with respect to reported strategy use than were the independent and integrated tasks. Furthermore, the integrated tasks elicited a wider variety of reported strategy use than the independent tasks. Overall, we found no relationship between the total number of reported strategic behaviors and total test score on the SSTiBT.

We conclude that strategy use is integral to performing SSTiBT tasks and should therefore be considered as part of the construct of communicative performance. However, the relationship between strategy use and test performance is varied and is due to complex interactions among test-taker characteristics, tasks, and contexts.

Key words: Academic speaking, second-language speaking, strategic behaviors, speaking tasks, speaking tests, TOEFL iBT

The Test of English as a Foreign Language™ (TOEFL®) was developed in 1963 by the National Council on the Testing of English as a Foreign Language. The Council was formed through the cooperative effort of more than 30 public and private organizations concerned with testing the English proficiency of nonnative speakers of the language applying for admission to institutions in the United States. In 1965, Educational Testing Service (ETS) and the College Board® assumed joint responsibility for the program. In 1973, a cooperative arrangement for the operation of the program was entered into by ETS, the College Board, and the Graduate Record Examinations® (GRE®) Board. The membership of the College Board is composed of schools, colleges, school systems, and educational associations; GRE Board members are associated with graduate education. The test is now wholly owned and operated by ETS.

ETS administers the TOEFL program under the general direction of a policy board that was established by, and is affiliated with, the sponsoring organizations. Members of the TOEFL Board (previously the Policy Council) represent the College Board, the GRE Board, and such institutions and agencies as graduate schools of business, two-year colleges, and nonprofit educational exchange agencies.

❖ ❖ ❖

Since its inception in 1963, the TOEFL has evolved from a paper-based test to a computer-based test and, in 2005, to an Internet-based test, TOEFL iBT. One constant throughout this evolution has been a continuing program of research related to the TOEFL test. From 1977 to 2005, nearly 100 research and technical reports on the early versions of TOEFL were published. In 1997, a monograph series that laid the groundwork for the development of TOEFL iBT was launched. With the release of TOEFL iBT, a TOEFL iBT report series has been introduced.

Currently this research is carried out in consultation with the TOEFL Committee of Examiners. Its members include representatives of the TOEFL Board and distinguished English as a second language specialists from the academic community. The Committee advises the TOEFL program about research needs and, through the research subcommittee, solicits, reviews, and approves proposals for funding and reports for publication. Members of the Committee of Examiners serve four-year terms at the invitation of the Board; the chair of the committee serves on the Board.

Current (2009-2010) members of the TOEFL Committee of Examiners are:

To obtain more information about the TOEFL programs and services, use one of the following:

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

Acknowledgments

We are grateful to the following people who contributed in various ways to this study: Glenn Fulcher and Mary Enright played key roles in facilitating the research project; Yan Wang assisted in data collection; and Dan Jiang and Yongfeng Jia transcribed and coded the data. Thanks are also due to our participants.

We wish to express our gratitude for the support of, and to acknowledge with thanks the timely and helpful collaboration of ETS personnel throughout the project. We also wish to thank the anonymous reviewers for their useful feedback and Xiaoming Xi for her thorough reading and detailed comments on earlier versions of the report.

Table of Contents

Page

List of Tables

List of Figures

Page

Introduction

The present study investigated test-takers' reported strategic behaviors when taking the new *Test of English as a Foreign Language*™ (TOEFL®) speaking test, the Speaking section of the TOEFL iBT™ (SSTiBT). Second-language acquisition (SLA) research on learner strategies has demonstrated that learners' strategy use is associated with second-language acquisition and performance (see Oxford, 2001; Oxford & Burry-Stock, 1995). However, from the language testing (LT) perspective, test-takers' strategic behaviors have not been given sufficient attention (Bachman, 1990, 2002; Kunnan, 1995; Purpura, 1998), even though they have been included in the language-ability models and communicative-competence models proposed by theorists in the field.

This project responds to the TOEFL research agenda concerning the need to understand the processes and knowledge test-takers utilize, by examining their reported strategy use. The project also responds to the acknowledgment in the LT field that researchers need to consider the strategies test-takers use when participating in second-language testing, in order to demonstrate that inferences about the academic speaking ability based on test-takers' performance are valid. This consideration is needed in order to address concerns about the construct validity of language tests (e.g., Bachman, 1990; Cohen, 1994, 1998, 2007; Kunnan, 1998), particularly if strategic competence is part of the construct definition (Fulcher, 2003). As Rosenfeld, Oltman, and Sheppard (2004) noted, as long as "the development of a new TOEFL continues, there will be a continuing need for test validation" (p. 1).

Research in the area of variation in tasks and contexts, as well as their effects on language use, has supported the hypothesis that both test performance (Bachman & Cohen, 1998) and strategy use (Poulisse, 1990) differ across tasks and across different proficiency levels (Purpura, 1999; Yoshida-Morise, 1998). As Cohen and Olshtain (1993) pointed out, "[N]ot all speaking tasks are created equal . . . there are tasks which make far greater demands on learners than do others" (p. 50). Butler, Eignor, Jones, McNamara, and Suomi (2000) also recognized how task characteristics and performance factors can influence test-takers' output on speaking tasks.

In addition to examining test-takers' reported strategic behaviors, we investigated how the reported strategic behaviors vary across three SSTiBT task groups and six individual SSTiBT speaking tasks,[1] and the relationship between respondents' reported strategic behaviors and their performance on the SSTiBT as indicated by their test scores.

Background

This section includes five parts. We begin by defining strategic behaviors and then discuss the meaning of the construct of strategic competence as found within models of communicative competence, followed by an overview of strategy taxonomies. We then present research on learner strategies within SLA. Finally, we introduce the present study.

Defining Strategic Behaviors

There is still much debate regarding how to define learner strategies, and different terminology is used within the field of SLA (Cohen, 1998; Ellis, 1994; Huang, 2004; Purpura, 1999). LT research focuses mainly on the test-taking strategies learners use to perform the task and deal with their communication needs during the test-taking process, rather than the strategies individuals employ when learning to communicate. We are aware of the lack of consensus about how processes and strategies are differentiated. In our view, strategy use is closely linked to cognitive processes[2] because strategies are the deliberate thoughts and behaviors used to manage or carry out cognitive processes with the goal of successful test performance. Based on this conceptualization, we examine strategic behaviors as those behaviors test-takers use to regulate their cognitive processes during a test or the behaviors they use to reflect on those cognitive processes.

For this study, *strategic behaviors* refers to the conscious thoughts and actions test-takers report using to acquire or manipulate information, such as attending, predicting, translating, planning, monitoring, linking, and inferencing (O'Malley & Chamot, 1990; Oxford, 1990; Phakiti, 2003); they are directly related to the test-taking process. Operationally, these strategic behaviors are the reported actions and thought processes used by test-takers. In principle, these strategies are defined as the conscious, goal-oriented thoughts and behaviors test-takers use to regulate cognitive processes, with the goal of improving their language use or test performance.

Strategic Competence as Part of the Speaking Construct

Language-testing researchers have become increasingly concerned about the various sources of variability that might influence performance on language tests, including the role strategic behaviors might play (Bachman, 1990; Bachman & Palmer, 1996; McNamara, 1996; Purpura, 1999). The strategic component "plays . . . a central role in the processing of communication" (Douglas, 1997, p. 6) and mediates between the context and its interpretation by test-takers (Douglas, 2000). Even though researchers and theorists view the second-language construct as multidimensional (e.g., Chamot, Küpper, & Impink-Hernandez, 1988; Purpura, 1998; Wesche, 1987), as Kunnan (1998) and Douglas (2000) pointed out, we have yet to identify and support evidentially the specific components underlying this multidimensional construct and how the dimensions interact in language use. Among these components are the strategies test-takers use.

Speakers' ability to use communication strategies to deal with communication breakdowns has been referred to as their *strategic competence*, which is a component of Canale and Swain's (1980) theoretical framework of communicative competence. Canale (1983) later expanded this component to include both compensatory and enhancement strategies. Bachman (1990) further broadened the model to include components of assessment, planning, and execution; this broadening is consistent with Widdowson's (1983) *communicative capacity*. Bachman and Palmer's (1996) conception of strategic competence includes "a set of metacognitive components or strategies," such as goal setting, assessment, and planning (p. 70). Douglas (1997) discussed the importance of the strategic component and includes three types of *processes* (Chapelle & Douglas, 1993) in the model of speaking in academic contexts: metacognitive strategies, language strategies, and fundamental cognitive strategies (see Chapelle & Douglas, 1993; Douglas 1997). In their COE (Committee of Examiners) model of communicative language proficiency in academic contexts, Chapelle, Grabe, and Berns (1997) termed strategic competence the *procedural competence* for enhancing communication or compensating for communication problems. Adapting Bachman and Palmer's (1996) model, Fulcher's (2003) most recently refined framework for describing the speaking construct includes strategic capacity, which features both achievement strategies and avoidance strategies.

In language testing, models of language have been a frequent focus of attention over the lifetime of the Language Testing Research Colloquium (Hamp-Lyons & Lynch, 1998). For the past two decades, much systematic research has examined the construct validation of the concept of *communicative competence* in second language education (e.g., Bachman & Palmer, 1996; Harley, Allen, Cummins, & Swain, 1990; Jamieson, Jones, Kirsch, Mosenthal, & Taylor, 2000; Palmer, Groot, & Trosper, 1981) and in language testing (e.g., Milanovic, Saville, Pollitt, & Cook, 1996; Swain, 1985; Wesche, 1981). Whether it is considered within Canale and Swain's (1980) communicative competence framework, Bachman's (1990) and Bachman and Palmer's

(1996) communicative language ability model, or the social-cognitive construct representation (see Chalhoub-Deville, 2003), strategic competence remains critical and has been recognized as interacting with other components of communicative competence.

While acknowledging that no single accepted representation of competence exists, and that the specific nature of the components remains debatable, *strategic competence* remains absent from the operational assessment framework in the scoring rubric for the SSTiBT tasks. Although there is growing recognition that these strategies and the interaction among strategies and tasks may affect performance, and that test-takers' strategy use can provide insights concerning test validity, research has been lacking with regard to the precise nature of strategic competence as applied to LT contexts.

As Swain (2001) stated, "Whoever is doing the task is engaging in construct-relevant processes while doing so" (p. 298). The field requires more empirical evidence about the actual strategies test-takers employ, in order to substantiate claims about the validity of inferences based on second language (L2) speaking test scores, and this evidence, as Fulcher (2003) explained, "has been one of the most difficult aspects of validity to study" (p. 195). Douglas (2000) stated that validation is "a dynamic process in which many different types of evidence are gathered and presented" and through which we can begin to obtain a better understanding of what a particular test is actually testing (p. 258). Examining strategy use is integral to gaining insights relevant to the construct. Chalhoub-Deville (2001) also called for language researchers and test constructors to "expand their test specifications to include the knowledge and skills that underlie the language construct" (p. 225). The effort to understand test-takers' strategic behaviors when they respond to assessment tasks is an important source of construct-validity evidence (e.g., Bachman, 2002; Chalhoub-Deville, 2001; McNamara, 1996), and the subject warrants in-depth investigation.

Strategy Taxonomies

In the early days of communication strategies research, communication strategies generally were regarded as strategies that individuals employed to deal with their communication needs while producing the target language, rather than during the course of learning to communicate in general. Færch and Kasper (1980) placed communication strategies in a processing model of speech production and defined them as "potentially conscious plans for solving what to an individual presents itself as a problem in a search for a particular communicative goal" (p. 81). While there are many varied taxonomies and theoretical approaches to communication strategies, there are also some overlaps among the strategy groups within each system, as well as among various systems. Several taxonomies already exist and are widely utilized for research and teaching purposes.

In the study of communication strategies, the development of the strategic component in the various frameworks of communicative-language skills led to numerous studies on the use of communication strategies in communicative tasks or situations (e.g., Færch & Kasper, 1980, 1983; Paribakht, 1985; Poulisse, 1987, 1990; Yoshida-Morise, 1998). Previous studies have revealed that numerous factors may affect the use of communication strategies. These factors include, but are not limited to, test/task differences in terms of proficiency level, language background, and instructional experiences.

The empirical basis of taxonomies is self-report data (interviews, questionnaires, and verbal protocols). Thus, taxonomies rely on participants' *reported use* of strategies rather than observations of learner/test-taker behavior. In terms of accuracy of reporting, think-aloud and stimulated recall are more focused and specific than are interview or questionnaire data with respect to a specific event. We are aware of the criticisms concerning the methodology used to elicit, measure, and classify strategies (e.g., LoCastro, 1994; Selinger, 1983; Skehan, 1991) but consider stimulated recalls as one of the best available means to achieve our goal of gaining greater understanding of the strategic behaviors test-takers use during a speaking test, while minimizing possible effects on speaking performance.

Research on Learner Strategies in Second-Language Acquisition

In the 1970s, much research was conducted on learner strategies and the relation between strategy use and second language performance. Much of this earlier work was devoted to descriptive studies that identified learner strategy type, variety, and frequency (e.g., Naiman, Fröhlich, Stern, & Todesco, 1978; Rubin, 1975). The generation of learner strategy lists has led to different ways of organizing and classifying learner strategies into frameworks and to differing opinions about how learner strategies should be categorized (e.g., Cohen, 2002; O'Malley & Chamot, 1990; Oxford, 1990).

Since the 1980s, the focus has shifted from a product to a process orientation. This shift in focus has generated much interest in the study of strategy use in language acquisition (e.g., Cohen, 1984; Cohen & Aphek, 1981; Homburg & Spaan, 1981; O'Malley & Chamot, 1990;

Wenden & Rubin, 1987). During the past decade or so, SLA researchers (e.g., O'Malley & Chamot, 1990; Oxford, 1990, 1996) have been developing an empirically based framework for analyzing learning strategies. Research on language-learning strategies has established the role learner strategies play in making language learning more efficient and successful (e.g., Chamot, 1993; Cohen, 1998; O'Malley & Chamot, 1990; Oxford, 1990; Rubin, 1975, 1987; Wenden & Rubin, 1987). Studies also have shown a positive association between proficiency level and the use of certain types of strategies, especially metacognitive (e.g., Bialystok, 1981; Flaitz & Feyten, 1996; Huang, 2004; Purpura, 1999), cognitive (e.g., Oxford & Ehrman, 1995), and compensation strategies (e.g., Dreyer & Oxford, 1996).

In the area of speaking, several studies have addressed how strategies can help learners develop their oral communication ability (e.g., Cohen & Olshtain, 1993; Cohen, Weaver, & Li, 1996; Dadour, 1995; Huang, 2004, Nunan, 1996; O'Malley & Chamot, 1990; Oxford, 1990). Much research has demonstrated the positive effects of strategy instruction on proficiency in speaking (e.g., Dadour & Robbins, 1996; Dörnyei, 1995; Feyten, Flaitz, & LaRocca, 1999; Nunan, 1996; O'Malley & Chamot, 1990; Oxford, 1990). Oxford and Ehrman's (1995) study also established a significant positive correlation between cognitive strategy use and speaking proficiency. Although some studies have concluded that learners with more proficiency use a greater variety and number of strategies (Anderson, 2005; Bruen, 2001; Green & Oxford, 1995; O'Malley & Chamot, 1990; Oxford & Burry-Stock, 1995; Wharton, 2000), the relationship between reported strategy use and performance is not clear-cut. While some researchers (e.g., Politzer & McGroarty, 1985) have found that some individual strategies correlate with language performance, they have found few statistically significant correlations between overall strategy use and language performance.

Research on Test-Taker Strategies in Language Testing

Similar to the findings from SLA, in the context of testing, the relationship between reported strategy use and proficiency and/or test performance is equally unclear. In Phakiti's (2003) study, test-takers' reported strategy use had a positive albeit weak relationship with performance on a reading test. In the context of a reading and limited-production writing test, Purpura (1998) found that high- and low-proficiency test-takers may use similar strategies but may perform differentially when using the same strategies. Song (2005) concluded that while the use of some strategies may enhance test performance, the use of others may have a negative impact on test performance; the use of still others may have no effect. In one of the few studies investigating strategic behaviors in a speaking test, Yoshida-Morise (1998) found that higher-proficiency test-takers used fewer communication strategies than did the lower-proficiency test-takers, who tended to use the strategies to compensate for their more limited speaking skills.

To the best of our knowledge, no research has examined the interaction among language proficiency level, reported strategic behaviors, and test performance in L2 speaking tests. The present study helps to fill this gap by providing empirical information concerning the relationships among these variables.

Present Study and Research Questions

Instead of focusing on only metacognitive and cognitive strategies, as the few strategy-use studies in the LT field have tended to do (e.g., Phakiti, 2003; Purpura, 1997, 1998; Song, 2005), we examine all speaking strategies[3] used during the communicative event (i.e., for the purpose of performing the six speaking tasks). The analysis uses a strategy-classification scheme based on Fulcher's (2003) summary of strategies for speaking in testing, the taxonomies and frameworks proposed by O'Malley and Chamot (1990) and Oxford (1990), and the work by Kæsper and Kellerman (1997), Paribakht (1985), Pressley and Afflerbach (1995), Purpura (1998), Yoshida-Morise (1998), and Yule and Tarone (1997). A synthesis of these strategies drawn from both the SLA and the LT fields was used as a starting point for this research (see Appendix A).

This study investigates the following four research questions:

1. Reported Strategic Behaviors
   When test-takers perform the SSTiBT, what strategic behaviors do they report using?

2. Reported Strategic Behaviors by Test-Taker Study and Proficiency Levels
   When test-takers perform the SSTiBT, are there differences in reported strategic behaviors, depending on their study level (graduate vs. undergraduate) and proficiency level (intermediate vs. advanced)?

3. Reported Strategic Behaviors by Task Groups
   When test-takers perform the SSTiBT, are there differences in reported strategic behaviors across task groups (A, B, and C)?[4]

4. Reported Strategic Behaviors and Test Performance

   When test-takers perform the SSTiBT, is there a relationship between their reported strategic behaviors and their test scores?

Method

Participants

The main study involved four groups of international students in Canada. As Figure 1 shows, Groups A and B included graduate students with advanced and intermediate levels of English-language proficiency, respectively, and Groups C and D consisted of undergraduate students with advanced and intermediate levels, respectively. This initial grouping of students in terms of English-language proficiency was based on a language-proficiency test administered at the beginning of the study (details about the test follow). Figure 1 shows the study's overall design.

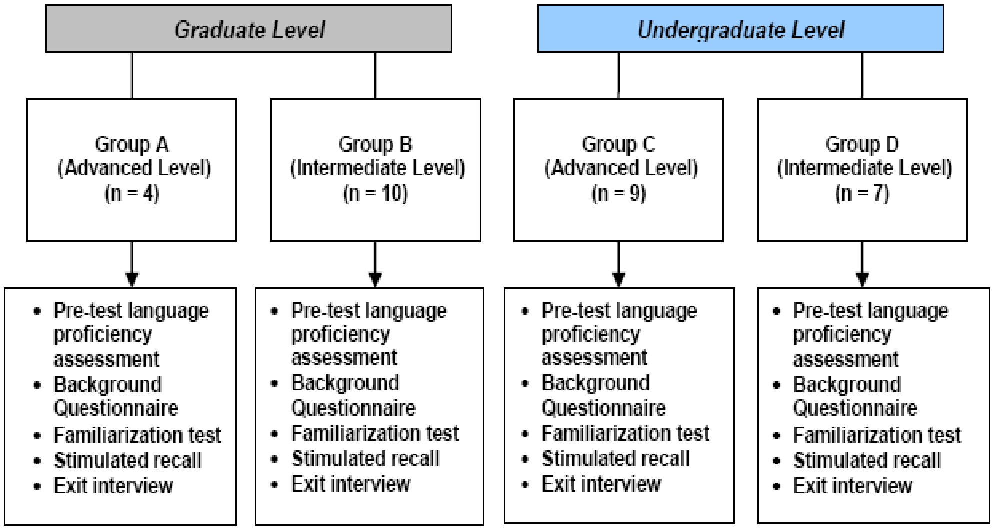

Graduate Level
Undergraduate Level
Group A (Advanced Level) (n = 4)
Group B (Intermediate Level) (n = 10)
Group C (Advanced Level) (n = 9)
Group D (Intermediate Level) (n = 7)
• Pre-test language proficiency assessment
• Background Questionnaire
• Familiarization test
• Stimulated recall
• Exit interview

***Figure 1.* Research design.**

Thirty individuals (14 graduate students and 16 undergraduate engineering students whose first language was Chinese) volunteered to participate in the main study. As Table 1 shows, the participants varied in terms of age (from 19 to 36 years), gender (19 males and 11 females), average length of stay in English-speaking countries, and English proficiency level (17 intermediate and 13 advanced). Table 2 reports descriptive statistics for test scores across participant groups and tasks. (Appendix B reports further analyses of the test scores in the present study.) Note that the scores included in the tables that report test scores were obtained by the participants when they took the research version of the SSTiBT; that is, they are the test scores obtained when the strategy data were collected. In other words, we did not use the pretest scores collected to categorize our data, but rather the scores obtained by the participants when they took the research version of the SSTiBT.

Table 1

Participants' Background

		Undergraduate (*n* = 16)	Graduate (*n* = 14)
Age range in years		19–22	25–36
Average length of stay in English-speaking countries		4.8 years	2.3 years
Gender	Female	6	5
	Male	10	9
English-proficiency level	Intermediate	7	10
	Advanced	9	4

Table 2

Descriptive Statistics for Test Scores by Student Study and Proficiency Level

Study level	Proficiency level	*N*	M^a	SD	Min	Max
Undergraduate	Intermediate	7	2.60	.30	2.17	2.92
	Advanced	9	3.41	.29	3.00	3.75
	Total	16	3.05	.50	2.17	3.75
Graduate	Intermediate	10	2.18	.49	1.33	2.92
	Advanced	4	3.02	.04	3.00	3.08
	Total	14	2.42	.56	1.33	3.08
Total	Intermediate	17	2.35	.46	1.33	2.92
	Advanced	13	3.29	.30	3.00	3.75
	Total	30	2.76	.61	1.33	3.75

[a] The test scores were averaged across six tasks. (For the SSTiBT, ETS sums the scores across tasks and then converts the total to a 0–30 point scale.)

We realize that it is important to collect data from members of different language groups. However, in order to (a) minimize learner variability, (b) enhance the strength of the conclusions that may be drawn with the resources available to us, and (c) deal with the issue of the representative nature of the respondents, we focused on participants from the same discipline (engineering) and whose first language is Chinese.[5] Our decision was based on the following considerations: (a) historically, Chinese-speaking international students have comprised the largest group of international students enrolled in the undergraduate and graduate programs from which the participants were drawn, (b) based on the TOEFL assessments' most recently published data summary, one of the largest groups of examinees has Chinese as its first language, and (c) since the second author, Huang, is proficient in Chinese and the two research assistants' first language is Chinese, we were able to elicit as much information as possible from the participants by allowing them to use their first language during the stimulated recall process and interviews.

Instruments

Language proficiency pretest. Two trained examiners assessed the oral proficiency of all participants in order to arrange the participants into intermediate and advanced groups, using the instrument in Appendix C. The same examiners independently rated the speech samples from Pilot Study 2 (see Data Collection, which follows) according to the scoring rubrics for the Speaking section of the TOEFL iBT(ETS, 2004) assessment established by the ETS. Next, the scores were checked for agreement between raters. Any disagreements were discussed until a 100% level of consistency was achieved. In the main study, the two raters independently scored the entire speech data set, and there were only three instances in which the scores showed a 0.5 range of difference. In the first case, one rater assigned a score of 2.5, and the other rated the proficiency level as being within the range of 2.0 and 2.5. In the second case, one assigned a score of 4, while the other gave 3.5. In the third case, the ratings were 2.0 and between 2.0 and 2.5. In those three cases, the test-takers' responses to the questions were discussed in order to establish agreement. The minor disagreements in these three cases did not affect the participant groupings because the advanced group members were those with scores of 3.0 and above, and the intermediate group members had scores from 2.0 to 2.5. Note that the test scores in Table 2 and Appendix B are from the research version of the SSTiBT and not from this language-proficiency pretest.

Background questionnaire. A questionnaire (Huang, 2004; see Appendix D) was distributed to all participants to collect information about their backgrounds and histories (e.g., gender, age, knowledge of other languages, educational experience, length of stay in English-speaking countries, oral test-taking experience).

The Speaking Section of the TOEFL iBT (SSTiBT). The SSTiBT is a speaking assessment tool that was developed to measure test-takers' oral communication skills in relation to their readiness for studies in colleges and universities in English-speaking countries. The test was delivered over the Internet and consisted of six speaking tasks classified into three groups in terms of the language skills they required. Table 3 lists the six tasks, their task groups, and the language skills each task required. The independent speaking tasks, Tasks 1 and 2 (Task Group A) required test-takers to respond to a question that elicited their thoughts or opinions on familiar topics that arose from their personal experience or background. Tasks 3 and 4 (Task Group B) integrated reading, listening, and speaking. These tasks included a short reading passage and a short talk and required test-takers to combine information from both the reading and listening material in their responses. Tasks 5 and 6 (Task Group C) integrated listening and speaking skills by having test-takers respond to listening material including a conversation or short lecture. Questions in Task Group C required test-takers to summarize key ideas from what they heard.

Table 3

Speaking Section of the TOEFL iBT (SSTiBT) Tasks and Language Skills Required

Task group	Task	Language skills required	Topic	TPT (in seconds)	TRT (in seconds)
A	1	Speaking	Familiar topic	15	45
	2	Speaking	Familiar topic	15	45
B	3	Speaking, Listening, & Reading	Campus-life situations	30	60
	4	Speaking, Listening, & Reading	Academic course content	30	60
C	5	Speaking & Listening	Campus-life situations	20	60
	6	Speaking & Listening	Academic course content	20	60

Note. TPT = total preparation time, TRT = total response time.

We used two different versions of the SSTiBT: (a) a familiarization version (a complete, timed form) administered to the participants so that they could become familiar with the test and the task types, and (b) a research version, which allowed us to pause after each task to facilitate stimulated recalls. All participants took the same familiarization and research versions of the SSTiBT. The six tasks were administered in the same order (as listed in Table 3) to all the participants.

Data Collection

Prior to the main study, we conducted two pilot studies. The first pilot study aimed to test the equipment and the organization of the sessions, while the second pilot study aimed to field-test the data-collection instruments and procedures of the main study. Based on the results of these two pilot studies, several changes were made in the design and implementation of the main study.

Pilot Study 1. In May 2005, we conducted a full-length, research version of the SSTiBT with one participant to test the computer system, the video and television equipment setup, and the seating arrangement, as well as to try out the stimulated recall session after each task. The test-taker was encouraged to use either English or Chinese during the stimulated recall session. The second author, Huang, trained a research assistant on all data-collection procedures prior to Pilot Study 1, and the research team was present to observe the entire process and provide feedback on areas needing change. We decided to implement the following three changes in Pilot Study 2.

1. We modified the physical setup to make it easier for test-takers to view the video playback. In Pilot Study 1, the computer was placed directly in front of the test-taker, who was easily distracted by the computer screen. This distraction diverted the test-taker's attention from viewing the test-taking process being shown on the television screen in the stimulated recall.

2. We adapted the computer configuration to enable the recording of test prompts. The test-taker said that he found it difficult to engage in stimulated recalls when he viewed himself listening to a dialogue or a lecture without any sounds that would provide the stimulus needed to recall what he was doing and thinking. Recording the test prompts (including the questions and listening-comprehension passages) helped

facilitate test-takers' recall of their thinking processes while they were listening to the prompts.

3. We changed the procedures to provide test-takers with an opportunity to practice doing stimulated recall. As practice, we used a short question (much like the first question in the SSTiBT) and asked the participant to recall what he was thinking before, during, and after he responded to the question. We found that the participant needed a question that would require greater processing than one that could be easily answered in a few sentences in order to practice doing the stimulated recall. We decided that, for Pilot Study 2, each participant would practice doing the stimulated recall immediately after completing the sixth task of the familiarization version and have an opportunity to ask any questions about what he or she would be asked to do in the research version, which was administered approximately one week later.

Pilot Study 2. The second pilot study was conducted in June 2005 in order to simulate the main study. We implemented the changes listed in the preceding section and performed a field test of all data-collection instruments and procedures. Six individuals volunteered to participate in the second pilot study and provided consent before the pretest proficiency screening. The pretest screening showed that two volunteers did not qualify to participate because their proficiency was at a beginner's level. The remaining four participants took the familiarization version of the SSTiBT approximately one week before taking the research version. At the end of the sixth task of the familiarization version, each participant engaged in a practice session of stimulated recall regarding the final task. For the research version, the testing time frame of 20 minutes for the SSTiBT was expanded to facilitate stimulated recall immediately after each task. Three participants returned for a semistructured exit interview, during which any areas that needed clarification were followed up.

In the second pilot study, we observed and noted the questions that participants raised while completing the questionnaire, performing the familiarization version of the test, doing the stimulated recall after completing the final item of the familiarization version, and answering the questions during the exit interview. As a result of Pilot Study 2, we made the following additional modifications in order to fine-tune the methodology.

1. During the stimulated recall, we let participants self-initiate replays and choose segments, which enabled them to verbalize freely in reaction to the tape. The research assistant also chose additional segments from the video and asked the participants to talk about what they were thinking at the time, as well as to clarify and expand on the information they provided.
2. We clarified the instructions the research assistant would give to the participants before and during the stimulated recall sessions to ensure that the participants would fully understand what to do, and that the research assistant would neither direct nor provide concrete reactions to the participants' responses. Also, the instructions were given in English and then translated into Chinese to ensure participants' full comprehension.
3. We modified the questions to be asked during the exit interviews to ensure that the participants would find the questions clear and understandable. Also, we stated the questions in both English and Chinese to make sure that the participants understood them.
4. We eliminated the static disturbances associated with the audio output and recording.
5. We moved to a new location and implemented the physical setup to better record the test-taking process and stimulated recall sessions, and to enhance the test-takers' viewing of the video playback. The setup is illustrated in Figures 2 and 3. Figure 2 shows the test-taker's position when he or she performed each of the six tasks in the SSTiBT. Figure 3 shows that the test-taker moved away from the computer after completing each task in the SSTiBT and turned to the researcher and television in order to engage in the stimulated recall of the task that he or she just performed. The first camera in Figures 2 and 3 was set up to capture the entire test-taking process, which then was played back on the television immediately after the test-taker completed each task. The second camera on the right captured all the stimulated recall sessions.

Main study. The main study was conducted from June to September 2005. Thirty participants volunteered to participate in the main study. First, we asked the respondents to give their informed consent to participate. We then administered the pretest proficiency assessment,

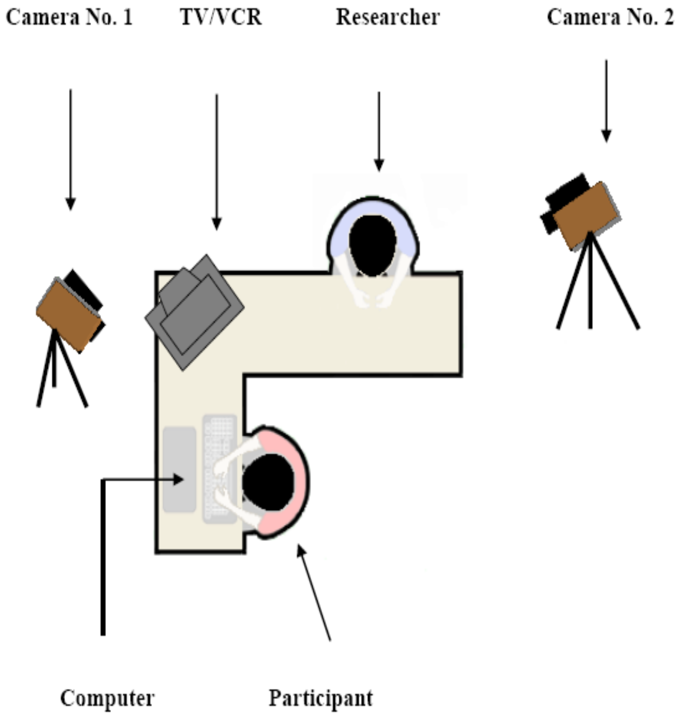


***Figure 2.* A graphic illustration of the physical setup during the test.**

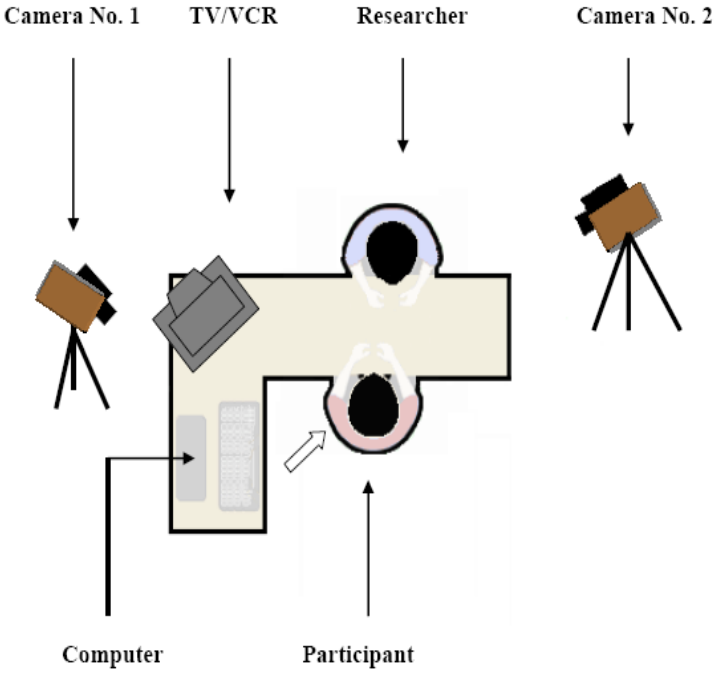


***Figure 3.* A graphic illustration of the stimulated recall sessions.**

and the participants completed the background questionnaire. Next, each participant took the familiarization version of the SSTiBT and engaged in a practice session of stimulated recall. Approximately one week after the familiarization version, we administered the research version of the SSTiBT to the participants. All the participants engaged in verbal reports through stimulated recall immediately after performing each of the six tasks contained in the SSTiBT, and they were offered an opportunity to take a break between Tasks 3 and 4, but no participant took up this offer. During the stimulated recall session, individual participants again were encouraged to speak in English or in Chinese, whichever came naturally when they were recalling their thoughts about what they did before, during, and after each speaking task. The participants also were reminded that they should report what they were thinking at the time, not what they thought they should have thought or done, or how they thought they should have responded (see Appendix E for the stimulated recall instructions). All testing sessions were completed in August 2005, and the responses from the research version of the SSTiBT were scored by ETS. Approximately two weeks after the research version was administered, all the participants returned for a semistructured exit interview, which addressed any of the test-takers' final thoughts. We also followed up on any areas that were not clear in the recordings or that needed clarification or elaboration.

Figure 4 provides a diagrammatic overview of the data-collection procedures implemented in the main study.

Coding Scheme

The coding scheme of the respondents' strategic behaviors was developed, drawing on the classification systems found in the literature across language skills and language testing, learning, and use contexts (see Appendix A). The strategies in the coding scheme (see Appendix F) were not limited to the categories listed in Appendix A, but rather emerged from the data of our pilot and main studies.

The coding scheme in Appendix F consists of five main categories of strategies: approach, communication, cognitive, metacognitive, and affective. Within each strategy category are individual strategies. For example, the approach strategy category includes individual strategies such as *recalling the task type*, *recalling the question*, *generating choices*, et cetera, that were coded as instances of strategies reported to approach the question. While developing the coding scheme, when we identified a strategy that did not exist on our list, we added it to the appropriate category along with a definition and an example for reference. Some individual strategies, such as *paraphrasing*, in the communication category, are further arranged into substrategies such as (a) *test-taker restating in another form or with other words to clarify meaning* and (b) *test-taker restating the thought in another form or with other words to avoid repetitions*. However, although the data were in some cases coded at the level of substrategy, for the data analyses, the substrategies were collapsed into their respective individual strategies.

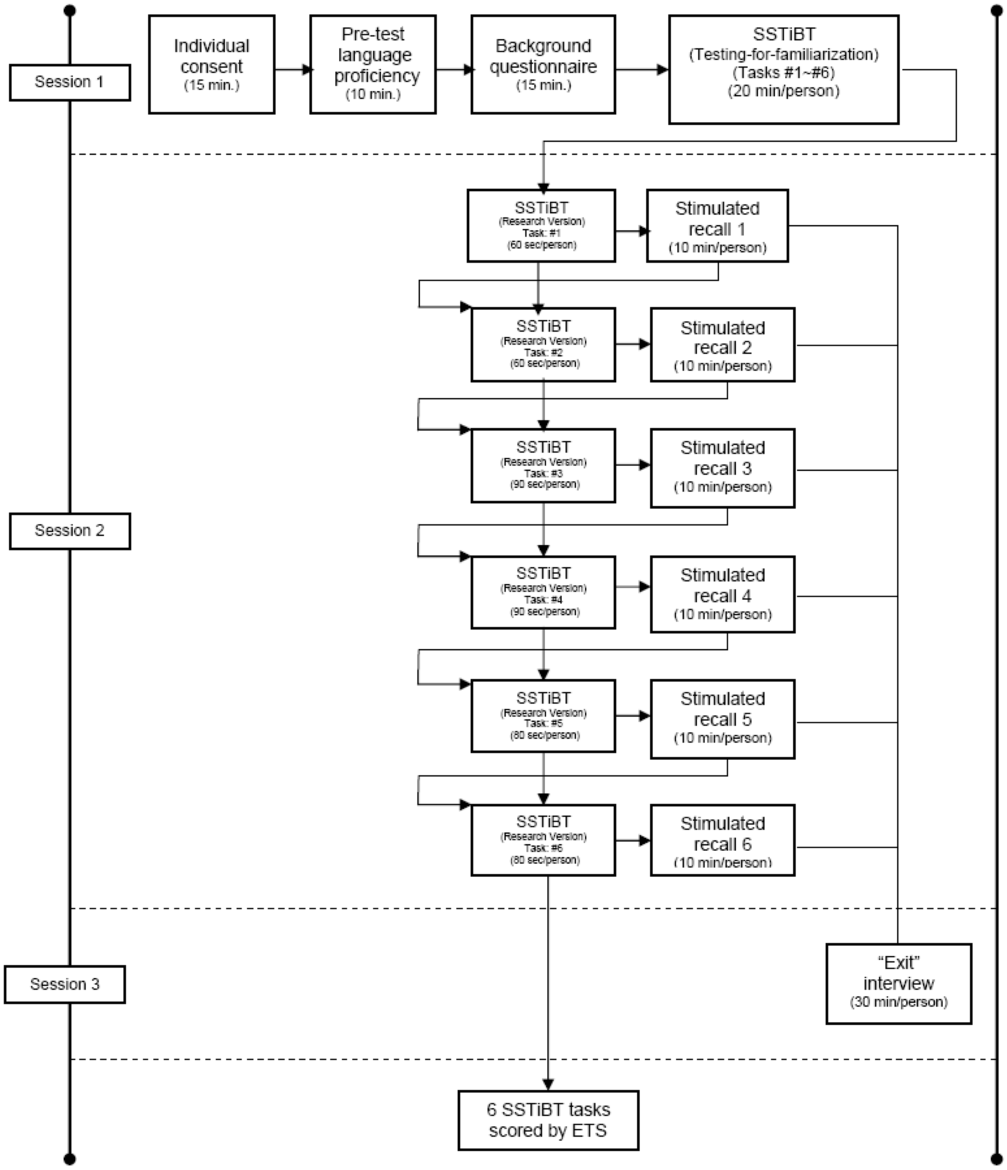

Session 1
Individual consent (15 min.)
Pre-test language proficiency (10 min.)
Background questionnaire (15 min.)
SSTiBT (Testing-for-familiarization) (Tasks #1~#6) (20 min/person)
Session 2
SSTiBT (Research Version) Task: #1 (60 sec/person)
Stimulated recall 1 (10 min/person)
SSTiBT (Research Version) Task: #2 (60 sec/person)
Stimulated recall 2 (10 min/person)
SSTiBT (Research Version) Task: #3 (90 sec/person)
Stimulated recall 3 (10 min/person)
SSTiBT (Research Version) Task: #4 (90 sec/person)
Stimulated recall 4 (10 min/person)
SSTiBT (Research Version) Task: #5 (80 sec/person)
Stimulated recall 5 (10 min/person)
SSTiBT (Research Version) Task: #6 (80 sec/person)
Stimulated recall 6 (10 min/person)
Session 3
"Exit" interview (30 min/person)
6 SSTiBT tasks scored by ETS

***Figure 4.* Data-collection procedures.**

In coding the data, when more than one code seemed to apply to a segment, we took the following actions:

1. We refined the coding. The coding scheme was refined to achieve a balance between being specific and being general—specific in capturing the strategic behaviors that participants used when completing the six tasks of the SSTiBT, and general in representing the strategic behaviors of more than one test-taker. For example, the strategy of *elaborating* was fine-tuned and expanded to two individual strategies to account for different reasons for elaboration: *elaborating to fill time* and *elaborating to clarify meaning*.
2. We split the segment and coded it as two segments. For example:

   我开始说的时候， 我就先把那题目 repeat 了一下，/ → *Borrowing*

   (Translation:[6] At the beginning of responding [to the question], I repeated the question again,/)

   后来我想我为什么要 repeat 它呢，浪费了我好多时间。 → Evaluating language production

   (Translation: Then I thought about why I repeated the question—it wasted so much of my time.)

This sentence involves two individual strategies: *borrowing* and *evaluating language production*. Here we used the symbol / to denote a segment boundary.

3. When they were sufficiently similar, we combined the codes into one individual strategy. For example:

   我这时候就瞄了一眼, 我觉得肯定是太多时间，/ → *Monitoring*: Test-taker monitoring production while it is occurring.

   (Translation: I peeked [at the clock], and I felt that there would be too much time left for sure . . ./)

   所以一边想一边讲，就是说，我在看下面还有几秒钟的时间，在想着剩余时间

> 我还可以讲点什么东西。→ *Monitoring*: Test-taker monitoring production vis-à-vis the clock while speaking.
>
> (Translation: So I was thinking and speaking at the same time; that is, I was looking at the number of seconds left and thinking about what else I could say in the time remaining.)

These two segments were fused into one individual strategy of *monitoring*, which is defined as "test-taker monitoring the clock while reading, listening, preparing, or speaking."

Data Coding

The verbal data generated from the stimulated recall sessions were fully transcribed and coded. We provided training for two research assistants (RAs) on data coding using the coding scheme, as well as coding using computer-assisted qualitative data-analysis software—NVivo. Having established intercoder agreement, the two RAs independently coded the verbal report responses for strategic behaviors. We based the inter-coder agreement on three tasks[7] in one transcript by calculating the number of agreements divided by the total number of coding decisions. The inter-coder agreement percentages (between the second author and the respective RAs) were 90% for RA1 and 93% for RA2. We discussed the coding decisions for which there was disagreement and resolved any discrepancies. Most disagreements occurred when there was more than one strategy in one segment, as described in 2 in the preceding section, or when the same strategies were counted more than once when the test-taker elaborated or repeated the same thought. The two RAs then each coded all the transcripts. Once the data coding was complete, the second author coded 10% of the transcripts randomly selected from each RA's set, and the overall inter-coder agreement percentage was an average of 86%.

Data Analysis

The coded data were tallied and percentages of reported individual strategies within each strategy category were computed for each test-taker for each task as follows: counts of coded individual strategies (e.g., *setting goals*) were summed for each test-taker for each task and then divided by the total number of instances of reported individual strategies for that particular test-taker for that particular task, to obtain a percentage of times that code occurred. These percentages served as the data for comparison across student groups and tasks.

Two issues that we had to address before conducting any statistical analyses on the coded data concerned (a) whether the coded data meet the statistical assumptions (e.g., normality of distribution) for such parametric tests as the *t*-test and ANOVA, and (b) the level of analysis for each research question.

In terms of statistical assumptions, Shapiro-Wilk tests on the percentages of the reported strategies by task indicated that the distributions were significantly different from normal for some categories (see Tables G3 and G4 in Appendix G). The distribution of test scores for some tasks (e.g., Tasks 1, 2, 5, and 6) were also not normally distributed, as the Shapiro-Wilk tests in Tables G1 and G2 indicate. As a result, a decision was made to use nonparametric statistical tests to address all research questions of the study.

To address Research Question 2 about differences across student groups in terms of reported strategies, we conducted Kolmogorov-Smirnov two-sample tests, a nonparametric equivalent of the two-sample *t*-test, with student group (advanced vs. intermediate, graduates vs. undergraduates) as the independent variable, and percentage of strategies reported as the dependent variable.

To answer Research Question 3 concerning the differences in percentages of reported strategies across the three task groups, we conducted a Friedman test, a nonparametric equivalent of a repeated-measures ANOVA, with task group as the independent variable and percentage of strategies reported as the dependent variable. Where a significant difference was detected, the Friedman test was followed by pairwise comparisons across task groups using Wilcoxon signed-rank tests, a nonparametric equivalent of the matched-pairs *t*-test.

To address Research Question 4 concerning the direction and magnitude of the relationship between the percentages of strategies reported and test scores, we conducted correlational analyses using the Spearman *rho* coefficient. All analyses were carried out using SPSS Version 14.

Because all the nonparametric statistical tests we used rely on rank rather than the value of scores and percentages, the measures of central tendency and dispersion that we report throughout the study (unless otherwise indicated) are the median (the midpoint in a distribution of values) and the range (the highest value minus the lowest value in a distribution), instead of the mean and standard deviation, which are usually reported with parametric tests.

The second issue we faced concerned the level of analysis for the different research questions. Because each participant performed six tasks, we had six percentages for each student for each individual strategy (i.e., one percentage per task). To be able to run the different statistical analyses described earlier, we needed to average these percentages in different ways depending on the research question we were addressing. Thus, for Research Question 1, where we compare strategy categories and reported individual strategies within and across strategy categories, we averaged the individual strategy percentages across the six tasks and all test-takers.

For Research Question 2, where we compare reported strategic behaviors across student groups, we averaged the individual strategy percentages across the six tasks for each student. For example, the percentage of the individual strategy *monitoring* for Student 21 was obtained by summing the percentages of this strategy for Student 21 across all six tasks and then dividing the total by 6. These averages were then used as the dependent variable in the Kolmogorov-Smirnov two-sample tests.

For Research Question 3, we wanted to examine whether there were differences in reported strategy use across task groups, comprised of pairs of tasks requiring the same language skills. Therefore, to address Research Question 3, Tasks 1 and 2, involving speaking only, were grouped together to form Task Group A; Tasks 3 and 4, involving reading, listening, and speaking, were grouped together as Task Group B; and Tasks 5 and 6, involving listening and speaking, were grouped together as Task Group C. For these analyses, the strategy percentages were averaged across *pairs* of tasks within each task group for each test-taker. For example, the percentages of individual strategies reported by each student for Tasks 1 and 2 were summed and then divided by 2 to obtain an average strategy percentage for Task Group A for each student. The Friedman test was then run using these averages as the dependent variable.

Finally, for Research Question 4, both aggregated (averaged) and unaggregated data were used. When examining the relationship between total test scores (i.e., averages of the six task scores) and percentages of reported strategies, we used averaged percentages of strategies across the six tasks for each student (i.e., as in Research Question 2). In examining the relationship between percentages of reported strategies and the task scores across pairs of tasks within task groups, aggregated data were used (i.e., as in Research Question 3). However, when examining the relationship between scores and percentages of reported strategies at the individual task level, we used unaggregated percentages of strategies reported by students while doing each individual task. The following section reports the results of these different analyses.

Results

Research Question 1: Reported Strategy Use

To answer the first research question, the frequencies of the individual strategies that test-takers reported using were analyzed by strategy category.[8] Overall, the test-takers used 49 different individual strategies across all tasks (see Table 4). The column labeled *raw frequency* lists the number of times test-takers reported using individual strategies. The column labeled *range* provides the maximum number of strategies reported minus the minimum (which in all cases is 0). The column labeled *% in relation to total number of strategies reported* indicates the percentage of each individual strategy in relation to the total number of strategies reported. The final column labeled *% in relation to strategy category* indicates the percentage of each individual strategy within its respective strategy category.

The highest percentage of reported strategy use by strategy category to the lowest was:

- The metacognitive category (33.42%)
- The communication category (26.48%)
- The cognitive category (25.04%)
- The approach category (11.43%)
- The affective category (3.63%)

As the last column in Table 4 shows, the most frequently reported individual strategy within the approach category was *developing reasons* (29.77%). The most frequently reported strategy within the communication category was *organizing thoughts* (26.02%). The most frequently reported individual strategy within the cognitive category was *using mechanical means to organize* (44.39%). The most frequently reported individual strategy within the metacognitive category was *evaluating the content of what was read/heard* (18.67%). The most frequently reported individual strategy within the affective category was *justifying performance* (45.13%).

Table 4

Frequencies and Percentages of Reported Use of Individual Speaking Strategies

	Raw frequency[a]		% in relation to total number of strategies reported	% in relation to strategy category
	Total	Range		
Approach	309	8	11.43	
Recalling the task type	26	2	.74	8.41
Recalling the question	49	4	1.63	15.86
Recalling the text	8	2	.18	2.59
Recalling the dialogue	43	3	1.37	13.92
Recalling the lecture	35	3	1.03	11.33
Generating choices	18	1	1.01	5.83
Making choices	38	2	2.02	12.30
Developing reasons	92	4	3.45	29.77
Communication	757	14	26.48	
Simplifying the message	21	2	.70	2.77
Avoiding	15	3	.59	1.98
Using Chinese	10	3	.34	1.32
Paraphrasing	15	2	.44	1.98
Approximating	9	2	.33	1.19
Linking to prior experiences/knowledge	**184**	**5**	**6.06**	**24.31**
Borrowing	35	3	1.01	4.62
Reviewing notes	39	2	1.25	5.15
Referring to notes	56	2	1.94	7.40
Organizing thoughts	**197**	**5**	**7.46**	**26.02**
Guessing	23	3	.77	3.04
Repeating	26	2	1.12	3.43
Rehearsing	13	1	.41	1.72

(Table continues)

Table 4 (continued)

	Raw frequency [a]		% in relation to total number of strategies reported	% in relation to strategy category
	Total	Range		
Reading ahead	36	1	1.12	4.76
Restructuring	12	1	.46	1.59
Slowing	16	3	.61	2.11
Thinking ahead	5	1	.11	0.66
Elaborating to fill time	31	2	1.22	4.10
Elaborating to clarify meaning	14	2	.52	1.85
Cognitive	748	13	25.04	
Attending	**151**	**5**	**5.17**	**20.19**
Anticipating the content	97	3	2.96	12.97
Anticipating the structure	87	3	2.69	11.63
Using imagery	11	2	.32	1.47
Using mechanical means to organize information	**332**	**7**	**11.68**	**44.39**
Memorizing	9	1	.29	1.20
Summarizing	34	2	1.06	4.55
Translating	5	3	.18	0.67
Inferencing	21	2	.66	2.81
Processing inductively	1	1	.03	0.13
Metacognitive	932	13	33.42	
Setting goals	**113**	**4**	**3.95**	**12.12**
Identifying the purpose of the task	75	2	2.92	8.05
Planning	**149**	**5**	**5.88**	**15.99**
Monitoring	**139**	**4**	**5.19**	**14.91**
Self-correcting	26	3	.88	2.79
Evaluating previous performance	39	3	1.14	4.18

(Table continues)

Table 4 (continued)

	Raw frequency [a]		% in relation	% in relation
	Total	Range	to total number of strategies reported	to strategy category
Evaluating the content of what was read/heard	**174**	**7**	**5.45**	**18.67**
Evaluating performance	**109**	**4**	**4.23**	**11.70**
Evaluating language production	**108**	**4**	**3.78**	**11.59**
Affective	113	5	3.63	
Lowering anxiety	28	2	.86	24.78
Encouraging self	34	3	1.14	30.09
Justifying performance	51	4	1.63	45.13

Note. Because the coding scheme was developed in part using the data from the pilot study, not all of the individual strategies listed in Appendix F appear in Table 4. In addition, only the individual strategies (not the substrategies found in Appendix F) are listed in this table. Individual strategies in bold are the 10 most frequently reported.

[a] The total number of all individual strategies reported across all tasks and test-takers was 2,859 strategies (min = 5, max = 35).

As the last column in Table 4 shows, the most frequently reported individual strategy within the approach category was *developing reasons* (29.77%). The most frequently reported strategy within the communication category was *organizing thoughts* (26.02%). The most frequently reported individual strategy within the cognitive category was *using mechanical means to organize* (44.39%). The most frequently reported individual strategy within the metacognitive category was *evaluating the content of what was read/heard* (18.67%). The most frequently reported individual strategy within the affective category was *justifying performance* (45.13%).

The 10 most frequently reported individual strategies (bolded in Table 4) were:

1. Cognitive: *using mechanical means to organize information* (11.68%)
2. Communication: *organizing thoughts* (7.46%)
3. Communication: *linking to prior experiences/knowledge* (6.06%)

4. Metacognitive: *planning* (5.88%)

5. Metacognitive: *evaluating the content of what was read/heard* (5.45%)

6. Metacognitive: *monitoring* (5.19%)

7. Cognitive: *attending* (5.17%)

8. Metacognitive: *evaluating performance* (4.23%)

9. Metacognitive: *setting goals* (3.95%)

10. Metacognitive: *evaluating language production* (3.78%)

Of the 10 most frequently reported strategies, 6 fall into the metacognitive category (28.48%), 2 fall into each of the cognitive (16.85%) and communication (13.52%) categories, and none falls into the approach or affective categories.

Finally, to examine the relationships among the strategy categories, we calculated their intercorrelations. As shown in Table 5, the only significant relationships were negative and occurred in three cases: the communication and cognitive categories were significantly negatively correlated, as were the approach and metacognitive categories and the communication and metacognitive categories. These negative and significant correlations indicate that, overall, participants who reported more communication strategies tended to report fewer cognitive and metacognitive strategies, and vice versa. Similarly, participants who reported more approach strategies tended to report fewer metacognitive strategies, and vice versa.

Table 5

Correlations Among Strategy Categories

	Approach	Communication	Cognitive	Metacognitive	Affective
Approach	1.00				
Communication	.09	1.00			
Cognitive	−.23	−.37*	1.00		
Metacognitive	−.43*	−.72**	−.05	1.00	
Affective	−.04	−.32	−.26	.27	1.00

Note. Spearman *rho*, $N = 30$.

* Correlation is significant at $p < .05$ (2-tailed). ** Correlation significant at $p < .01$ (2-tailed).

Research Question 2: Reported Strategy Use by Test-Taker Proficiency and Study Levels

To answer the second research question concerning differences in reported strategic behaviors depending on test-takers' study level and proficiency level, we compared the reported strategic behaviors between groups of test-takers based on averaged strategy percentages across tasks for each student. We compared students across study levels (undergraduate vs. graduate) and proficiency levels (intermediate vs. advanced). The results are presented in three parts: (a) undergraduate vs. graduate groups, (b) intermediate- vs. advanced-level groups, and (c) the subgroupings (i.e., undergraduate–intermediate vs. undergraduate–advanced; graduate–intermediate vs. graduate–advanced; intermediate–undergraduate vs. intermediate–graduate; advanced–undergraduate vs. advanced–graduate).

Reported strategies by test-taker study level. Table 6 presents the descriptive statistics for test-takers' reported strategy use at the undergraduate and graduate levels by strategy category. The largest difference in the medians between the study-level groups is in the communication category, followed by the cognitive and metacognitive categories. To examine whether these differences in medians are statistically significant, we conducted a two-sample Kolmogorov-Smirnov test on the medians of strategy categories across test-taker study levels. The results are reported in Table 7.

Table 6

Reported Strategy Use by Strategy Category and Test-Taker Study Level

Study level		Approach	Communication	Cognitive	Metacognitive	Affective
Undergraduate	Median	11.63	29.83	22.57	29.51	2.08
(*n* = 16)	Range	16.42	27.20	16.06	20.37	8.24
Graduate	Median	10.00	20.36	27.55	33.79	3.83
(*n* = 14)	Range	13.78	27.14	17.61	36.13	7.68
Total	Median	11.32	28.98	25.15	31.54	3.19
(*n* = 30)	Range	20.43	38.47	20.58	36.13	9.23

Note. Medians and ranges are based on percentage of reported strategy use.

As shown in Table 7, there are significant differences between the study groups for three strategy categories: communication, cognitive, and affective. For the communication category, undergraduates reported significantly more communication strategies ($z = 1.37$, $p < .05$). This is due mainly to the difference between the medians for the individual strategy *organizing thoughts* (*Mdn* = 10.86 for undergraduates vs. 2.71 for graduates; see Appendix H).

For the cognitive category, the graduates reported significantly more cognitive strategies ($z = 1.42$, $p < .05$) than the undergraduates reported. The individual strategy that shows the greatest difference between the two groups is *attending* (*Mdn* = 6.50 for graduates and 2.31 for undergraduates; see Appendix H).

Table 7

Two-Sample Kolmogorov-Smirnov Test for Reported Strategy Use by Test-Taker Study Level

		Approach	Communication	Cognitive	Metacognitive	Affective
Most	Absolute	.31	.50	.52	.43	.50
extreme	Positive	.00	.00	.52	.43	.50
differences	Negative	−.31	−.50	.00	−.07	−.11
Kolmogorov-Smirnov *Z*		.85	1.37	1.42	1.17	1.37
Asymp. sig. (2-tailed)		.459	.048	.036	.129	.048
Effect size (*r*) [a.]		.16	.25	.26	.21	.25

[a.] Following Field (2005), we used Pearson’s correlation coefficient *r* as a measure of effect size. This coefficient is constrained to lie between 0 (*no effect*) and 1 (*a perfect effect*). Following Cohen (1988), Field suggested the following guidelines for interpreting effect sizes: small effect: $r = .10$, medium effect: $r = .30$, and large effect: $r = .50$ (Field, 2005, p. 32).

For the affective category, the graduates reported significantly more affective strategies ($z = 1.37$, $p < .05$) than the undergraduates reported. The individual strategy that shows the greatest difference between the two groups is *justifying performance* (*Mdn* = 2.50 for the graduates and .88 for the undergraduates; see Appendix H). Table 7 also reports the effect size for each strategy category. Note that for the three strategy categories (communication, cognitive, and affective), the effect size is less than .30, suggesting a small effect of study level on reported strategy use.

In the metacognitive and approach categories, there were no significant differences between the study-level groups. However, an examination of individual strategies in the metacognitive category (Appendix H) shows that, of the nine individual strategies in this category, the undergraduate group has higher medians in four categories (*identifying the purpose of the task, monitoring, self-correcting,* and *evaluating language production*), while the graduate group has higher medians in five (*setting goals, planning, evaluating the content of what was read/heard, evaluating previous performance,* and *evaluating performance*). These differences across student groups in terms of individual strategies seem to cancel out any differences and to explain the lack of significant differences in terms of the overall metacognitive category across test-taker study levels.

Reported strategies by test-taker proficiency level. Table 8 presents the descriptive statistics for test-takers' reported strategy use at the intermediate and advanced proficiency levels. It shows no large median differences across the two student groups. As shown in Table 9, the two-sample Kolmogorov-Smirnov test on the medians of strategy categories detected no statistically significant differences between the proficiency-level groups.

Table 8

Reported Strategy Use by Strategy Category and Test-Taker Proficiency Level

Proficiency level		Approach	Communication	Cognitive	Metacognitive	Affective
Intermediate	Median	11.64	28.96	25.56	31.54	3.68
(*n* = 17)	Range	15.47	37.68	17.89	35.29	8.44
Advanced	Median	10.22	29.01	24.26	32.09	2.71
(*n* = 13)	Range	20.43	31.83	19.78	29.32	7.76

Note. Medians and ranges are based on percentage of reported strategy use.

Table 9

Two-Sample Kolmogorov-Smirnov Test for Reported Strategy Use by Test-Taker Proficiency Level

		Approach	Communication	Cognitive	Metacognitive	Affective
Most	Absolute	.34	.24	.20	.18	.25
extreme	Positive	.16	.24	.17	.13	.15
differences	Negative	−.34	−.16	−.20	−.18	−.25
Kolmogorov-Smirnov *Z*		.92	.65	.55	.48	.68
Asymp. sig. (2-tailed)		.36	.79	.92	.98	.75

Reported strategies by test-taker proficiency level and study level. Table 10 presents the descriptive statistics for the following subgroupings:

- Undergraduate–intermediate versus undergraduate–advanced
- Graduate–intermediate versus graduate–advanced
- Intermediate–undergraduate versus intermediate–graduate
- Advanced–undergraduate versus advanced–graduate

We ran a Kolmogorov-Smirnov (K-S) two-sample test for each of these pairs of student groups. None was significant at the $p < .05$ level.

Table 10

Reported Strategy Use by Strategy Category and Test-Taker Study Level and Proficiency Level

Study level	Proficiency		Approach	Communication	Cognitive	Metacognitive	Affective
Undergraduate	Intermediate	Median	15.27	29.95	25.18	28.43	1.45
	(*n* = 7)	Range	12.35	26.41	10.79	18.38	7.44
	Advanced	Median	11.27	29.13	21.99	30.55	2.71
	(*n* = 9)	Range	15.06	25.67	15.25	19.14	7.76
Graduate	Intermediate	Median	11.44	20.36	27.55	33.79	3.83
	(*n* = 10)	Range	12.90	27.14	14.91	31.87	7.68
	Advanced	Median	9.34	20.47	31.39	37.93	3.74
	(*n* = 4)	Range	7.76	18.54	12.85	29.32	2.69

Note. Medians and ranges are based on percentage of reported strategy use.

Research Question 3: Reported Strategy Use by Task Group

In this section we examine the relationships between task groups in the SSTiBT and the strategies that the test-takers reported. As indicated in Table 3, the six tasks in the SSTiBT fall into three groups, A (Tasks 1 and 2), B (Tasks 3 and 4), and C (Tasks 5 and 6), that differ in terms of the language skills they require. Task Group A requires only speaking skills, while Task Groups B and C integrate two or more language skills each. Task Group C requires listening and speaking skills, while Task Group B requires listening, reading, and speaking skills. All analyses reported in this section were conducted on averaged strategy percentages across tasks within each task group for each test-taker (e.g., the percentages of individual strategies reported by each student for Tasks 3 and 4 were summed and then divided by 2 to obtain an average strategy percentage for Task Group B for each student).

Task group and reported strategic behaviors. Table 11 provides the medians and ranges of the averaged percentages of reported strategy use across task groups. It shows that Task Group A resulted in a higher median in terms of both approach and communication strategies than Task Groups B and C. Task Group A elicited slightly more metacognitive strategies than Task Group B as well. Task Groups B and C, on the other hand, elicited more cognitive and affective strategies than Task Group A. Note also that, as the last column in Table 11 shows, the participants reported more strategies while doing tasks in Group B than when doing tasks in Groups C and A.

Table 11

Overall Reported Strategy Use by Task Group

Task group		Approach	Communication	Cognitive	Metacognitive	Affective	Total [a]
A	Median	19.09	32.89	9.06	35.19	.00	10.00
	Range	34.76	50.89	41.67	52.53	21.43	25.50
B	Median	8.19	23.58	32.24	28.23	4.26	20.75
	Range	18.00	38.48	28.63	46.35	11.69	21.00
C	Median	7.08	25.02	31.97	32.12	2.51	15.50
	Range	22.40	43.56	43.13	52.08	15.00	15.50

Note. Medians and ranges are based on percentage of reported strategy use.

[a] Figures in this column are based on raw frequencies, not percentages, of strategies reported.

To examine whether the differences in the medians of the five strategy categories across task groups are statistically significant, we conducted Friedman tests using task group as the independent variable and averaged percentages of reported strategy use as the dependent variables. The results are reported in Table 12. The test was significant for the approach (X^2 (2, N = 30) = 24.82, $p < .01$), communication (X^2 (2, N = 30) = 8.60, $p < .05$), and cognitive (X^2 (2, N = 30) = 39.98, $p < .01$) categories. Follow-up pairwise comparisons were conducted using Wilcoxon signed-rank tests. The results of these tests are presented in Table 13. A Bonferroni correction was applied, so all effects are reported at a .0167 (.05/3) level of significance. Table 13 also reports the effect size for each pairwise comparison for each strategy category.

Table 13 shows that the median for the approach category for Task Group A was significantly higher than the medians for Task Groups B and C (*p* < .0167); in both cases *r* > .50, indicating a large effect size. The median for communication for Task Group A was also significantly higher than for Task Group B (*p* < .0167), with a medium effect size (*r* = .40), but not Task Group C (*p* > .0167). Finally, the medians for the cognitive category for Task Groups B and C were both significantly higher than for Task Group A (*p* < .0167); in both cases *r* = .60, indicating a large effect size. There were no significant differences (*p* > .0167) between the medians for Task Group B and Task Group C in terms of the three strategy categories: approach, communication, and cognitive.

Table 12

Friedman Tests for Strategy Category by Task Group

	Approach	Communication	Cognitive	Metacognitive	Affective
Chi-square	24.82	8.60	39.98	3.80	3.94
Df	2	2	2	2	2
Asymp. sig.	.00	.01	.00	.15	.14

Table 13

Follow-Up Tests for Strategy Category by Task Group

Task group		Approach	Communication	Cognitive	Metacognitive	Affective
A vs. B	Z [a]	−4.29	−3.08	−4.66	−2.38	−1.98
	Sig. [b]	.00	.00	.00	.02	.05
	r [c]	.55	.40	.60	.31	.26
A vs. C	Z	−4.33	−2.34	−4.64	−1.59	−.71
	Sig.	.00	.02	.00	.11	.48
	r	.56	.30	.60	.21	.09
B vs. C	Z	−.14	−.90	−1.31	−.57	−1.66
	Sig.	.89	.37	.19	.57	.10
	r	.02	.12	.17	.07	.22

[a] Wilcoxon signed-rank test. [b]Asymp. sig. (2-tailed). [c] Effect size.

An examination of the individual strategies (Appendix I) indicates that three individual strategies within the approach category (*generating choices, making choices,* and *developing* *reasons*) have higher medians for Task Group A than for Task Groups B and C. Appendix I shows also that Task Group A has a higher median for one communication strategy, *organizing thoughts*, while Task Group B led to a slightly higher median for *reading ahead*, under the communication category. Task Group C generated more *referring to notes*. In terms of the cognitive category, Task Groups B and C have higher medians than Task Group A for four individual strategies: *using mechanical means to organize information, anticipating the structure, anticipating the content,* and *attending*.

Finally, while the Friedman test did not detect any significant differences between the medians of task groups for the metacognitive and affective categories (Table 12), there were some relatively large differences in the medians of some individual strategies across task groups. For example, Task Group A has higher medians for *planning*, *monitoring,* and *evaluating performance* (*Mdn* = 5.44, 6.63, and 6.30, respectively) than Task Groups B (*Mdn* = 4.01, 4.12, and 2.22, respectively) and C (*Mdn* = 3.71, 2.79, and 2.36, respectively), while Task Groups B and C have higher medians for *evaluating the content of what was read/heard* (*Mdn* = 6.37 and 5.75) than Task Group A (*Mdn* = .00). Task Group C elicited three other metacognitive strategies more frequently than the other two task groups as well: *setting goals* (*Mdn* = 4.01), *identifying the purpose of the task* (*Mdn* = 3.59), and e*valuating language production* (*Mdn* = 4.50). Finally, Appendix I shows that, under the affective category, Task Group B elicited more *justifying performance* (*Mdn* = 1.70) than were elicited by Task Groups A and C (*Mdn* = .00 each). It is also worth noting that the medians for several individual strategies for Task Group A were 0 (Appendix I), unlike those for Task Group C and, particularly, Task Group B. This suggests that Task Group B typically elicited a wider variety of individual strategies than were elicited by Task Group C, which in turn seems to have elicited a wider variety of individual strategies than were elicited by Task Group A.

Table 14 lists the five individual strategies that have the highest medians for each task group, as reported in Appendix I. Among them, two strategies (*organizing thoughts* and *using mechanical means to organize information*) are common across the three task groups, while three are unique to Task Group A. Note also that four individual strategies are listed for both Task Group B and Task Group C, though in a slightly different order.

Table 14

Top Five Individual Strategies by Task Group

Task group	Skills required	Individual strategies	Median
A (Tasks 1–2)	Speaking	Communication: Organizing thoughts Cognitive: Using mechanical means to organize information Metacognitive: Monitoring Metacognitive: Evaluating performance Approach: Making choices	9.71 6.90 6.63 6.30 5.65
B (Tasks 3–4)	Reading, Listening & Speaking	Cognitive: Using mechanical tools to organize information Communication: Linking to prior experiences/knowledge Metacognitive: Evaluating the content of what was read/heard Cognitive: Attending Communication: Organizing thoughts	12.71 6.92 6.37 5.22 5.61
C (Tasks 5–6)	Listening, & Speaking	Cognitive: Using mechanical tools to organize information Communication: Organizing thoughts Metacognitive: Evaluating the content of what was read/heard Communication: Linking to prior experiences/knowledge Metacognitive: Evaluating language production	12.86 6.36 5.75 5.28 4.50

The next section provides examples of individual strategies that were reported frequently. The first excerpt is an example of *using mechanical means to organize or remember information* from student GYGW[9] while doing Task 4:

Excerpt 1: 但是因为这个 topic 我一看对我来说太不熟悉，(???)太不熟悉，我就知道它是个 tough task 对我来说，因为这个<social interactive>我就开始记，我怕我忘掉了，因此我在这里做的笔记包括, 在读的过程当中，social interaction,　influence behavior because 我不知道它后来它会 focus on which part and audience effect，他们对我说是比较陌生的东西，我可能，从理解的角度上会更难以理解一些，所以我就会[举起笔记示意]记了这么多。把中间要用到的词，跟我想到的词记下来。

(GYGW, Task 4) (Translation: Because I was not familiar with this topic, I knew that it would be a tough task for me. So I started writing down notes because I was afraid that I would forget. While I was reading, I wrote down here “social interaction,” “influence behavior” because I did not know which part would be the question’s focus. Due to unfamiliarity with the subject matter, it was more difficult for me to comprehend. So I wrote down so many notes. [showing the notepad] I wrote down words that I would need and words that I could think of at the time.)

Organizing thoughts was also common to the three task groups. For example, GYL reported for Task 3:

Excerpt 2: 这时候就想，接下来就说 his reason 吧，因为没有组织好，就上来就开 始 reasons, 然后一想，这 reason 也得一、二、三说。前头先说一个，它一共有几 个。嗯。

(GYL, Task 3) (Translation: At that time, I was thinking that I should talk about the reason next, but I did not organize the points well. I started to talk about the reasons. Then I thought that the reasons should also have points one, two, and three. So I then mentioned how many reasons there were at the beginning.)

For Task Group A (Tasks 1 and 2), which did not require test-takers to listen to a passage or read a text, but for which they needed to self-generate a response by drawing on their own knowledge or ideas, the following strategies were reported most often: *evaluating performance* (e.g., Excerpt 3), *monitoring* (e.g., Excerpt 4), and *making choices* (e.g., Excerpt 5).

Excerpt 3: 这个题目太- 太 unpredictable，我觉得，我第二题做得不好，觉得我不该花 15 到 25 秒的时间来重复它的题目。我考虑的可能没有必要，可能因为有个自我评估嘛。这个不是一个好的解题方式，我就在想我可能在下面会寻求一些变化。(GWL, Task 2) (Translation: This task is so, so unpredictable. I felt that I did not do well on the second task. I felt that I shouldn't have spent 15 to 25 seconds on restating the question. I thought that that use of time was unnecessary. This thought may have arisen because I was self-evaluating how I did and realized that my method of responding to the question was not a good one. I was thinking that I would probably make some changes in the way I responded to the subsequent tasks.)

Excerpt 4: 可是我后来看了一下时间，好像还 ok。我说干脆就，我就不想 organize 这个 point，我就干脆讲的慢一点，把他的 detail point 讲出来。(UBT, Task 2) (Translation: Then I took a look at the clock and thought that it was fine. I then thought that I might as well not organize the point. I would just go slowly in order to deliver the point in detail.)

Excerpt 5: 我就在想一个 place，如果我有经常去的地方,我还有东西可说。可我没有经常去的地方，我说什么呢? 我想了三个地方，一个 museum，一个 LIBRARY，一个 shopping mall。然后我想 museum 其实我不经常去。但刹那间我觉得 museum 词汇太复杂了，整理起来比较麻烦，我想! 算了,说一个能说出点东西的地方，一个可拿点分的地方，我想我还是选 library。可以 find article，上网…。(GSX, Task 1) (Translation: I was trying to think of a place. If there were a place where I often go, then I would have something to say. But, there wasn't any place where I often go. So I thought, What can I talk about? I thought of three places: the museum, the library, and the shopping mall. Then I thought that, in fact, I don't go to the museum all that often. Then, all of a sudden, I felt that the vocabulary needed for me to talk about going to the museum was too complicated, and it would be too troublesome to organize what I wanted to say. I thought that I should just forget it! I would choose a place that I could say something about, and so be certain that my answer would get me some points. I thought that I would choose the library, a location that would enable me to talk about finding articles, surfing the Internet, and so on.)

Task Groups B and C (Tasks 3–6), which both required test-takers to listen to a dialogue or a monologue before responding, resulted in the more frequent report of two strategies: *evaluating the content of what was read and/or heard* (e.g., Excerpt 6) and *linking to prior experiences or knowledge* (e.g., Excerpt 7):

Excerpt 6: 脑子里闪过，他讲的这些东西到底是不是 true。(ULS, Task 4)

(Translation: A thought flashed through my mind about whether what the speaker said was true or not.)

Excerpt 7: 他这个 topic 出来之后，看了一下之后，突然觉得挺高兴的， 因为这个东西以前在 society 课上学过，我在想说，因为像这种考试的话，他给你一个topic，你如果知道的话，肯定你就 understand。我就想说，这个挺不错的，知道怎么回事，而不是出来一个东西，不知道是什么。(UJZ, Task 4) (Translation: When the topic came up, I was pretty happy after seeing it, because I had learned about it in a sociology class before. I was thinking that, if you are given a topic that you are familiar with in a test like this, you can understand everything. I was thinking that this is not bad, knowing what's going on, rather than not knowing what it's about when a topic comes up.)

The similarity across Task Groups B and C in terms of the strategies reported by the participants is most obvious when we compare Tasks 4 and 6, which both required test-takers to listen to a lecture. In both tasks, the test-takers often made associations between their personal experience and knowledge and what they were reading and/or hearing, as the following three excerpts show:

Excerpt 8: 后来我想说，他在讲那种绑鞋带那个, 会 tend to make 更多 mistakes 这样的, 我在想，实际上是说, 他讲的这些东西, 也不一定是 audience effect, 我本来做事做快一点，我的 possibility to make mistakes 就高一些。(UBT, Task 4) (Translation: Then I was thinking . . . when he was talking about tying shoelaces and about the tendency to make more mistakes, I was thinking that, in fact, what he talked about might not necessarily be related to *audience effect*. When I try to do things faster, the possibility of making mistakes is correspondingly higher.)

Excerpt 9: 那个时候我就想到-这个好象是自己经历过的。不一定就是自己系鞋带 什么，就是这种情景自己经历过的。如果有人看着你，就一定要做好怎么样怎么样，但是那个时候我又想，既然他这么说呢，我也在旁人监视下，然后呢我就想我就慢慢 地自己做。(UJG, Task 4) (Translation: At that time, I was thinking that this seemed to be what I had experienced before. It was not necessarily about my tying shoelaces, but that I also had experienced similar situations before. That is, if someone was watching you, you would want to do well and whatnot, but, then again, I was thinking that, after listening to what the speaker had said, I realized that I was also being watched, and then I thought that I would want to do things SLOWLY [when I performed the test, to avoid making mistakes].)

Excerpt 10: 是听过他讲以后，我就会联想到以前，就是说，关于这样的东西，我知道的一些。然后，我会做一下很短的回忆，就是，比如听到这个 money， 讲钱的东西，我会，就是我会想起，我以前听到过讲钱的东西是什么。(UJZ, Task 6) (Translation: Listening to the speaker led me to think about things related to the talk and some things that I already knew. Then I did a very brief thinking back—for example: when I listened to the talk about money, I would think of what money-related talks I had heard of before.)

In addition, Task Group B (Tasks 3 and 4) elicited more reported instances of the strategy *attending* (e.g., Excerpt 11), while Task Group C led to frequent use of the strategy *evaluating language production* (e.g., Excerpt 12).[10]

> Excerpt 11: 我就写男生的 . . . 因为我发现女生讲话很少，她只是在 continue 那个 conversation . . . 。女生好像只是 repeat 那个 argument . . . 。(UBT, Task 3)
>
> (Translation: I was writing about what the male speaker said . . . because I noticed that the female speaker said very little; she was merely continuing the conversation . . . 。She seemed to be only repeating that [the male speaker’s] argument.)
>
> Excerpt 12: 就是在我说一件事儿的时候，嗯，(. . .) 我应该 focus on answer whatever they ask . . . 。(UMW, Task 5) (Translation: When I was stating the event, er . . . I should focus on answering whatever I was asked.)

Overall, these findings indicate that the integrated tasks (Task Groups B and C) were more similar to each other but differed from the independent tasks (Task Group A) in terms of the strategies they elicited. In addition, the integrated tasks typically elicited a wider variety of individual strategies than the independent tasks elicited.

Reported strategy use by task. Table 15 reports the descriptive statistics for reported strategy use across individual tasks. It shows that, in general, the integrated tasks (Tasks 3–6) elicited more reported strategy use than the independent tasks (Tasks 1 and 2) elicited. Table 15 shows also that there were some differences across tasks within task groups in terms of percentage of reported strategy use. For example, Task 1 has a higher median than Task 2 has in terms of the approach category, while Task 2 elicited more communication, cognitive, and metacognitive strategies overall. Similarly, Tasks 3 and 4 (Task Group B) and Tasks 5 and 6 (Task Group C) show different median percentages of reported strategies, indicating that tasks within each task group elicited slightly different percentages of individual strategies under each category. Appendix J reports the median of individual strategies across the six tasks in this study. In examining Appendix J for the greatest differences in medians of reported strategy use across tasks within task groups, we note that Task 1 has a higher median (*Mdn* = 9.09) for the strategy *making choices* than Task 2 (*Mdn* = .00). Similarly, Task 4 has a higher median for *linking to* *prior experiences/knowledge* (*Mdn* = 8.39) than Task 3 (*Mdn* = 4.45) has, whereas the median for *identifying the purpose of the task* for Task 5 (*Mdn* = 5.26) is higher than that for Task 6 (*Mdn* = .00; see Appendix J).

Overall, however, tasks within the same task group (i.e., requiring the same language skills) are more similar to each other in terms of percentage of reported strategy use than they are to tasks in other task groups (i.e., requiring different or additional language skills). The only exception is the large difference between Task 1 and Task 2 in terms of the approach category. It is worth noting here that a Friedman test comparing scores across tasks detected a significant difference between the test-takers' scores for Tasks 1 and 2. This was the only significant difference in scores across the six tasks (see Tables B1, B2, and B3 in Appendix B). Task 1 was the first task that the participants in this study encountered. It is possible that the differences in scores and reported use of approach strategies are due to the fact that Task 1 was the first task to be administered, rather than to any characteristics of the task itself.

Table 15

Overall Strategy Use by Task

Task		Approach	Communication	Cognitive	Metacognitive	Affective	Total[a]
Task 1	Median	23.30	28.57	8.39	30.38	.00	10.00
	Range	60.00	60.00	50.00	41.82	42.86	26.00
Task 2	Median	12.50	34.52	11.44	35.50	.00	10.00
	Range	28.57	87.50	42.86	83.33	15.38	27.00
Task 3	Median	7.74	20.00	34.58	26.49	4.36	18.50
	Range	25.00	44.44	39.29	62.25	14.29	25.00
Task 4	Median	5.88	24.62	30.77	29.80	1.72	21.50
	Range	17.24	42.31	35.71	44.64	15.00	29.00
Task 5	Median	8.01	26.14	26.14	32.46	.00	13.50
	Range	29.41	46.67	51.97	55.56	12.50	16.00
Task 6	Median	5.01	25.83	33.10	30.00	3.85	16.00
	Range	18.75	54.05	47.60	50.00	30.00	18.00

Note. Medians and ranges are based on percentage of reported strategy use.

[a] Figures in this column are based on raw frequencies, not percentages, of strategies reported.

Reported strategy use by task group and test-taker study and proficiency levels. Overall, there does not seem to be any significant interactions between test-taker study level and task group in terms of percentages of strategies reported, except for the approach category. As Table 16 shows, the undergraduate students have a slightly higher median for the approach category for Task Groups B and C than the graduate students have, but the medians of both student groups are equal for Task Group A. In terms of interaction between test-taker proficiency level and task group, Table 17 shows that there might be an interaction effect for the approach, communication, and affective categories. In other words, the differences in median percentages for these strategy categories vary depending on *both* task group and examinee proficiency level. For example, the median for the affective category for Task Group B is higher for the test-takers at the intermediate level than for their advanced counterparts, but for Task Group C, it is higher for the test-takers at the advanced level. However, these differences in medians across student and task groups were often not very large.

Table 16

Reported Strategy Use by Task Group and Test-Taker Study Level

Task group	Study level		Approach	Communication	Cognitive	Meta-cognitive	Affective
A	Undergraduate	Median	16.03	38.80	7.14	30.77	.00
	(*n* = 16)	Range	60.00	77.50	28.57	62.50	20.00
	Graduate	Median	16.23	15.48	12.92	38.75	.00
	(*n* = 14)	Range	42.86	45.45	50.00	69.05	42.86
B	Undergraduate	Median	8.01	24.19	31.01	25.95	.00
	(*n* = 16)	Range	25.00	43.75	41.25	41.35	15.00
	Graduate	Median	5.00	19.17	32.74	33.33	5.28
	(*n* = 14)	Range	20.00	40.00	44.05	62.25	14.29
C	Undergraduate	Median	9.55	28.57	26.14	29.71	1.92
	(*n* = 16)	Range	29.41	57.89	45.03	48.58	30.00
	Graduate	Median	6.46	22.90	33.33	35.57	.00
	(*n* = 14)	Range	22.22	46.67	56.25	66.67	15.38

Note. Medians and ranges are based on percentage of reported strategy use.

Overall, there were no large interaction effects between task group and test-takers' study and proficiency levels on the percentage of reported strategy use. In terms of test scores, Tables B4 and B5 in Appendix B show that the undergraduate students obtained significantly higher scores than the graduate students on Task 2. As Table 16 shows, Task 2 resulted in higher medians for the communication category and lower medians for the cognitive and metacognitive categories for the undergraduate group than for the graduate group.

Table 17

Reported Strategy Use by Task Group and Test-Taker Proficiency Level

Task group	Proficiency level		Approach	Communication	Cognitive	Meta-cognitive	Affective
A	Intermediate	Median	17.16	29.29	11.76	35.29	.00
	(*n* = 17)	Range	44.44	87.50	33.33	83.33	42.86
	Advanced	Median	15.04	36.16	7.14	34.17	.00
	(*n* = 13)	Range	60.00	71.43	50.00	72.73	15.79
B	Intermediate	Median	5.57	23.21	32.67	27.89	4.36
	(*n* = 17)	Range	21.43	50.00	43.33	62.25	15.00
	Advanced	Median	8.01	22.25	31.70	28.35	.00
	(*n* = 13)	Range	25.00	41.12	41.96	44.64	14.29
C	Intermediate	Median	7.42	23.30	28.99	30.38	.00
	(*n* = 17)	Range	29.41	50.00	48.30	55.56	15.38
	Advanced	Median	6.70	28.57	28.47	31.25	4.26
	(*n* = 13)	Range	20.83	57.89	48.21	55.56	30.00

Note. Medians and ranges are based on percentage of reported strategy use.

Research Question 4: Reported Strategy Use and Test Performance

To answer the fourth research question concerning the relationship between test-takers' reported strategic behaviors and their test scores, we conducted correlational analyses to examine whether there was a relationship between the test-takers' reported strategic behaviors and their SSTiBT test scores. The results in this section are presented from the broadest level of analysis (correlations between strategy categories and total test scores) to the narrowest level of analysis (correlations between reported individual strategies and task scores by task). In the analyses involving total test scores in this section (the "Overall reported strategy use and total test scores," "Strategy categories and total test scores," and "Individual strategies and total test scores" subsections below), we ran correlations between the aggregated (averaged) percentages of reported strategies across the six tasks for each student and the total test score, which is an average of the six task scores.[11] The correlations for task groups (the "Strategy categories and test scores by task group" and "Individual strategies and test scores by task group" subsections below) were run between the average reported strategy use and the task score averages across pairs of tasks within task groups. For individual task scores, the correlations were run between the scores for a given task and the strategies reported by the students while doing that particular task (the "Strategy categories and test scores by task" and "Individual strategies and test scores by task" subsections below). The results of the analyses (except for the "Strategy categories and test scores by task group" and "Individual strategies and test scores by task group" subsections below) are presented in Table 18.

Overall reported strategy use and total test scores. As shown in the second row of the last column of Table 18, there was no significant correlation between the total number of reported strategies and total test scores. Although not significant, the Spearman rho (r_s) coefficient (−.02) was negative. Since the test-takers were organized into proficiency groups based on their SSTiBT scores, this finding supports the results from our second research question, in which we found no significant differences in reported strategy use between intermediate and advanced test-takers (see Tables 8 and 9). This suggests that there was a great deal of variation in the number of reported strategies regardless of test-taker proficiency level.

Strategy categories and total test scores. In examining the correlations between the percentages of each of the five strategy categories and the total test scores, the results in the last column of Table 18 show no significant correlations for the approach, communication, cognitive, and metacognitive categories, but there is a significant negative correlation ($r_s = -.37$, $p < .05$) between the percentage of reported affective strategies and the total test score. Although the affective strategy category represented only a small percentage of the total strategies reported by test-takers (see Table 4), with an increased percentage of reported use, test scores tended to decrease.

Strategy categories and test scores by task group. When correlations were run on the average percentages of reported strategies and the average of the task scores within each task group, no significant correlations were found.[12] This means that the average of reported strategies in Task Group A did not correlate with the average score within that task group; similarly, the average of reported strategies in Task Groups B and C did not correlate significantly with their respective average task group scores.

Table 18

Correlations Between Percentage of Reported Strategy Use and Task and Test Scores

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total score
Total individual strategies[a]	−.11	−.06	−.22	.33	−.29	.08	−.02
Approach	−.10	−.22	.26	.07	.04	−.05	.15
Recalling the task type	.12	.23	.22	.10	−.05	N/A	.17
Recalling the question	.16	−.06	.00	.00	.19	.19	.19
Recalling the text	N/A	N/A	.19	.07	N/A	N/A	.09
Recalling the dialogue	N/A	N/A	.22	N/A	.15	N/A	.25
Recalling the lecture	N/A	N/A	N/A	−.06	N/A	−.02	.02
Generating choices	−.19	−.24	N/A	N/A	−.07	N/A	.03
Making choices	.14	.16	.07	N/A	−.07	N/A	.34
Developing reasons	−.22	−.34	−.08	.13	−.27	−.08	−.35
Communication	.02	.28	.06	.22	.01	.22	.30
Simplifying the message	−.07	−.15	−.02	.03	−.07	−.02	−.14
Avoiding	−.00	.28	.30	N/A	.03	.32	.21
Using Chinese	−.19	−.09	−.24	N/A	−.07	−.26	−.24
Paraphrasing	N/A	.01	.07	.14	.27	−.14	.06
Approximating	−.19	−.03	N/A	−.19	.14	−.04	.06
Linking to prior experiences/knowledge	−.04	.32	.08	.12	−.04	.19	.23
Borrowing	−.06	−.02	.03	.41*	.04	.32	.10
Reviewing notes	N/A	N/A	.25	.09	.25	.07	.24

(Table continues)

Table 18 (continued)

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total score
Referring to notes	−.08	.35	.09	−.06	−.03	−.26	−.03
Organizing thoughts	.12	.36*	.05	.20	.34	.06	.34
Guessing	N/A	−.24	.20	.12	−.32	−.13	−.20
Repeating	.28	−.26	−.32	−.23	−.14	−.04	.05
Rehearsing	.12	−.03	−.16	.19	−.07	.23	−.10
Reading ahead	−.06	.29	.15	.04	−.07	.06	.16
Restructuring	−.19	−.05	−.09	.13	−.07	.23	−.05
Slowing	−.18	−.05	−.04	.31	−.01	.04	.00
Thinking ahead	.05	.28	.07	N/A	.28	N/A	.23
Elaborating to fill time	.17	.19	−.01	−.06	−.19	N/A	.08
Elaborating to clarify meaning	.14	.00	−.09	.13	−.20	.21	.28
Cognitive	.05	−.09	.07	−.33	.23	−.04	−.05
Attending	.14	−.19	−.34	−.26	−.29	−.10	−.42*
Anticipating the content	−.19	−.03	.06	−.01	−.17	.01	−.13
Anticipating the structure	N/A	.40*	.38*	−.33	.21	−.02	.24
Using imagery	−.06	.16	.35	.08	−.07	.29	.36
Using mechanical means to organize	−.11	.01	−.05	.04	.38*	.09	.14
Memorizing	N/A	N/A	.20	.07	.20	.07	.25
Summarizing	.15	−.02	−.32	.15	.29	−.01	.08
Translating	.12	−.24	−.24	N/A	N/A	N/A	−.16
Inferencing	N/A	N/A	.24	−.06	.30	−.07	.17
Processing inductively	N/A	N/A	N/A	.27	N/A	N/A	.28

(Table continues)

Table 18 (continued)

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total score
Metacognitive	.24	−.19	−.10	−.03	−.28	−.20	−.26
Setting goals	−.12	−.23	−.29	.45*	−.08	−.20	−.43*
Identifying the purpose of the task	−.23	.34	−.43*	.06	.22	.34	.19
Planning	.19	−.22	−.03	−.09	−.25	−.05	−.20
Monitoring	−.10	.05	.03	.11	−.17	−.04	.05
Self-correcting	.32	.15	.23	−.18	.01	.26	.39*
Evaluating previous performance	−.27	.04	.18	.14	−.41*	.19	.08
Evaluating the content of what was read/heard	−.03	.01	−.21	.01	−.05	−.37*	−.35
Evaluating performance	.29	−.23	.22	−.28	−.12	−.04	−.24
Evaluating language production	.22	−.03	.08	.20	.07	.07	.21
Affective	−.37*	−.14	−.33	.10	.23	.04	−.37*
Lowering anxiety	−.16	N/A	.02	.11	.10	.38*	.02
Encouraging self	−.44*	N/A	−.07	.10	.18	−.10	−.22
Justifying performance	−.38*	−.14	−.38*	−.15	.14	−.21	−.48**

Note. Spearman *rho*, $N = 30$

[a] Grand total number of strategies reported based on raw frequencies, not percentages.

* Correlation is significant at $p < .05$ (2-tailed). ** Correlation significant at $p < .01$ (2-tailed).

Strategy categories and test scores by task. In Table 18, columns 2 to 7 (under the headings Task 1 through Task 6) show the results of examining the strength and direction of the relationship between the strategy categories within each task and the test score for that task. All of the correlations were nonsignificant except under Task 1; the affective strategy category negatively correlated with the task score ($r_s = -.37$, $p < .05$).[13] Suggestive of a practice effect, since it was the first task the participants encountered, the average test score for Task 1 (see Table B1 in Appendix B) had the lowest mean score ($M = 2.57$, $SD = .78$) of all the task scores. It was also the only task in which a participant received a score of 0 (see Table B1). It is important to note that this same participant reported 42.86% of the total affective strategies in Task 1, and explains why a significant negative correlation was found.

Individual strategies and total test scores. In examining the last column of Table 18 for significant correlations between the percentages of reported use of individual strategies and the total test score, we found four significant correlations. The cognitive strategy of *attending* negatively correlated with the total test score ($r_s = -.42$, $p < .05$) as did the metacognitive strategy of *setting goals* ($r_s = -.43$, $p < .05$). Both of these individual strategies direct attention toward aspects of the task itself (see Appendix F) and away from the online language processing for the task, so these correlations suggest that as reported use of those strategies increased, the test-takers' performance, as measured by their test scores, decreased. *Self-correcting*, a metacognitive strategy, positively correlated with test score ($r_s = .39$, $p < .05$), which suggests that test-takers who reported an awareness of their errors, at least in the stimulated recall,[14] tended to have higher test scores. Finally, the percentage of the reported affective strategy of *justifying performance* correlated negatively with the total test score ($r_s = -.48$, $p < .01$). As the example in Appendix F suggests, test-takers may have justified their performance only in the stimulated recall rather than during the test itself, but overall the test scores tended to decrease as their reported use of that individual strategy increased.

Individual strategies and test scores by task group. When we ran correlations between the average of reported individual strategies within each task group (i.e., Task Groups A, B, and C) and the average test scores with each task group, there were no significant correlations in Task Group A. In Task Group B, the communication strategy of *repeating* negatively correlated with test scores ($r_s = -.40$, $p < .05$), as did the affective strategy of *justifying performance* ($r_s = -.42$,

$p < .05$). In Task Group C, the cognitive strategy of *attending* negatively correlated with test scores ($r_s = -.36$, $p < .05$).

Individual strategies and test scores by task. In this subsection, we report on the results in Table 18, in which there were significant correlations between individual strategies and the test score for each respective task. A total of 13 significant correlations were found, and they are listed in Table 19 along with the strategy category and the direction of the relationship between the reported individual strategy use and task score.

Table 19

Significant Correlations Between Reported Individual Strategies and Task Scores by Task

Task	Strategy category	Individual strategy	Direction of correlation
Task 1	Affective	Encouraging self	Negative
	Affective	Justifying performance	Negative
Task 2	Communication	Organizing thoughts	Positive
	Cognitive	Anticipating the structure	Positive
Task 3	Cognitive	Anticipating the structure	Positive
	Metacognitive	Identifying purpose of the task	Negative
	Affective	Justifying performance	Negative
Task 4	Communication	Borrowing	Positive
	Metacognitive	Setting goals	Positive
Task 5	Cognitive	Using mechanical means	Positive
	Metacognitive	Evaluating previous performance	Negative
Task 6	Metacognitive	Evaluating content of what heard/said	Negative
	Affective	Lowering anxiety	Positive

As shown in Table 19, of the 13 significant correlations, the two in the communication strategy category were positive, as were the three in the cognitive strategy category. In the metacognitive strategy category, one correlation was positive (*setting goals*) while three were negative. In the affective strategy category, one correlation was positive (*lowering anxiety*) while the remaining three were negative. It is worth noting that two of the three significant negative correlations in the affective category were in Task 1, again suggestive of the test-takers having more of an affective response at the beginning of the SSTiBT and also possibly because of the one participant who scored 0 on Task 1 and reported a high percentage of the total affective strategies for that task (42.86%).

Discussion

Key Findings and Implications

Test-takers reported using a wide range of strategies (49 in all) when completing the test. These strategies are applicable to both learning and testing contexts. All participants reported using at least five strategies for each task.

In general, tasks within each task group are similar to each other with respect to reported strategy use. This supports grouping the tasks by language skills. Currently, the TOEFL speaking tasks are grouped by the sub-domains (e.g., everyday familiar topics, campus-life situations) that are expected to affect some components of students' speaking performance in important ways (e.g., vocabulary usage, fluency; X. Xi, personal communication, February 17, 2008).

The integrated task groups (Task Groups B and C) were more similar to each other than they were to Task Group A. First, Task Groups B and C elicited a wider variety of reported strategies than Task Group A elicited. Second, there were more significant differences in terms of strategy categories between Task Group A on the one hand and Task Groups B and C on the other hand, than between Task Groups B and C. Including integrated tasks thus broadens the scope of strategies called upon in the SSTiBT speaking tasks. The integrated task group involving three language skills (Task Group B) elicited greater reported strategy use than the integrated task group that involved two language skills (Task Group C), and both Task Group B and Task Group C elicited more reported strategy use than did independent Task Group A. This suggests that the more language skills involved in a task, the higher the frequency of reported strategy use. The inclusion of integrated tasks is intended to simulate typical communication in an actual academic setting. Our findings indicate that integrated tasks elicit strategic behaviors that are different from those used in independent tasks, and thus support the use of both types of tasks in the SSTiBT.

The findings that all test-takers reported using a variety of strategies and that strategy use varied significantly across task groups imply that strategy use is integral to performing SSTiBT tasks, and therefore should be considered as part of the construct of communicative performance. We propose three versions of this argument. First, a weak version, supported by the findings of the current study, indicates that test-takers do engage in a variety of strategic behaviors when performing the SSTiBT tasks. Given that many of these strategies are in some ways obvious given the task types (e.g., *generating ideas, planning, attending*), one can conclude that these *are* part of the construct in that the task designers must have had these strategies in mind when they designed the SSTiBT tasks. This empirical evidence about the actual strategies that test-takers reported employing can be used to substantiate claims about the validity of inferences based on SSTiBT scores.

Second, the finding that the more complex the task is, the more strategies the test-takers report using supports a slightly stronger version of the argument. According to this version, strategic behavior mediates the relationship between task (complexity) and performance (scores). In other words, strategies compensate for the complexity or difficulty of the task. As tasks become more complex or difficult, test-takers use more strategies to achieve the same level of performance. It is also possible that some features of the more complex tasks may require the use of (additional) specific types of strategy (e.g., *using mechanical means to organize*, *anticipating the structure, attending*), leading to the use of more strategies overall.

Finally, a strong version of the argument is that strategy use should be part of the scoring criteria and claims based on SSTiBT scores. However, to be included as part of the scoring criteria, two conditions need to be met: (a) the use of strategies has to be observable in the product (i.e., raters must be able to identify it), and (b) the amount or type of strategy use has to differ across score levels. Since the focus of the current study was on test-takers' stimulated recalls of their performance, rather than the spoken performances themselves, we are unable to address this strong version of the argument. Additionally, we are aware that some strategies are inherently unobservable (e.g., *rehearsing*, *using imagery*) and cannot be included as part of a scoring rubric.

The undergraduate group reported using significantly more communication strategies than the graduate group reported, whereas the graduates reported using significantly more cognitive and affective strategies than the undergraduates reported. In our sample, the undergraduates had spent more time than the graduates in an English-speaking country, and may therefore more readily have used communication strategies. We wonder if the difference in length of residence is a typical difference between undergraduate and graduate populations in

North America. If it is, what should be the implications for test reporting and admissions practices?

In this study we examined the relationship between the reported use of strategic behaviors and test and task scores. However, based on the finding that there is no relationship ($r_s = -.02$) between the total number of reported strategic behaviors and total test score on the SSTiBT, we would argue that the reported use of strategic behaviors is indirectly related to performance. Our consideration of the total data set convinces us that strategic behaviors mediate the relationship between task/test and spoken performance. However, in our study it was the spoken performance that was rated, not the strategic behaviors that the test-takers reported using to perform the task. As a result, many of the correlations between reported strategic behaviors and scores in this study were weak or mixed. When faced with tasks that were more complex or difficult, test-takers tended to report using more strategies, and this increased use of strategies may have led to their obtaining the same scores on tasks that differ in terms of difficulty.

In addition, the finding that the same reported strategic behavior may be effective with one task but not with another makes it challenging to link reported strategy use to test performance, because a desirable strategy in one instance may negatively impact performance in another context. This may be because the resources allocated to the execution of any particular strategy may impact other aspects of speech production that need attentional resources at the same time. In turn, this tendency may be related to the difficulty or complexity of the task, as well as to test-taker second/foreign language learning and test-taking histories. In other words, the effectiveness of a particular strategy may be task, context, and individual dependent.

While the total number of reported strategic behaviors did not correlate significantly with total test score ($r_s = -.02$), there is one significant correlation at the level of strategy category. That significant correlation is negative and is between reported affective strategy use and total test score ($r_s = -.37$), due mostly to students' justifying their performance. This means that the students with low proficiency (i.e., those obtaining lower test scores) tended more often to try to explain their poor performance.

Of the 13 individual strategies for which there were significant correlations with task scores, the cognitive and communication strategies correlated positively, whereas the metacognitive and affective strategies tended to correlate negatively. The learning strategy literature suggests a positive relationship between performance and three of these strategy categories (cognitive, communication, and metacognitive), so the negative correlations between some of the individual metacognitive strategies and task scores found in the current study are somewhat surprising. The use of metacognitive strategies would seem essential for the successful completion of a speaking task. On reflection, however, we suggest that speaking (perhaps like listening) is a skill that has special requirements (relative to writing and reading) because of the immediate, online nature of a speaking performance. Making use of metacognitive strategies may simply use up too much of the attentional resources required to produce a speaking performance that is fluent, linguistically satisfactory (use of correct morphology, syntax, and vocabulary), and contains acceptable content.[15] In other words, because of the unique features of a speaking task, the use of metacognitive strategies may negatively affect performance because it consumes the limited mental resources available, but needed, to successfully carry out the task at hand. Given the findings of this study, we would like to suggest that the use of some metacognitive strategies (e.g., *setting goals*) but not others (e.g., *self-correcting*) may interfere with successful performance on a timed speaking test.

The results from our study have implications for strategy training. One of the goals of strategies-based instruction is to increase students' awareness and repertoire of strategies (Brown, 2007) so that they can determine the right combination of strategies that works well for them on a given task. For training test-takers, the use of stimulated recall could serve the purpose of raising students' awareness of the strategies they use in a speaking task and elicit from them other strategies they could add to their repertoires.

Limitations

Although all test-takers were asked to take the test seriously, as if their admission to a university depended on it, the fact that the test did not take place under real examination conditions and had no real consequences for the students might have produced different test results and/or elicited different strategic behaviors than those that would have occurred during an actual administration of the SSTiBT.

Stimulated recalls may represent only a partial list of the possible strategies test-takers could have tapped while performing the SSTiBT, or that they tapped but did not report (Cohen, 1998; Gass & Mackey, 2000; Pressley & Afflerbach, 1995; Russo, Johnson, & Stephens, 1989). In addition to the argument that some strategies are automatic and, thus, not conscious or cannot be verbalized, it is possible that some strategic behaviors that are more global (e.g., *identifying* *the purpose of the task*, 3%) are reported less frequently than other, more localized behaviors (e.g., *using mechanical means to organize*, 12%) because, although they can affect the process and outcome of performance significantly, they need to be employed only a few times (e.g., at the beginning of the process). In addition, participants can be selective in terms of what they report, given the large number of behaviors they may employ at a given time and/or their awareness of an audience for their verbal reports (Cohen, 1998; Pressley & Afflerbach, 1995; Russo et al., 1989).

The research version of the SSTiBT allowed us to pause after each task to facilitate stimulated recalls. Some participants reported that having a chance to reflect on what they had done after each task affected their use of strategic behaviors and test performance in the subsequent tasks (see Appendix K for some examples). Although we list this as a limitation, we wish to make it clear that these examples also illustrate the *value of stimulated recall* in a teaching and learning context, and (we would argue) its value in helping students understand which strategic behaviors might help them in a test-taking context according to the task type and language skill(s) involved, as compared to a naturalistic context, where the language skills that might be needed will more likely be mediated by a broader range of strategic behaviors.

Recategorizing our sample of participants according to the research version of the SSTiBT (as compared to the initial categorization based on our pretest) yielded a small subsample of four (graduate/advanced). This made achieving statistical significance difficult. The range of proficiency we targeted appears to have been too limited, perhaps leading to findings of no differences.

The inclusion of only two tasks per task group (imposed by the structure of SSTiBT) limited our ability to generalize with confidence. In addition, the test tasks were administered in the same order to all students (imposed by the structure of SSTiBT), so that counterbalancing task order across students was not possible.

As is often the case for other studies in the field, we considered only frequencies (percentages) of reported strategic behaviors. We did not consider sequencing (e.g., metacognitive strategies may tend to be used initially; cognitive and communication strategies may tend to be used during performance), quality (i.e., what works for an individual test-taker on a given task), and the global/local nature inherent in each individual strategy (i.e., the importance of the strategy to the task as a whole versus its local application).

The taxonomies of strategies are atheoretical. When there is no theory to inform coding decisions within and across studies, they are rather arbitrary. We consider this a general weakness of all studies investigating strategy use.

Future Research

The current data set could be explored further in at least two ways: (a) through case studies, examine the relationships among test-takers' strategic behaviors while taking the SSTiBT (e.g., Do they self-correct?), their actual performance (e.g., their self-correction), the quality and sequencing of their strategy use, their test scores, and their stimulated recalls (e.g., Did they report that they self-corrected?); and (b) through within-group comparisons (e.g., comparing test-takers who achieved high scores with each other), examine the differences in patterns of reported strategy use. The goal of these group comparisons would be to explore the variability among test-takers who obtain similar test scores.

This study could be replicated minimizing the limitations we have identified (e.g., including more tasks per task group, counterbalancing the order of task presentation) and including different samples of test-takers (e.g., participants with differing language backgrounds and having a wider range of proficiency in the L2).

More research is needed to assess whether and how strategic competence should be incorporated into the scoring rubric. Such research could explore whether and what strategic behaviors are observable (i.e., heard by the rater) in spoken performance, whether they vary across proficiency levels, whether they can and should be considered separately from other aspects of performance (e.g., grammar, discourse), whether and how they can be identified and evaluated accurately and consistently, and if and how test users can interpret and use such information appropriately.

To obtain a clearer picture of the indirect relationships between strategic behaviors and scores, future studies/analyses need to examine how strategic behaviors affect spoken performance (e.g., linguistic and discourse features) and how the spoken performance affects test and task scores through a multilayered analysis, such as multilevel modeling (Raudenbush & Bryk, 2002).

Conclusions

This study constitutes a response to the TOEFL program's research agenda concerning the need to understand the processes and knowledge that test-takers use when responding to the speaking tasks in the TOEFL iBT assessment. To do this, we asked test-takers to report the strategies they used while completing each task of a version of the TOEFL iBT Speaking test. They were asked to report the strategies they used as they viewed a video of themselves completing each of six tasks (stimulated recall). The stimulated recalls were conducted immediately after each task. To our knowledge, this study is the first to collect data regarding the strategies test-takers report using while taking a speaking test. Furthermore, it is the first to use stimulated recall immediately following the completion of each test item (or task, in the case of SSTiBT).

The perspective taken in this research is that strategic behaviors are the goal-directed actions taken by test-takers to regulate their cognitive processes in preparing to respond to a task, in responding to the task, or in reflecting on how they responded to the task. The actions taken by an individual test-taker reflect his or her background characteristics (in the case of this study, their proficiency level [intermediate and advanced] and study level [graduate and undergraduate]), their goals (to do well on the test and get some practice on how to do it well) in interacting with the tasks (six SSTiBT tasks), and the context (university research project) in which the testing takes place.

Given this complexity, we remain unsurprised but disappointed that we found few significant correlations between the strategies test-takers reported using and their test scores. We do, however, remain convinced that a relationship exists, but that it is much more complex than a simple linear relationship. Our current view—one supported by the stimulated protocol data—is that strategies are mediating tools; that is to say, strategies mediate between the test-taker and his or her performance (as reflected in the score he or she obtains). The test-takers' reports make it strikingly clear that the use of strategies is an integral aspect of taking the SSTiBT and, in that sense, should be considered as part of the construct of communicative performance. Furthermore, it appears that the more complex a task is from the perspective of the demands made on test-takers' language skills (i.e., integrated tasks), the wider the variety of the reported use of strategies. This may partially account for the fact that there were no significant test score differences among the task types (groups).[16] In other words, as the tasks became more complex, the test-takers compensated by using a greater variety of strategies. Our results also show that, in the context of a speaking test, the reported use of communication, cognitive, and metacognitive strategies are negatively correlated. We argue that this is due to the online nature of speaking, particularly under test conditions where limited mental resources must be used to undertake a task with unique, online characteristics.

As we have noted, this study is the first to examine reported strategy use in a speaking test context. A next step is to examine actual strategy use in relation to both test scores and reported strategy use. However, in order to move forward, we need to revisit the fact that frequency is the basis of all analyses in strategic behavior studies. We believe that going beyond simple frequency counts will lead to a reconceptualization of underlying constructs. That is to say, we need to deconstruct our construct of each individual strategy: what, by whom, why, where, when, and how. As part of that process, we will better understand the use of strategies as a mediating tool between task characteristics and performance in a particular context. By *context*, we mean not only the setting but also test-takers' characteristics, including language learning and test-taking histories.

We believe the strategy research field needs shaking up. Having worked with the set of strategies found in the literature, we see conceptual and empirical overlap, with little attention paid to the specifics of the context of use. Our view is that the lists of strategies need to be deconstructed, culled, and reformulated into a theoretically based framework that takes account of the history of the strategy user, the tasks to which the strategies are being applied, and the broader context of use. Microgenetic analysis of change over time with respect to task and context is key to this understanding. Needless to say, we see the area as ripe for further research and theorizing.

References

Anderson, N. (2005). L2 learning strategies. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning.* (pp. 757–771). Mahwah, NJ: Erlbaum.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford, England: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476.

Bachman L. F., & Cohen, A. D. (Eds.). (1998). *Interfaces between second language acquisition and language testing research.* Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford, England: Oxford University Press.

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes*. Unpublished doctoral thesis, University of Toronto, ON, Canada.

Bialystok, E. (1981). The role of conscious strategies in second language proficiency. *Modern Language Journal*, *65*, 24–35.

Brown, H. D. (2007). *Teaching by principles: An interactive approach to language pedagogy*. White Plains, NY: Pearson Education.

Bruen, J. (2001). Strategies for success: Profiling the effective learner of German. *Foreign Language Annals, 34*(3), 216–225.

Butler, F. A., Eignor, D., Jones, S., McNamara, T., & Suomi, B. K. (2000). *TOEFL 2000 speaking framework: A working paper* (TOEFL Monograph Series Rep. No. 20) Princeton, NJ: ETS.

Bygate, M., Skehan, P., & Swain, M. (Eds.). *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow, England: Longman.

Canale, M. (1983). On some dimensions of language proficiency. In J. W. Oller, Jr. (Ed.), *Issues in language testing research* (pp. 333–342). Rowley, MA: Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1–47.

Chalhoub-Deville, M. (2001). Task-based assessments: Characteristics and validity evidence. In M. Bygate, P. Skehan, & M. Swain (Eds.). *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 210–228). Harlow, England: Longman.

Chalhoub-Deville, M. (2003). Second language interaction: Current perspectives and future trends. *Language Testing, 20,* 369–383.

Chamot, A. U. (1993). Student responses to learning strategy instruction in the foreign language classroom. *Foreign Language Annals*, *26*, 308–321.

Chamot, A. U., Kϋpper, L., & Impink-Hernandez, M. V. (1988). *A study of learning strategies in foreign language instruction: Findings of the longitudinal study*. McLean, VA: Interstate Research Associates.

Chapelle, C., & Douglas, D. (1993, March). *Interpreting L2 performance data.* Paper presented at the Second Language Research Colloquium, Pittsburgh, PA.

Chapelle, C., Grabe, W., & Berns. M. (1997). *Communicative language proficiency: Definition and implications for TOEFL 2000* (TOEFL Monograph Series Rep. No. 10). Princeton, NJ: ETS.

Cohen, A. D. (1984). On taking language tests: What the students report. *Language Testing*, *1*(1), 70–81.

Cohen. A. D. (1994). *Assessing language ability in the classroom* (2nd ed.). Boston: Newbury House/Heinle & Heinle.

Cohen, A. D. (1998). *Strategies in learning and using a second language*. London: Longman.

Cohen, A. D. (2002). Preparing teachers for styles- and strategies-based instruction. In V. Crew, C. Davison, & M. Barley (Eds.), *Reflection language in education* (pp. 49–69). Hong Kong: The Hong Kong Institute of Education.

Cohen, A. D. (2007). The coming of age for research on test-taking strategies. In J. Fox, M. Wesche, D. Bayliss, L. Cheng, C. E. Turner, & C. Doe (Eds.), *Language testing reconsidered* (pp. 89–111). Ottawa, Canada: University of Ottawa Press.

Cohen, A. D., & Aphek, E. (1981). Easifying second language learning. *Studies in Second Language Acquisition, 3*(2), 221–235.

Cohen, A. D., & Olshtain, E. (1993). The production of speech acts by EFL learners. *TESOL Quarterly, 27*, 33–58.

Cohen, A. D., Weaver, S., & Li, T.-Y. (1996). *The impact of strategies-based instruction on speaking a foreign language* (CARLA Working Papers Series No. 4). Minneapolis, MN: Center for Advanced Research on Language Acquisition.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York: Academic Press.

Dadour, S. (1995). *The effectiveness of selected learning strategies in developing oral communication of English department students in faculties of education*. Unpublished doctoral thesis, Mansoura University, Damietta, Egypt.

Dadour, S., & Robbins, J. (1996). University-level studies using strategy instruction to improve speaking ability in Egypt and Japan. In R. L. Oxford (Ed.), *Language learning motivation: Pathways to the new century* (pp. 157–166). Mānoa, HI: University of Hawaii Press.

Dörnyei, Z. (1995). On the teachability of communication strategies. *TESOL Quarterly*, *29*, 55–85.

Douglas, D. (1997). *Testing speaking ability in academic contexts: Theoretical considerations* (TOEFL Monograph Series Rep. No. 8.) Princeton, NJ: ETS.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge, England: Cambridge University Press.

Dreyer, C., & Oxford, R. (1996). Learning strategies and other predictors of ESL proficiency among Afrikaans speakers in South Africa. In R. L. Oxford (Ed.), *Language learning strategies around the world: Cross-cultural perspectives* (pp. 61–74). Mānoa, HI: University of Hawaii Press.

Ellis, R. (1994). *The study of second language acquisition.* Oxford, England: Oxford University Press.

ETS. (2004). *iBT/Next Generation TOEFL Test: Independent speaking rubrics.* Retrieved May 20, 2009, from http://www.ets.org/Media/Tests/TOEFL/pdf/Speaking_Rubrics.pdf

Færch, C., & Kasper, G. (1980). Processes and strategies in foreign language learning and communication. *Interlanguage Studies Bulletin*, *5*, 47–118.

Færch, C., & Kasper, G. (Eds.). (1983). *Strategies in interlanguage communication*. New York: Longman.

Feyten, C. M., Flaitz, J., & LaRocca, M. (1999). Consciousness raising and strategy use. *Applied Language Learning*, *10*(1 & 2), 15–38.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). Thousand Oaks, CA: Sage.

Flaitz, J., & Feyten, C. (1996). A two-phase study involving consciousness raising and strategy use for foreign language learners. In R. L. Oxford (Ed.), *Language learning strategies around the world: Cross-cultural perspectives* (pp. 211–225). Mānoa, HI: University of Hawaii Press.

Fulcher, G. (2003). *Testing second language speaking.* London: Longman/Pearson Education.

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum.

Green, N. M., & Oxford, R. (1995). A closer look at learning strategies, L2 proficiency, and gender. *TESOL Quarterly, 29*, 261–297.

Hamp-Lyons, L., & Lynch, B. K. (1998). Perspectives on validity: A historical analysis of language testing conference abstracts. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 253–276). Mahwah, NJ: Lawrence Erlbaum.

Harley, B., Allen, P., Cummins, J., & Swain, M. (1990). *The development of second language proficiency*. Cambridge, England: Cambridge University Press.

Homburg, T. J., & Spaan, M. C. (1981). ESL reading proficiency assessment: Testing strategies. In M. Hines & W. Rutherford (Eds.), *On TESOL '81* (pp. 25–33). Washington, DC: TESOL.

Huang, L.-S. (2004). Focus on the learner: Language learning strategies for fostering self-regulated learning. *Contact* [Special Research Symposium Issue], *30*, 37–54.

Jamieson, J., Jones, S., Kirsch, I., Mosenthal, P., & Taylor, C. (2000). *TOEFL 2000 framework: A working paper* (TOEFL Monograph Series Rep. No. 16). Princeton, NJ: ETS.

Kæsper, G., & Kellerman, E. (1997). (Eds.). *Communication strategies: Psycholinguistic and sociolinguistic perspectives*. London: Longman.

Kunnan, A. J. (1995). Theoretical models and empirical studies. In M. Milanovic (Ed.), *Test taker characteristics and test performance* (Studies in Language Testing 2). Cambridge, England: University of Cambridge Local Examinations Syndicate.

Kunnan, A. J. (Ed.). (1998). *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah, NJ: Lawrence Erlbaum.

LoCastro, V. (1994). Learning strategies and learning environments. *TESOL Quarterly*, *28*, 409–414.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Milanovic, M., Saville, N., Pollitt, A., & Cook, A. (1996). Developing rating scales for CASE: Theoretical concerns and analyses. In A. Cumming & R. Berwick (Eds.), *Validation in language testing* (pp. 15–38). Clevedon, England: Multilingual Matters.

Naiman, N., Fröhlich M., Stern H. H., & Todesco, A. (1978). *The good language learner. Research in Education Series 7.* Toronto, ON: Ontario Institute for Studies in Education.

Neisser, U. (1976). *Cognition and reality: Principles and implications of cognitive psychology*. San Francisco, CA: Freeman.

Norris, J. M., Brown, J. D., Hudson, T., & Yoshioka, J. (1998). *Designing second language performance assessments* (SLTCC Technical Rep. 8). Mānoa, HI: University of Hawaii Press.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge, England: Cambridge University Press.

Nunan, D. (1996). *The effect of strategy training on student motivation, strategy knowledge, perceived utility and deployment*. Unpublished manuscript, University of Hong Kong.

O'Malley, M. J., & Chamot, A. U. (1990). *Learning strategies in second language acquisition.* Cambridge, England: Cambridge University Press.

Oxford, R. L. (1990). *Language learning strategies*. New York: Newbury House.

Oxford, R. L. (Ed.). (1996). *Language learning strategies around the world: Cross-cultural perspectives* (SLT&CC Technical Rep. No. 13). Mānoa, HI: University of Hawaii Press.

Oxford, R. L. (2001). Language learning styles and strategies. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (pp. 359–366). Boston: Heinle & Heinle.

Oxford, R. L., & Burry-Stock, J. (1995). Assessing the use of language learning strategies worldwide with the ESL/EFL version of the Strategy Inventory for Language Learning. *System, 23*, 1–23.

Oxford, R. L., & Ehrman, M. E. (1995). Adults' language learning strategies in an intensive foreign language program in the United States. *System, 23*, 38–45.

Palmer, A. S., Groot, P. J. M., & Trosper, G. A. (Eds.). (1981). *The construct validation of tests of communicative competence.* Washington, DC: TESOL.

Paribakht, T. (1985). Strategic competence and language proficiency. *Applied Linguistics, 6,* 132–146.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test. *Language Testing*, *20*, 26–56.

Politzer, R., & McGroarty, M. (1985). An exploratory study of learning behaviors and their relationship to gains in linguistic and communicative competence. *TESOL Quarterly*, *19*, 103–124.

Poulisse, N. (1987). Problems and solutions in the classification of compensatory strategies. *Second Language Research*, *3*, 141–153.

Poulisse, N. (1990). *The use of compensatory strategies by Dutch learners of English*. Dordrecht, Holland: Foris.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading: The nature of constructively responsive reading*. Hillsdale, NJ: Lawrence Erlbaum.

Purpura, J. E. (1997). An analysis of the relationships between test-takers' cognitive and metacognitive strategy use and second language test performance. *Language Learning*, *47*, 289–325.

Purpura, J. E. (1998). Investigating the effects of strategy use and second language test performance with high- and low-ability test-takers: A structural equation modeling approach. *Language Testing, 15*, 333–379.

Purpura, J. E. (1999). *Learner strategy use and performance and language tests: A structural equation modeling approach.* Cambridge, England: University of Cambridge Local Examinations Syndicate and Cambridge University Press.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Rosenfeld, M., Oltman, P. K., & Sheppard, K. (2004). *Investigating the validity of TOEFL: A feasibility study using content and criterion-related strategies* (TOEFL Research Rep. No. 71). Princeton, NJ: ETS.

Rubin, J. (1975). What the "good language learner" can teach us. *TESOL Quarterly*, *9*, 41–51.

Rubin, J. (1987). Learner strategies: Theoretical assumptions, research history, and typology. In A. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 15–29). Englewood Cliffs, NJ: Prentice-Hall International.

Russo, J. E., Johnson, E. J., & Stephens, D. L. (1989). The validity of verbal protocols. *Memory and Cognition, 17*, 759–769.

Selinger, H. W. (1983). The language learner as linguist: Of metaphors and realities. *Applied Linguistics*, *4*, 179–191.

Skehan, P. (1991). Individual differences in second-language learning. *Studies in Second Language Acquisition*, *13*, 275–298.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics, 17*, 38–62.

Song, X. (2005). Language learner strategy use and English proficiency on the Michigan English Language Assessment Battery. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *3*, 1–26.

Swain, M. (1985). Large-scale communicative language testing: A case study. In Y. P. Lee, A. C. Y. Fok, R. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35-46). Oxford, England: Pergamon Press.

Swain, M. (2001). Examining dialogue: Another approach to content specification and to validating inferences drawn from test scores. *Language Testing, 18*, 275–306.

Wenden, A., & Rubin, J. (1987). *Learner strategies in language learning.* Englewood Cliffs, NJ: Prentice-Hall International.

Wesche, M. B. (1981). Communicative testing in a second language. *Canadian Modern Language Review, 37*, 551–571.

Wesche, M. B. (1987). Second language performance testing: The Ontario test of ESL as an example. *Language Testing*, *4*, 28–47.

Wharton, G. (2000). Language learning strategy use of bilingual foreign language learners in Singapore. *Language Learning*, *50*, 203–243.

Widdowson, H. (1983). *Learning purpose and language use.* Oxford, England: Oxford University Press.

Yoshida-Morise, Y. (1998). The use of communication strategies in LPIs. In R. Young & W. He (Eds.), *Talking and testing: Discourse approaches to the assessment of oral proficiency* (pp. 205–238). Amsterdam: John Benjamins.

Yule, G., & Tarone, E. (1997). Investigating L2 reference: Pros and cons. In G. Kasper & E. Kellerman (Eds.), *Advances in communication strategy research* (pp. 17–30). New York: Longman.

Notes

[1] We are aware that there is little consensus regarding how to define *tasks* (e.g., Bachman, 2002; Bachman & Palmer, 1996; Bygate, Skehan, & Swain, 2001; Norris, Brown, Hudson, & Yoshioka, 1998; Nunan, 1989; Skehan, 1996). Here, *tasks* refers specifically to the six speaking tasks in the SSTiBT (see Table 3).

[2] The term *cognitive processes*, which is taken from cognitive psychology, refers to all processes by which sensory input is transformed, reduced, elaborated, stored, recovered, and used (Neisser, 1976).

[3] Since the SSTiBT does not require dialogical exchanges between the tester and the test-taker, *social strategies,* which entail interacting with others to improve language learning or language use (e.g., asking for correction, cooperating, and empathizing with others; O'Malley & Chamot, 1990; Oxford, 1990), were not included in the list of strategies.

[4] See Table 3.

[5] In this study, *Chinese* refers to modern standard Chinese (commonly known as Mandarin or Putonghua), which is the official language of government and education in the People's Republic of China and Taiwan.

[6] All translations in this report were done by the second author, Huang, who is a professional translator certified by the National Accreditation Authority for Translators and Interpreters (NAATI) and the Canadian federal government's Translation Bureau.

[7] Tasks 1, 3, and 6 were selected to establish inter-coder agreement because they represent the main task groups (see Table 3).

[8] Refer to Appendix F for the definitions of the five strategy categories.

[9] The first letter stands for graduate (G) or undergraduate (U) and subsequent letters are the initials of the participant.

[10] *Recalling the text* was reported surprisingly infrequently. This may be related to how the test-takers perceived the importance of the reading segment in Tasks 3 and 4. Several participants commented during the stimulated recall sessions and exit interviews that they learned during the familiarization session that comprehending the reading segment did not play a role in

facilitating or hindering their speaking performance; the content that had the most direct relevance to their speaking was in the listening portion of the tasks. Thus, the participants found little need to recall the text that they had read for Tasks 3 and 4 during the preparing-to-speak stage.

[11] As explained in the Data Analysis section, coded data (i.e., strategic behaviors) were converted to percentages before any statistical analyses were conducted. Percentages of reported individual strategies were computed for each test-taker for each task as follows: counts of coded individual strategies (e.g., *setting goals*) were summed for each test-taker for each task and then divided by the total number of instances of reported individual strategies for that particular test-taker for that particular task to obtain a percentage of times that code occurred.

[12] There is no table for this section because none of the correlations were significant.

[13] Note that significant correlations were found (a) between total test score and percentage of reported affective strategies and (b) between Task 1 scores and percentage of reported affective strategies for Task 1, but not between (c) Task Group A scores and percentage of reported affective strategies. This may be due to the different ways of aggregating the scores and percentages of reported strategies as described in the Data Analysis section.

[14] An examination of the test transcripts would reveal how often the test-takers actually self-corrected during the SSTiBT vs. whether they reported thinking about self-correcting during the stimulated recall.

[15] A reviewer suggested that for the speakers with more advanced levels of proficiency, using metacognitive strategies may be more automatic and subconscious, so these proficiencies were not reported as frequently as they were for speakers of lower proficiency. However, we think it is just as likely that students of higher proficiency who have a greater repertoire of strategies simply may not be able to verbalize them all and, given limited time, may select some among those to verbalize (Barkaoui, 2008).

[16] It should be noted that the performance descriptors for the same scores for the integrated vs. independent tasks in the rubric are slightly different. In addition, the same score level for an integrated task may not require the same level of performance as that for an independent task given the greater complexity of the former.

List of Appendixes

Page

Appendix A
A List of Strategic Behaviors

This is a compilation of L2 use, learning, test-taking, and communication strategies found in the literature.

Communication Strategies: Involving conscious plans for solving a linguistic problem in order to reach a communicative goal

Reduction Strategies:

Topic avoidance: Avoiding topic areas that pose linguistic difficulties

Message abandonment: Leaving a message unfinished because of linguistic difficulties

Semantic reduction: Changing a message (e.g., reducing the scope of message) rather than abandoning the message

Achievement Strategies:

Guessing using linguistic or other clues

Approximation: Use of such strategies as lexical substitution, over-generalization, and exemplification

Paraphrase: Use of circumlocution, synonym, word coinage, and morphological creativity

Interlingual strategies: Use of such strategies as borrowing and "foreignizing" literal translation

Stalling/time-gaining strategies: Use of verbal fillers or formulaic expressions

Restructuring: Reconstruction of the sentence to deal with linguistic limitations

Cognitive Strategies: Involving manipulating the target language for understanding and producing language

Selecting (attending)

Comprehending

Clarifying or verifying

Translating

Inferencing

Analyzing contrastively

Analyzing inductively

Reasoning deductively

Storing or memory

Repeating

Associating

Linking with prior knowledge

Summarizing

Using imagery

Using mechanical means to store information

Retrieval or using

Recombining

Applying rules

Transferring

Translating

Practicing naturalistically

Using outside resources

Rehearsing

Metacognitive Strategies: Involving a conscious examination of the learning/test-taking process in order to organize, plan, and evaluate efficient ways of learning/test taking

Goal formation

Organizing

Planning

Evaluating

Affective Strategies: Involving self-talk or mental control over affect

Lowering anxiety

Encouraging self

Appendix B

Analysis of Test Scores

Table B1

Descriptive Statistics for Task and Test Scores (N = 30)

	Min	Max	Mean	SD
Task 1	.0	4.0	2.57	.78
Task 2	1.5	4.0	2.98	.72
Task 3	1.5	4.0	2.82	.69
Task 4	1	4	2.60	.71
Task 5	1.0	4.0	2.72	.83
Task 6	1.5	4.0	2.87	.67
Total Test Score	1.33	3.75	2.76	.61

Table B2

Friedman Test for Comparing Scores Across Tasks

N	30
Chi-square	15.51
Df	5
Asymp. sig.	.01

Table B3

Follow-Up Tests for Comparing Scores Across Tasks (Wilcoxon Signed-Rank Test) (p < .01)

	Z	Asymp. sig. (2-tailed)
Task 2 – Task 1	−2.97	.00
Task 3 – Task 1	−2.10	.04
Task 4 – Task 1	−.25	.80
Task 5 – Task 1	−1.25	.21
Task 6 – Task 1	−2.42	.02

(Table continues)

Table B3 (continued)

	Z	Asymp. sig. (2-tailed)
Task 3 – Task 2	−1.37	.17
Task 4 – Task 2	−2.59	.01
Task 5 – Task 2	−2.17	.03
Task 6 – Task 2	−.91	.37
Task 4 – Task 3	−1.94	.05
Task 5 – Task 3	−1.13	.26
Task 6 – Task 3	−.52	.60
Task 5 – Task 4	−1.13	.26
Task 6 – Task 4	−1.95	.05
Task 6 – Task 5	−1.17	.24

Table B4

Descriptive Statistics for Scores by Task and Test-Taker Study Level

Study level		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Undergraduate	*N*	16	16	16	16	16	16
	M	2.75	3.31	3.09	2.84	3.16	3.16
	SD	.66	.57	.64	.68	.70	.57
Graduate	*N*	14	14	14	14	14	14
	M	2.36	2.61	2.50	2.32	2.21	2.54
	SD	.89	.71	.62	.67	.67	.63

Table B5

Two-Sample Kolmogorov-Smirnov Test for Task Scores by Test-Taker Study Level

		Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Total Score
Most extreme differences	Absolute	.18	.51	.37	.25	.44	.46	.45
	Positive	.00	.00	.00	.00	.00	.00	.00
	Negative	−.18	−.51	−.37	−.25	−.44	−.46	−.45
Kolmogorov-Smirnov Z		.49	1.39	1.00	.68	1.20	1.24	1.22
Asymp. sig. (2-tailed)		.97	.04	.27	.74	.12	.09	.10

Table B6

Correlations Among Task Scores

	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6
Task 1	1.00					
Task 2	.60	1.00				
Task 3	.70	.64	1.00			
Task 4	.65	.50	.70	1.00		
Task 5	.58	.66	.81	.64	1.00	
Task 6	.63	.53	.55	.50	.56	1.00

Note. Spearman *rho*. *N* = 30. All correlations are significant at *p* < .01 (2-tailed).

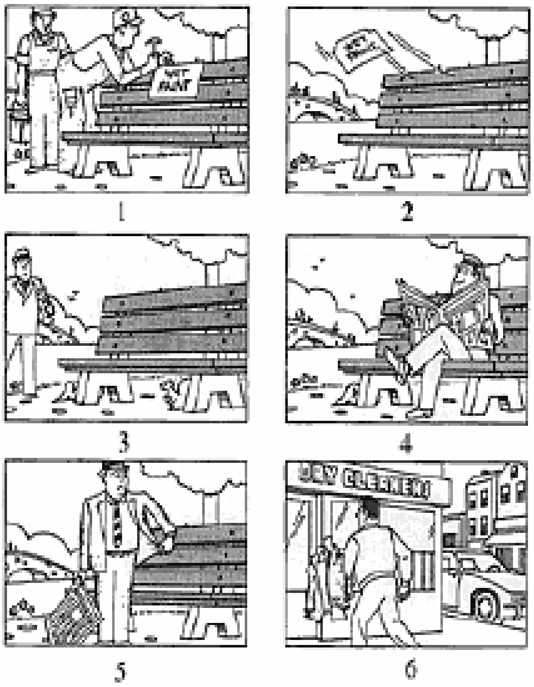
Appendix C

Pretest Proficiency Screening

Part One Story Telling and Answering Questions

Source: Test of Spoken English™ (TSE®).

Please look at the six pictures.



1. I would like you to tell me the story that the pictures show, starting with picture number 1 and going through picture number 6.

   Preparation time: 15 seconds

   Response time: 1 minute

2. What could have been done to prevent this situation?

   Preparation time: 15 seconds

   Response time: 1 minute

3. The man in the pictures is reading a newspaper. Both newspapers and television news programs can be good sources of information about current events. What do you think are the advantages and disadvantages of each of these sources?

Preparation time: 1 minute

Response time: 2 minutes

Part Two Free Choice (Source: SSTiBT)

Examiner: Describe a class you have taken in school and explain why the class was important to you. Include details and examples to support your explanation.

Preparation time: 15 seconds

Response time: 45 seconds

Part Three Pair Choice (Source: SSTiBT)

Some universities require first-year students to live in dormitories on campus. Others allow students to live off campus. Which policy do you think is better for first-year students and why? Include details and examples in your explanation.

Preparation time: 15 seconds

Response time: 45 seconds

Appendix D
Individual Profile Questionnaire

Your name, e-mail address and personal information will be kept confidential. Thank you very much for your time.

Name:	Gender: □ Male □ Female
E-mail Address:	Age:

1. What is the highest level of education you have completed? Please list degrees or certificates that you have received.
2. Do you have any special linguistic qualifications? That is, are you a translator or interpreter, a foreign language teacher, or anything of that kind?
3. Where were you born?
4. What language did you learn first?
5. What was the first foreign language you learned? How old were you when you learned it?
6. At what age did you have your first contact with English?
7. For how long have you been learning English?
8. What countries have you stayed or lived in where English is spoken? How long did you stay in each one of them?
9. Please list any other languages you have learned, with the age that each was learned, and indicate (by circling one of the levels provided below) how well you can speak these languages now.

Language	Age	Level
		Elementary/Intermediate/Advanced/Native-like
		Elementary/Intermediate/Advanced/Native-like
		Elementary/Intermediate/Advanced/Native-like
		Elementary/Intermediate/Advanced/Native-like

10. For how long have you been in Canada?

11. Which language(s) do you use at home?

12. Which language(s) do you use at work?

13. Which language(s) do you use socially?

14. How many hours per day do you speak with people who are fluent English speakers?

15. Have you taken any of the following proficiency tests? If not, please mark "N/A," and, if yes, please provide the score(s) and the year you took the test.

Test	**Score**	**Year**
TOEFL (Test of English as a Foreign Language)		
TSE (Test of Spoken English)		
MELAB (Michigan English Language Assessment Battery)		
IELTS (International English Language Testing System)		
Other		

Appendix E
Stimulated Recall Instructions

A . Instructions for the Research Assistant to keep in mind while conducting the stimulated recall sessions:

1. Make the participant feel comfortable.
2. Explain the purpose of the session.
3. Provide instructions about what the participant is expected to do.
4. Ensure that the video camera is working.
5. Do not direct the participant's responses. Do not go beyond "What you were thinking then?"
6. If the response is that he/she does not remember, do not pursue the matter.
7. Do not provide definite reactions to participants' responses. Use back-channelling or nonresponses—oh; I see; okay; uh-huh, etc.
8. Pause the tape when the participant is talking during playback.

B . Script:

In this study, I am interested in learning what you think about as you carry out the six speaking tasks administered over the internet. To do this, I am going to first record you during the test-taking process. After each task you complete, I am going to play back the video clip and ask you to think aloud. By "think aloud," I mean that I want you to recall and say out loud everything that came into your mind before, during, and after you completed each task. It is important that you do not plan or try to explain to me what you are thinking, and it is important that you keep talking. You may speak in English or in Mandarin, whichever comes naturally to you when you are recalling the thoughts you had before, during, and after completing each speaking task. It is important that you keep talking in English or in Mandarin. If you are silent for any period of time, I will remind you to keep talking.

Do you understand what I am asking you to do? Do you have any questions?

Now we are going to watch the video clip of you performing the 1st/2nd/3rd/4th/5th/6th task. I'd like you to tell me what you were thinking before, during, and after you completed the task. I am interested in what was in your mind as early as when you began working on the task up until

the time when you finished the task. Please do not think about what you think you may have or should have thought or done. I do not want you to try to perform the speaking task again.
I am going to put the remote control on the table here, and you can pause the video anytime that you want. So, if you want to tell me something about what you were thinking, you can push the pause button. If I have a question about what you were thinking or what you have said, then I will push pause and ask you to clarify that part of the video. Please go ahead and tell me what you can remember.

[Model stopping the video and asking a question.]

C . Possible questions to ask during the stimulated recall session:

- What were you thinking here at this point?
- What were you thinking just then?
- Can you tell me what you were thinking at that point?
- I see you appear confused. What were you thinking then?
- I see you are. . . . . What were you thinking then?
- Is there anything else that comes to your mind?
- Do you remember thinking anything when you . . . ?
- Can you remember what you were thinking when . . . ?

Appendix F

Coding Scheme

	Definition/ substrategy	Example(s)
Approach strategies : What the test-taker does to orient him- or herself to the task		
Recalling the task type	Test-taker thinking about the task’s format	这一题我一听，也是 read a passage, 跟第三题的形式是一样的。 (UHTS, Task 4) (When I heard this question, I was asked to read a passage also; this was the same task type as the third question.)
Recalling the question	Test-taker thinking about the meaning of the question	就是，想明白这个题让我干嘛。是让我，就是重新，重述，复述一遍这两个 example 呢，还是让我评价这个？(GYL, Task 4) (That is, I was trying to understand what was asked of me in this question. Was the intention for me to restate the two examples or to evaluate this [what I had heard]?)
Recalling the text	Test-taker thinking about the reading	听到这个题目的时候，我就在想，回忆那个文章。(UJZ, Task 3) (When I heard the question, I was thinking…, that is, trying to recall the reading passage.)
Recalling the dialogue	Test-taker thinking about the dialogue	就在想，在整个 conversation 中到底说了些什么东西。然后就一直在回忆，go over, 大概，把那个人讲的在脑子里 go over 一遍。 (UJZ, Task 3) (I was thinking about what was mentioned in the entire conversation. Then I kept recalling and going over… I roughly went over once in my head what the speaker had said.)
Recalling the lecture	Test-taker thinking about the lecture	就是说，他这段话，到底想说什么，他的- 他的中心论点到底是什么。(UMW, Task 4) (I was thinking what the speaker was trying to say in his talk, what his main thesis/idea was.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Generating choices	Test-taker generating choices	我第一个我就开始想我有什么地方去的,… 后来我就，当时就犹豫了三个地方,一个 museum,一个 library, 还有就是 shopping 的。(GSX, Task 1) (First, I started thinking about where I could go. Then I thought about three places—a museum, a library, and a shopping [mall].)
Making choices	Test-taker narrowing down the choices regarding the topic	当时我就在想，常去的地方。脑子里就在想，有几个选择。当时我第一个反应就是，因为我家楼下就是一个图书馆。然后，第一个反应是那儿。但是，我又觉得可能挺难说的。然后，我又想去那个，就是，去那个，体育馆锻炼。因为在学校的时候，经常去。但是觉得，也不是很好说。然后，后来，又- 脑中又想象出一个，一个(???)park，然后，那个地方，让我感觉最舒服，而且感受最多，然后，而且在国内的时候也是喜欢去嘛。 (GUS, Task 1) (At the time, I was thinking about places I go often. I thought about several choices. At that time, my first reaction was the library, because there is a library right below where I live. But I felt that it [the library] was hard to talk about. Then, I thought of going to a gym to exercise, because I went there often when I was in school. Then I felt that that was also not easy to talk about. Then, an image of the […] park emerged in my head. I felt the most comfortable with that place and the place also evoked many thoughts and emotions; I also liked going there in China.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Developing reasons for choosing what to say/do	Things that the test-taker says indicate the reason for doing what he/she does	后来 - 后来当然这个时候有几个 choice 之后，肯定它有一个 priority level，嗯，当时我的想法是说找最简单的来说，所以我就不是说最简单的说，而是说自己最有的可以说的，但是又不能说太容易，太容易觉得表现不出水平来。所以说我觉得 Library 还是很好的，因为它有很多生活内容在里面，它可以借各种各样的东西，还可以学习…。(UHTS, Task 1) (Then… of course I had several choices; there is an order of preference. At that time, my thinking was to talk about the thing that was the easiest to talk about. So I… well, not the easiest, but what I could say the most [about]. It shouldn't be too easy, though, because if it's too easy, it will not show my level of competency. I thought that the library was a good [place to talk about], because it had much life-related content and a variety of things that could be borrowed. We could also learn there….)
Communication strategies: Involving conscious plans for solving a linguistic problem in order to reach a communicative goal		
Simplifying the message	Test-taker thinking about changing a message by simplifying it	因为我觉得最后一个这个观点，说的简单一点，因为没有前面那个那么清楚。清楚地东西我就想说得好一点，不是特别清楚的东西我就想说得简单明朗一些。(GXS, Task 6) (I felt that the final viewpoint should be simplified because, unlike the previous one, I was not clear what the point was. With things that I was clear about, I would say it better; for things that I was not especially clear about, I would simplify it.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Avoiding	Test-taker thinking about avoiding areas that pose linguistic difficulties	这个我碰到问题就是说，可能我已经想到东西了，但是我就不知怎么用英语表达出来。就比如说，我想，知识都是融会贯通的啊，或者那种，我其实很想说，但是我觉得，因为我限于这个，而且我又不可能在这里揣摩，因为它这 45 秒。我就想想，就尽量避免一些，但是呢，又尽量能够表达些我的意思，那种。 (GSX, Task 2) (The problem I faced was that I probably thought of things to say, but I did not know how to express them in English. For example, I thought of saying “thorough understanding” and whatnot. I really wanted to express those thoughts, but I felt that, because of this limitation, plus the fact that it’s impossible for me to mull over [how to express thoughts that I couldn’t express in English] due to the 45 seconds [of speaking time]. I then thought that I would try to avoid them, but would still try to express my meanings.)
Using Chinese	Test-taker using Chinese to formulate what to say	而且有的时候你要是用中文你要是再转成英文,就<慢/乱>了…中文想的。因为很久很久,我觉得好象自从离开高中以后就很少很少用英文去想东西了。(UDD, Task 2) (Sometimes when you have to translate from Chinese to English, it’s slow/messy)… I used Chinese to think, because it’s been a very long time… I felt that I very rarely think about things in English ever since I graduated from high school.)
Paraphrasing	Test-taker restating in another form or with other words to clarify meaning	嗯，就试着想，自己该怎么样，就是把他所说的内容用自己的语言再来重复一遍。(UDZ, Task 3) (er… I was trying to think about how I could use my own words to state the content mentioned by the speaker again.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Paraphrasing	Test-taker restating the thought in another form or with other words to avoid repetitions	主要想的就是那三个开头句，怎么替换掉。因为我们写作文的时候，都写一样的话，就完蛋了。(URX, Task 3) ([I was] mainly thinking about the three opening sentences—how to change it. Because if we were writing an English composition, we would be doomed if we used the same words.) 嗯，这个地方我有点卡住了。我在想怎么样，因为我要说三个那个，points，我在想，怎么样说，不会太重复了，整天用什么 he thinks 之类的话然后，怎么说呢， 当时紧张，根本就没有想。就是说如果事先想到了，好一点。 我三个卡住的地方就是卡这一点，我想怎么样不要太，重用一句话。(URX, Task 3) (um… I was a bit stuck here. Because I needed to state three points, I kept thinking about how to say them so that they don’t sound repetitive. If I kept using phrases like “he thinks,”… how should I put this… I was nervous and didn’t think much. It would’ve been better if I had thought the points through in advance. Of the three points, I got stuck on this one; I was thinking about how I can make it less repetitive.)
Approximating	Test-taker using lexical or grammatical substitution	这个时候我在想，就是她已经答应，答应那个我没想到应该用那个词，其实应该用promise。Promise 不是特别好，应该用更好的词嘛。后来就没想起来，就只能说talk to another prof 去帮助那个 museum。(UVQ, Task 5) (At this time, I was thinking that she had already promised... promised that... I couldn’t think of what word to use. I should use the word “promise.” The word “promise” was not a good word, either. There should be a better word. Then I couldn’t think of any word, so I could only say “talk to another professor” about assisting in the museum work.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Linking to prior experiences/ knowledge	Test-taker making connections between what is known or his/her previous experience and what he/she is reading or hearing	他这个 topic 出来之后，看了一下之后，突然觉得挺高兴的，因为这个东西以前在 society 课上学过，我在想说，因为像这种考试的话，他给你一个 topic，你如果知道的话，肯定你就 understand。我就想说，这个挺不错的，知道怎么回事，而不是出来一个东西，不知道是什么。(UJZ, Task 4) (When the topic came up, I was pretty happy after seeing it, because I had learned about it in a sociology class before. I was thinking that, if you are given a topic that you are familiar with in a test like this, you can understand everything. I was thinking that this is not bad, knowing what's going on, rather than not knowing what it's about when a topic comes up.)
Borrowing	Test-taker thinking about using words heard or read during the listening or reading activity in the response	第一句话，我不想太唐突，然后我就照着它东西先读一下。我想就是说，我觉得这是，我觉得，不知道是不是技巧，等于说，第一你可以说点东西，而且给人感觉你比较哦，就是说，fluency 那种，比较 smooth 那种感觉。(GSX, Task 1) (I didn't want to be too abrupt with the opening sentence, so I read what was written. I thought that… I don't know if this is a technique, but on the one hand, you would have something to say, and, on the other, you would give the impression that you are more fluent and smoother.)
	Test-taker borrowing phrases from the question in order to gain time	这段我是读的，读的上面的。I think it's better to— to—whatever，因为是屏幕上的那个题目嘛。因为我，那段时间我就在想，我就照读吧，我这段时间脑子里还可以想一想，救急想一遍，然后就这样。(UMW, Task 2) (This passage I read from what was on the [computer] screen. I thought that it would be better to… because it was the question on the screen … I was thinking that I would just read it to buy some time to think over [what to say]—a quick chance to think over once. That was it.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Reviewing notes	Test-taker reviewing the notes in order to remember/formulate what to say	当时就是再读一遍，让自己更加，更加，就似乎 familiar with sentences, familiar 刚才的内容…。(UDZ, Task 6) (At that time, I read the notes again in order to become more familiar with the sentences and the content [of the talk] earlier.)
Referring to notes	Test-taker referring to the notes in order to remember/ formulate what to say	就是看一下, 说一下。我发现我做这一题时老是往本子上看, 因为本子上写的东西很多。(UHTS, Task 4) (I read [from the notes] and spoke. I found that I kept reading from the notes for this task, because I wrote down a lot of notes.)
Organizing thoughts	Test-taker organizing ideas while speaking	这时候就想，接下来就说 his reason 吧，因为没有组织好，就上来就开始 reasons, 然后一想，这 reason 也得一丶二丶三说。前头先说一个，它一共有几个。 (GYL, Task 3) (At that time, I was thinking that I should talk about the reason next, but I did not organize the points well. I started to talk about the reasons. Then I thought that the reasons should also have points one, two, and three. So I then mentioned how many reasons there were at the beginning.)
Guessing	Test-taker guessing by using linguistic or other clues	后面那个部分就是说，嗯，就是因为前面那句话，cave 那句话没听懂，我就不知道，不知道那句话，那句话到底是算在后面这部分呢，还是算在前面那部分。然后，所以，当时我就主要是在猜他什么意思。 (GSP, Task 5) (Later in that part, because I did not understand the word "cave," I did not know what that sentence should have belonged to—the later segment or the earlier segment. So I was mainly guessing what the speaker meant at that time.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Repeating	Test-taker repeating unfamiliar words	然后我又想一想这 field trip，因为我在心里重复的一遍，因为这是一个，（…）就是相对于，怎么说呢，在，比较生疏，但是，听过的词。(GSP, Task 5) (Then I was thinking about this “field trip,” and I silently repeated the phrase to myself, because this is… how should I say it… a relatively unfamiliar phrase, but I have heard it before.)
	Test-taker repeating phrases in order to fill the time	就是这个地方我一下子，reorganize 好语言了，然后就是一边想一边说，当时想着时间<怎么这么多?>，[笑] 然后就一句话重复了那么多遍。(UJZ, Task 2) (At this spot, I suddenly finished reorganizing the sentence, and then was thinking and speaking at the same time. At that time, I looked at the clock and thought about why there would be <so much> time left. [laugh] Then I repeated the same sentence so many times.)
Rehearsing	Test-taker mentally rehearsing what to say	就是，就是没念出来，在嘴里念，其实是在嘴里说。因为它 30 秒的准备时间，你想它 30 秒的准备时间，60 秒的说的时间，然后我如果就用这 30 秒去说的话，我至少可以说一半。…相当于重说一遍吧。当然，这个，因为没有记下来，只是说，大概排演一遍，下面一遍的话，就看能记住多少是多少，它要是给 60 秒的时间就可以完全 rehearse 一遍了。(GJL, Task 3) (That is, I did not read it aloud, but silently read it. Because there were 30 seconds of preparation time and 60 seconds of speaking time, if I used the 30 seconds to talk, at least I could say half of the answer. It was like repeating it once. However, because I did not write down points, I roughly rehearsed it once, and then tried to remember it as much as I could. If I had been given 60 seconds, I would’ve been able to rehearse once fully.)
Reading ahead	Test-taker reading ahead in order to gain time	在『begin reading now』 那句话的前面就开始 reading 了。(GSP, Task 3) (Before the speaker said “Begin reading now,” I already had started reading.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Restructuring	Test-taker restructuring the sentence in order to deal with linguistic limitations	当时看了一下屏幕是因为，第一句话说到一半儿的时候，突然，语言 organize 上出问题了。然后，看那个，想问题到底原话是怎么说的。Try to answer- 就是第一句，我在想说怎么开这个头嘛。比如说，他既然提出一个问题，第一句话就是，先把那个 point state 出来，我一看那个题目是怎么，我就 try to answer。 (UJZ, Task 3) (At that time, I looked up at the computer screen because I was having problems organizing my response when I was halfway through the first sentence. Then I was thinking how the question was worded originally. I tried to respond…I was thinking how to start the first line. For example, a question was given, so I would try to state that point. I read the question and tried to respond [using the question].)
Stalling	Test-taker using verbal fillers, formulaic expressions, and so on to gain time	中间有个 pause，这 pause 的话，我就想把它填起来，因为我的话还没想好怎么说，我就把它填起来，然后，我就嘴上就说的是，说了几个，这个单词，好象跟那个，呲，跟那个，跟那个内容不相关的就是。(GSY, Task 3) (There was a pause in the middle. I was thinking about filling the pause. Because I hadn’t thought of how/what to say, I then said a few words that had nothing to do with the content in order to fill the pause.)
Slowing	Test-taker slowing down the speed of delivery to gain time to formulate speech	可是我后来看了一下时间，好像还 ok。我说干脆就，我就不想 organize 这个 point，我就干脆讲的慢一点，把他的 detail point 讲出来。(UBT, Task 2) (Then I took a look at the clock and thought that it was fine. I then thought that I might as well not organize the point. I would just go slowly in order to deliver the point in detail.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Slowing	Test-taker slowing down the speed of delivery to avoid making mistakes	刚才我在说这个的时候，因为又是单数复数的问题，我特别讨厌，然后我就- 所以后来，我就说得很慢，因为我怕，我怕，make a mistake，因为我总是单数复数这个，该加 s 还是不加 s，我总是会弄错。所以，当时我说，a group of people tend to tie their shoes，说得很慢，就是这样。 (UMW, Task 4) (Earlier, when I was talking about this, because it was the singular and plural problem that I hate the most, I talked very slowly, because I was afraid that I would make a mistake. It's always the singular and plural issue; I always make mistakes about whether to add –s or not. So when I got to "a group of people tend to tie their shoes," I said it very slowly, just like this.)
Summarizing the task	Test-taker making verbal summaries of the target information	先大概 summarize 一下他这个 task 到底是在问什么，然后我就想说，做一个 brief introduction...。 (UJZ, Task 5) (I roughly summarized what was asked in the task first. Then I thought of giving a brief introduction….
Thinking ahead	Test-taker thinking ahead	就是一看到那个题目，不管他在读什么，我就开始想我想说的一个 place。(UHTS, Task 1) (As soon as I read the question, I didn't care what he was reading. I had already started thinking about the place that I wanted to speak about.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Elaborating to fill time	Test-taker elaborating on the points in order to fill the time	但是我想我最后凑到了最后一秒，我一直在看着这个表，我在说的时候一直看着，中间我最后说的是，最，最后一段话是 redundant, 我，我说是 to do a good job on a research work, actually is redundant, 但是呢，我凑，当时还有 5 秒钟，因此我说的拉了一点长度，就是凑够那个时间。(GYGW, Task 1) (But I thought that I would fill the time until the last second. I kept looking at the clock. While I was speaking, I kept looking at it. What I said in the last segment—I said “to do a good job on research work” was actually redundant, but, when I finished my response, there were still five seconds left. So I elaborated further in order to fill the time.)
Elaborating to fill time	Test-taker elaborating on points that might not be relevant to the question in order to fill the time	在这地方就是，因为，嗯，在这地方我想呢是因为，嗯，因为每次我都是无话可说嘛，[笑]然后我就想干，干脆我就解释一下这两个人到底是谁。一，这样，一方面我不会犯错，另外一方面，对回答整个，对回答整个题来说也是，也不算是，一个完全不沾边的一个部分。然后我就把这句话加进来的，因为这样可以，就是不至于觉得没话可说。(GSP, Task 5) (Here, I thought that, because I had nothing to say [laughing], I thought that I might as well explain who the two speakers were. In that way, on the one hand, I wouldn’t make mistakes. On the other hand, in terms of responding to the question, what I said would not be totally irrelevant. I added that sentence, because then I wouldn’t feel that I had nothing to say.)
Elaborating to clarify meaning	Test-taker elaborating on the points in order to clarify meaning	第一个我说 study, 我本来就想到我去那 study，然后，我就想，我要再解释得清楚一点。就是，many people study there, so that- 我当时说的是，so that makes me study too。(UMW, Task 1) (First I said “study.” I originally thought about where I studied. Then I thought that I needed to explain more clearly, that is, that many people studied there, so that.... at that time, I said that “so that makes me study too.”)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
		Cognitive strategies: involving manipulating the target language to understand and produce language
Attending	Test-taker directing attention to or concentrating on a specific aspect of the task	我就写男生的… 因为我发现女生讲话很少，她只是在 continue 那个 conversation 女生好像只是 repeat 那个 argument…。 (UBT, Task 3) (I was writing about what the male speaker said… because I noticed that the female speaker said very little; she was merely continuing the conversation… She seemed to be only repeating that [the male speaker's] argument.)
Anticipating the content	Test-taker anticipating the content	阅读的时候，我一边读一边就是，一个人在想说，接下来那个 topic，那个 Prof. 会讲些什么东西。 (UJZ, Task 4) (While I was reading the passage, I was thinking: What would the topic be? What would the professor be talking about later?)
Anticipating the structure	Test-taker anticipating the structure of talk during a listening activity	当时有，就是因为从他说话的 pattern，我觉得他会举好几个 example，然后就在，就是在纸上写下第一个、第二个 point，这样的。 (UDZ, Task 4) (At that time, on the basis of the speaker's speech pattern, I felt that the speaker would provide several examples. I then wrote down on the paper the first and the second point.)
Anticipating the question	Test-taker anticipating the question	当时可能在，嗯，在- -在-在预测题目。我觉得可能，要么他就是提出，就两种情况，一种就是，这个男人说什么，怎么去，然后就或者这个女人说什么，怎么去。当然就这两种可能，然后我再想，出现哪个…。(GLW, Task 3) (At that time, I was predicting the question. I felt that there would probably be two types of scenarios—one would be about what the man said and so on, or it would be about what the woman said and so on. Of course, both are possible. Then I thought again about which question would be more likely to show up….)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Using imagery	Test-taker using visual images, either generated or actual, to understand, think, or remember information	后来可能潜意识里会想到说运动场馆那种东西， 当时想着有一个踢足球的场景…。(UHTS, Task 1) (After that, subconsciously I started thinking something about a stadium, imaging a soccer game….)
Using mechanical means to organize or remember information	Test-taker writing things down to organize or remember information	但是因为这个 topic 我一看对我来说太不熟悉，(???)太不熟悉，我就知道它是个 tough task 对我来说，因为这个<social interactive> 我就开始记，我怕我忘掉了，因此我在这里做的笔记包括，在读的过程当中，social interaction, influence behavior, because 我不知道它后来它会 focus on which part, and audience effect，他们对我说是比较陌生的东西，我可能，从理解的角度上会更难以理解一些，所以我就会 [举起笔记示意] 记了这么多。把中间要用到的词，跟我想到的词记下来。(GYGW, Task 4) (Because I was not familiar with this topic, I knew that it would be a tough task for me. So I started writing down notes because I was afraid that I would forget. While I was reading, I wrote down here "social interaction," "influence behavior" because I did not know which part would be the question's focus. Due to unfamiliarity with the subject matter, it was more difficult for me to understand in the aspect of reading comprehension. So I wrote down so many notes. [showing the notepad] I wrote down words that I would need and words that I could think of at the time.)
	Test -taker using symbols for drawing attention during delivery	[在笔记上示意画线] 然后把准备要说的画上，要都是 underline，或者是要不 circle， 表示我这里一会儿眼睛看着这要说。 (GYL, Task 3) [Showing drawing lines on the notes] (Then, I underlined those that I planned to say. I either underlined the words or circled them to remind myself to make those points when my viewing reached those points during delivery.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Using mechanical means to organize or remember information	Test-taker writing down information in numerical form during a listening or reading activity	然后我想，这点因为要先写，所以我就划了个1，把之前记下来的划了个2和3。这样就是，remind myself，待会儿我说的时候按顺序说，先说这个，然后再说这个。 (UMW, Task 3) Then I thought that, because this point must be written first, I drew a 1. I drew a 2 and 3 for what I had written down earlier. I did so to remind myself to follow the order when I respond later—to say this first and that later.)
	Test-taker mapping information to organize notes during a listening or reading activity	听的过程中非常认真地来记，我一般是这样记的，我把它分成两部分，因为它是一个conversation，我分成左边和右边，这边是<Marilyn>，那便是<Jason>，然后中间-- 如果有时间的话，最好能把这些idea connect起来，这样就会，就会比较好。怎么说呢，好整理这些东西。 (UZYH, Task 5) (I wrote down notes very seriously when I was listening. This is how I did it. I divided the page into the left and the right sides because this was a conversation. This side was about Marilyn and the other side was about Jason. When I had time, I used the middle section to connect the ideas from both sides. In this way, it would be easier for me to respond… how do I say it? —easier to organize information.)
Memorizing	Test-taker trying to memorize what was said in the dialogue/lecture or what was written in the text	我当时，听他们两个对话的时候，我就在想说，就尽量memorize，我本来想，本来是想take notes，他们说得太快了，我写的话，我一take notes可能会耽误听下面 的sentence。我就想一边听，一边在memorize他们俩到底在说什么东西。 (UJZ, Task 3) (At that time when I was listening to the dialogue, I was thinking that I should try to memorize… I thought about taking notes, but they spoke too quickly, and taking notes would have probably negatively affected my listening. So I thought that I would try to listen and memorize what they were saying at the same time.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Summarizing	Test-taker making mental summaries of the target information	我在想，我实际上要 summarize 它的那个主，它的东西嘛。它那么多东西，记也记不清楚，(…) 大致是个什么意思，它说就是，这个，（ … ）实际上就是，（ … ）别人跟你的出现，或者别人在旁边这些东西就影，就会影响你的行为。 (GZM, Task 4) (I was thinking… in fact, I was trying to summarize the main [point]. There were so many bits of information, [it was] hard to remember everything clearly…. What was the lecture about? In a nutshell, basically it was about how being watched would change one's behavior.)
Translating	Test-taker seeking to understand by translating from L1 to the target language during a speaking or listening activity	他说，increase speed, 我说对啊, 是干得快一些， 也会出很多错，然后我在想这个当然这都是中文啦。(UDD, Task 4) (The speaker said that the speed would increase. I thought that what he said was correct. One would do it faster and, as a result, would make many mistakes. When I was thinking about this, it was of course all in Chinese.)
	Test-taker seeking to formulate speech by translating from L1 to the target language	考试时候，在做这种题目的时候感觉，很，很模糊啊，很慢，然后，其实一段时间，英文，汉语组织句子再想英文，那我在想说，如果说，如果再让我，在过去的来考的话，那，那 20 秒的时间要让我先想要讲什么，再把它翻成英文，再讲的话，恐怕会比这个更难。(UBT, Task 6) (During the test, it all felt very blurry and slow to me when I was doing the task. Actually, I used Chinese to formulate sentences, and then I translated them into English. I was thinking that, if I were taking this test in the past, the 20 seconds would not have been enough for me to think about what to say, translate it into English, and then say it; it probably would have been more difficult than it is now.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Inferencing	Test-taker seeking to understand by using information in the text, dialogue, or monologue to guess the meanings of linguistic items or to make up missing information	当时是有一点太记笔记了，所以有些话可能没听到，你像她说 I've ridden the buses，当时我就想了一下，她说 I've read the buses, 是说从报纸上读 到的 『 Imitating the action of opening a newspaper』 ， 还是，我当时想的是 I've read the buses, 就是说 bus issue 那样。后来提到 There were only a few people, 我一想应该是在坐 buses。(UHTS, Task 3) (I took down too many notes then, so I didn't hear some words. For example, when she said "I've ridden the buses," I thought for a moment, and I thought she said that "I've read the buses," that is, read it from a newspapers [imitating the action of reading a newspaper] or, I thought that she said, "I've read the buses," that is, the bus issue. Then later, when it was mentioned that there were only a few people, I figured out that the sentence should had been about riding buses.)
Processing inductively	Test-taker seeking to make generalizations	因为我已经把所有的 key points 都记在我的本子上了，然后就在想如何去想，其中包括下面的准备阶段。因为我的习惯呢，一般是做一个 general conclusion 之类的。(GYJW, Task 4) (Since I already had jotted down all the key points on the notepad, I was thinking about how/what to think, including the preparation stage. This was because my habit usually is to make a general conclusion.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
		Metacognitive strategies: involving organizing, planning, and evaluating
Setting goals	Test-taker setting a goal for completing a task	第一我，基本上两个东西我肯定在规定的时间能说完，所以这个时候我就想，尽量把语言不要有那种，口语是什么啊，不加s啊，这种东西，万一有什么错，我就把它纠正，弄弄好那种感觉，就是说，给他感觉我句子都是正确的。不象前边我句子我就不care 了，但我要把我的观点说出来。 (GSX, Task 4) (First, I was certain that I could cover the two basic points within the given time. So I thought that I would do my best to avoid, for example, colloquial expressions, or missing the plural “s.” Also, if I made any mistakes while I was speaking, I would correct them, trying to make sure that the sentences were correct. Not like when I was doing the tasks earlier. Then I did not care about mistakes, but about getting my ideas across.)
Identifying the purpose of the task	Test-taker identifying the purpose of the task: purposeful listening, reading, and/or speaking	第一个呢想一想，他实际上有几条方案嘛，就两条方案，是吧？所以，把两条方案要找出来啊。…然后，他因为他还问了一句，你要，解释它的原因嘛，也就是说你，那么我们自己的，这个题我肯定要想，第一个，把两个原因要讲出来，是吧？(GZM, Task 5) (First, I was thinking how many cases. Two cases, right? So I need to look for the two cases. Then, the speaker stated that you [the test-taker] needed to explain the reasons…. For this question, I needed to first state the two reasons, right?)
Planning	Test-taker planning the parts, sequence, or main ideas to be expressed verbally	就是按照，就是说，先把它的argument，它的idea 说出来，然后再support 它的argument，最后我就想着再总结一下。(GYFZ, Task 3) (First, I would state the argument/its idea, and then support the argument. Finally, I thought that I would make a conclusion.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Monitoring	Test-taker monitoring the clock while reading, listening, preparing, or speaking	然后他那个给了45秒让读。我自己观察，特意观察了一下，等我读完的时候，还剩25 秒。(GYGW, Task 3) (Then I was given 45 seconds to read. I observed myself, and especially observed that when I finished reading the passage, I had 25 seconds left.) 这个地方停一下，我想跟你说一下。其实刚才我在看screen的时候，我就在找时间，因为我觉得，我当时想的是一个point说30秒钟，这样的。然后我已看已经到了30秒钟，就必须停下来，然后开始第二个point…。 (UZYH, Task 3) (I paused a while here. Let me tell you. Actually, when I was looking at the [computer] screen, I was looking for the clock. Because I felt that… I was thinking that I would use the 30 seconds to talk about one point. Then I saw that the clock had already reached 30 seconds, so I had to stop and start the second point.)
Self-correcting	Test-taker self-correcting errors in his/her own pronunciation, vocabulary, grammar, etc.	这个地方就是，我先想的是the reason he give, he gives，然后，后来想不对，我在想grammar，应该是reasons，因为他给了不止一个reason。然后脑子就是 grammar的问题，但是习惯性的，习惯性的中国人，不会考虑单数复数，因为我们没有，所有我总是make a mistake。然后，但是，我说过之后，我又改过来了…。(UMW, Task 3) (Here I first thought of the reason he gave. Then I thought that that was not correct. I was thinking about grammar—it should've been "reasons" because he gave more than one reason. Then I was thinking about grammatical problems. Chinese speakers habitually do not think about singular or plural, because we don't have such a distinction. As a result, I always make mistakes. But after I verbalized, I self-corrected.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Evaluating previous performance	Test-taker evaluating his/her performance in the previous task	上一次做的时候。觉得那个比较费时间，而且不work，所以就- 还是用key word。(UZYH, Task 1) (The last time when I did it [taking notes in full sentences], I felt that it was too time consuming and that it did not work. So I used key words instead.)
Evaluating the content of what was read/heard	Test-taker evaluating the content of what he/she read or heard	脑子里闪过，他讲的这些东西到底是不是true。(ULS, Task 4) (A thought flashed through my mind about whether what the speaker said was true or not.)
Evaluating performance	Test-taker evaluating language production while speaking	那个时候我想说，哇，其实我都没有写那种sentence 的notes，都是points，然后讲到一半儿的时候，发现整理这些东西怎么这么麻烦！因为又是research，又是那个group，然后被看，和没被看到。就是，它的description 都很长。(UBT, Task 4) (At that time, I was thinking that, wow, I didn’t write any notes with full sentences; they were all in point form. Later, halfway through my talk, I discovered that it was so troublesome to organize the points, because it was about the research, about which group was observed and which was not. That is, the description was very lengthy.)
Evaluating language production	Test-taker evaluating language production after completing a task	后来我想我为什么要repeat 它呢，浪费了我好多时间。 (UJZ, Task 2) (Then I thought about why I repeated it [the question] again—it wasted so much of my time.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
		Affective strategies: involving self-talk or mental control over affect
Lowering anxiety	Test-taker reducing anxiety by taking a break or using techniques	第一句话，我不想太唐突，然后我就照着它东西先读一下… 然后一方面说的时候我等于是在让自己一方面镇定一下…。(GSX, Task 1) (The opening line—I didn’t want to be too abrupt. I just read according to what was shown [on the computer screen]. Then I calmed myself down while I was speaking.)
Encouraging self	Test-taker encouraging him/herself through positive statements	我当时在想，那个，因为我觉得我前面，前几个问题回答的都很糟糕嘛，我还在想，哎呀，这个问题我要好好回答一下。 (UMW, Task 6) (I was thinking at that time that, because I felt that I performed very poorly on the previous few tasks, I thought that I would do my best on this question.)

(Table continues)

Table (continued)

	Definition/ substrategy	Example(s)
Justifying performance	Test-taker justifying his/her performance	我当时，就是，呃，觉得好象是，就脑子里觉得这东西好象很清楚，（???）我的意识上觉得，如果，我要用中文来说，可能会很清楚，或怎么着，然后的话，就没怎么太太太准备这，这个细节上怎么说，就是说。但是到中间的话，就是，(…) 到中间的话，一想，哎哟，这，这好象，这好象要把所有的话都说到这个，这个，清清楚楚名字来的话，还真是需要好好准备一下。把这个，要自己要说的哪些话，如果要是在准备过程中都，都象这样稍微准备一下的话，这样可能说起来会更好一点。 (GSY, Task 6) (At that time, I felt that I knew very clearly what to say. I was aware that, if I were to use Chinese, my response would be very clear. As a result, I did not prepare very much what to say in detail. Then, midway through my speaking, I thought that, whoa, to speak everything, all the names clearly really required much preparation. Had I prepared what I wanted to say [in English] or had I prepared a bit, I would have responded a bit better.) 但是因为紧张，而且因为语言上面还是不是那么的熟练，所以在想把自己想要表达的东西表达出来的过程当中挺费劲的。(UDZ, Task 4) (Quite a bit of effort was required during the process of expressing what I wanted to express.)

Appendix G

Results of Normality Tests

Table G1

Tests of Normality for Test Scores

Task	Shapiro-Wilk		
	Statistic	*df*	Sig.
Task 1	.90	30	.01
Task 2	.91	30	.01
Task 3	.93	30	.06
Task 4	.95	30	.14
Task 5	.93	30	.04
Task 6	.93	30	.04
Average total test score	.97	30	.44

Table G2

Tests of Normality for Test Scores by Test-Taker Study Level

Task	Student group	Shapiro-Wilk		
		Statistic	*df*	Sig.
Task 1	Undergraduate	.95	16	.45
	Graduate	.86	14	.03
Task 2	Undergraduate	.82	16	.01
	Graduate	.86	14	.03
Task 3	Undergraduate	.91	16	.10
	Graduate	.91	14	.16
Task 4	Undergraduate	.90	16	.09
	Graduate	.88	14	.05
Task 5	Undergraduate	.86	16	.02
	Graduate	.88	14	.06
Task 6	Undergraduate	.91	16	.11
	Graduate	.92	14	.22
Average total test score	Undergraduate	.95	16	.47
	Graduate	.91	14	.17

Table G3

Tests of Normality for Strategy Categories

Strategy category	Shapiro-Wilk		
	Statistic	*df*	Sig.
Approach	.98	30	.83
Communication	.95	30	.19
Cognitive	.96	30	.34
Metacognitive	.95	30	.13
Affective	.94	30	.08

Table G4

Tests of Normality for Strategy Categories by Task

Task	Strategy category	Shapiro-Wilk		
		Statistic	*df*	Sig.
Task 1	Approach	.97	30	.46
	Communication	.95	30	.22
	Cognitive	.81	30	.00
	Metacognitive	.93	30	.05
	Affective	.47	30	.00
Task 2	Approach	.96	30	.25
	Communication	.97	30	.47
	Cognitive	.90	30	.01
	Metacognitive	.97	30	.56
	Affective	.62	30	.00

(Table continues)

Table G4 (continued)

Task	Strategy category	Shapiro-Wilk		
		Statistic	*df*	*Sig.*
Task 3	Approach	.91	30	.02
	Communication	.94	30	.11
	Cognitive	.95	30	.18
	Metacognitive	.91	30	.02
	Affective	.84	30	.00
Task 4	Approach	.92	30	.03
	Communication	.97	30	.60
	Cognitive	.96	30	.36
	Metacognitive	.95	30	.14
	Affective	.81	30	.00
Task 5	Approach	.92	30	.02
	Communication	.94	30	.11
	Cognitive	.91	30	.01
	Metacognitive	.95	30	.14
	Affective	.72	30	.00
Task 6	Approach	.86	30	.00
	Communication	.98	30	.83
	Cognitive	.97	30	.51
	Metacognitive	.95	30	.21
	Affective	.69	30	.00

Appendix H

Descriptive Statistics for Individual Strategies by Test-Taker Study Level

	Undergraduate		Graduate	
	Median	Range	Median	Range
Approach				
Recalling task type	.81	3.36	.00	1.71
Recalling the question	1.63	6.46	.93	3.03
Recalling the text	.00	.83	.00	1.67
Recalling the dialogue	1.63	4.98	.80	2.33
Recalling the lecture	.89	4.17	.83	3.90
Generating choices	1.45	2.78	.00	3.33
Making choices	2.74	4.66	.00	4.23
Developing reasons	2.38	6.48	4.17	9.48
Communication				
Simplifying the message	.00	2.08	.85	3.34
Avoiding	.00	1.67	.00	4.38
Using Chinese	.00	1.28	.00	5.44
Paraphrasing	.00	1.83	.00	3.13
Approximating	.00	3.37	.00	1.19

(Table continues)

Table (continued)

	Undergraduate		Graduate	
	Median	Range	Median	Range
Linking to prior experiences/knowledge	6.95	13.96	4.69	10.80
Borrowing	.61	4.15	.79	5.03
Reviewing notes	1.66	3.98	.85	2.11
Referring to notes	1.87	4.24	1.86	4.14
Organizing thoughts	10.86	9.42	2.71	14.36
Guessing	.00	2.96	.49	2.98
Repeating	.90	3.75	.24	3.64
Rehearsing	.00	2.76	.00	2.15
Reading ahead	.88	4.63	1.17	3.09
Restructuring	.32	2.80	.00	.46
Slowing	.00	5.59	.00	1.85
Thinking ahead	.00	1.94	.00	.88
Elaborating to fill time	1.19	2.71	1.52	4.31
Elaborating to clarify meaning	.42	2.38	.00	3.33

(Table continues)

Table (continued)

	Undergraduate		Graduate	
	Median	Range	Median	Range
Cognitive				
Attending	2.31	7.54	6.50	12.13
Anticipating the content	2.32	7.05	2.92	6.14
Anticipating the structure	2.29	6.26	2.07	5.81
Using imagery	.00	2.42	.00	1.19
Using mechanical means to organize	11.32	7.60	10.46	25.65
Memorizing	.00	2.45	.00	.00
Summarizing	1.11	2.69	.85	3.28
Translating	.00	1.19	.00	4.33
Inferencing	.67	3.15	.00	1.28
Processing inductively	.00	.88	.00	.00

(Table continues)

Table (continued)

	Undergraduate		Graduate	
	Median	Range	Median	Range
Metacognitive				
Setting goals	2.54	5.74	4.46	10.79
Identifying the purpose of the task	3.57	7.46	1.48	4.80
Planning	3.34	7.12	7.38	15.48
Monitoring	6.70	9.66	2.50	10.42
Self-correcting	.81	4.88	.00	2.21
Evaluating previous performance	.74	2.66	1.37	4.99
Evaluating the content of what was read/heard	3.36	7.08	7.22	16.61
Evaluating performance	3.26	6.39	4.27	10.53
Evaluating language production	4.56	8.35	2.60	12.06
Affective				
Lowering anxiety	.76	4.68	.52	2.38
Encouraging self	.82	4.21	.26	4.47
Justifying performance	.88	3.49	2.50	3.88

Appendix I

Descriptive Statistics for Individual Strategies by Task Group

	Task Group A		Task Group B		Task Group C	
	Median	Range	Median	Range	Median	Range
Approach						
Recalling task type	.00	5.13	.00	6.51	.00	4.17
Recalling the question	.00	14.68	.00	4.85	.00	6.25
Recalling the text	.00	.00	.00	5.01	.00	.00
Recalling the dialogue	.00	.00	1.85	9.38	2.23	5.88
Recalling the lecture	.00	.00	.00	5.88	.00	8.33
Generating choices	2.12	10.00	.00	.00	.00	4.17
Making choices	5.65	17.14	.00	3.33	.00	2.94
Developing reasons	5.28	24.29	2.13	6.76	.00	7.29
Communication						
Simplifying the message	.00	6.90	.00	4.17	.00	6.25
Avoiding	.00	7.14	.00	1.56	.00	13.13
Using Chinese	.00	14.05	.00	2.27	.00	3.33
Paraphrasing	.00	3.13	.00	3.57	.00	9.40
Approximating	.00	10.10	.00	2.38	.00	5.26

(Table continues)

Table (continued)

	Task Group A		Task Group B		Task Group C	
	Median	Range	Median	Range	Median	Range
Linking to prior experiences/knowledge	4.77	21.43	6.92	15.29	5.28	15.61
Borrowing	.00	7.63	.00	7.89	.00	5.56
Reviewing notes	.00	.00	1.74	6.97	.00	9.17
Referring to notes	.00	8.33	1.56	7.39	2.78	7.63
Organizing thoughts	9.71	24.31	5.16	13.87	6.36	14.58
Guessing	.00	7.14	.00	6.67	.00	6.67
Repeating	.00	11.25	.00	3.85	.00	6.25
Rehearsing	.00	6.35	.00	3.33	.00	5.00
Reading ahead	.00	10.00	3.03	5.12	.00	3.13
Restructuring	.00	6.25	.00	1.61	.00	5.00
Slowing	.00	16.76	.00	5.56	.00	5.13
Thinking ahead	.00	3.18	.00	1.67	.00	2.63
Elaborating to fill time	.81	6.25	.00	6.62	.00	6.67
Elaborating to clarify meaning	.00	10.00	.00	2.50	.00	5.26
Cognitive						
Attending	.00	14.29	5.22	15.44	4.36	17.69
Anticipating the content	.00	5.56	3.23	12.27	3.85	14.17

(Table continues)

Table (continued)

	Task Group A		Task Group B		Task Group C	
	Median	Range	Median	Range	Median	Range
Anticipating the structure	.00	7.14	4.92	13.22	2.51	11.54
Using imagery	.00	2.50	.00	7.27	.00	2.38
Using mechanical means to organize	6.90	25.00	12.71	17.14	12.86	38.54
Memorizing	.00	.00	.00	4.58	.00	4.17
Summarizing	.00	8.33	.00	4.85	.00	4.55
Translating	.00	10.71	.00	2.27	.00	.00
Inferencing	.00	.00	.00	4.35	.00	9.45
Processing inductively	.00	.00	.00	2.63	.00	.00
Metacognitive						
Setting goals	.00	16.67	2.94	10.27	4.01	18.72
Identifying the purpose of the task	.00	13.85	1.81	7.14	3.59	11.81
Planning	5.44	30.00	4.01	18.50	3.71	13.57
Monitoring	6.63	23.81	4.12	13.25	2.79	13.85
Self-correcting	.00	7.14	.00	4.69	.00	5.00
Evaluating previous performance	.00	9.09	.00	5.88	.00	9.79

(Table continues)

Table (continued)

	Task Group A		Task Group B		Task Group C	
	Median	Range	Median	Range	Median	Range
Evaluating the content of what was read/heard	.00	10.48	6.37	27.21	5.75	19.17
Evaluating performance	6.30	25.00	2.22	12.29	2.36	11.01
Evaluating language production	3.23	23.81	2.39	18.16	4.50	11.54
Affective						
Lowering anxiety	.00	7.14	.00	5.41	.00	8.63
Encouraging self	.00	7.14	.00	7.14	.00	10.00
Justifying performance	.00	7.69	1.70	5.88	.00	5.77

Appendix J

Descriptive Statistics for Individual Strategies by Task

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range
Approach												
Recalling task type	.00	5.26	.00	7.14	.00	8.33	.00	5.88	.00	8.33	.00	.00
Recalling the question	.00	20.00	.00	22.22	.00	7.14	.00	5.56	.00	12.50	.00	12.50
Recalling the text	.00	.00	.00	.00	.00	5.00	.00	6.90	.00	.00	.00	.00
Recalling the dialogue	.00	.00	.00	.00	3.70	18.75	.00	.00	4.46	11.76	.00	.00
Recalling the lecture	.00	.00	.00	.00	.00	.00	.00	11.76	.00	.00	.00	16.67
Generating choices	4.24	16.67	.00	10.00	.00	.00	.00	.00	.00	8.33	.00	.00
Making choices	9.09	16.67	.00	20.00	.00	6.67	.00	.00	.00	5.88	.00	.00
Developing reasons	3.85	21.43	2.78	28.57	.00	10.00	.00	10.34	.00	11.11	.00	8.33
Communication												
Simplifying the message	.00	7.14	.00	12.50	.00	8.33	.00	6.25	.00	8.33	.00	12.50
Avoiding	.00	14.29	.00	3.13	.00	3.13	.00	.00	.00	20.00	.00	12.50
Using Chinese	.00	6.67	.00	21.43	.00	4.55	.00	.00	.00	4.35	.00	6.67
Paraphrasing	.00	.00	.00	6.25	.00	7.14	.00	5.00	.00	11.11	.00	7.69
Approximating	.00	12.50	.00	7.69	.00	.00	.00	4.76	.00	10.53	.00	7.14

(Table continues)

Table (continued)

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range
Linking to prior experiences/knowledge	.00	20.00	.00	28.57	4.45	15.79	8.39	17.39	.00	18.75	6.90	25.00
Borrowing	.00	9.09	.00	10.00	.00	15.79	.00	5.88	.00	11.11	.00	8.70
Reviewing notes	.00	.00	.00	.00	.00	7.14	.00	7.69	.00	9.09	.00	10.00
Referring to notes	.00	16.67	.00	14.29	.00	10.00	.00	9.52	.00	9.09	4.01	10.00
Organizing thoughts	10.56	28.57	9.55	37.50	4.46	19.05	5.02	17.65	8.01	16.67	5.72	16.67
Guessing	.00	.00	.00	14.29	.00	5.88	.00	13.33	.00	13.33	.00	5.88
Repeating	.00	14.29	.00	20.00	.00	2.86	.00	7.69	.00	12.50	.00	6.25
Rehearsing	.00	7.14	.00	10.00	.00	5.88	.00	5.00	.00	5.26	.00	10.00
Reading ahead	.00	3.23	.00	20.00	3.13	10.00	.00	7.69	.00	4.35	.00	6.25
Restructuring	.00	7.69	.00	12.50	.00	3.23	.00	2.78	.00	9.09	.00	10.00
Slowing	.00	27.27	.00	14.29	.00	11.11	.00	4.35	.00	8.33	.00	7.14
Thinking ahead	.00	5.26	.00	3.13	.00	3.33	.00	.00	.00	5.26	.00	.00
Elaborating to fill time	.00	12.50	.00	7.14	.00	7.14	.00	7.69	.00	13.33	.00	.00
Elaborating to clarify meaning	.00	14.29	.00	10.00	.00	3.23	.00	5.00	.00	5.88	.00	10.53

(Table continues)

Table (continued)

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range
Cognitive												
Attending	.00	20.00	.00	28.57	6.07	25.00	5.21	18.75	6.07	25.00	4.88	20.00
Anticipating the content	.00	6.67	.00	11.11	1.56	20.00	4.45	13.64	.00	18.18	5.26	20.00
Anticipating the structure	.00	.00	.00	14.29	5.88	18.75	3.45	13.04	.00	15.38	.00	10.00
Using imagery	.00	3.23	.00	5.00	.00	10.00	.00	7.69	.00	4.76	.00	3.85
Using mechanical means to organize	5.96	33.33	6.07	33.33	13.96	28.33	12.77	23.08	12.50	37.50	13.96	50.00
Memorizing	.00	.00	.00	.00	.00	5.00	.00	7.69	.00	8.33	.00	5.88
Summarizing	.00	11.11	.00	16.67	.00	8.33	.00	8.00	.00	9.09	.00	6.25
Translating	.00	7.14	.00	21.43	.00	4.55	.00	.00	.00	.00	.00	.00
Inferencing	.00	.00	.00	.00	.00	8.70	.00	7.69	.00	9.09	.00	11.76
Processing inductively	.00	.00	.00	.00	.00	.00	.00	5.26	.00	.00	.00	.00

(Table continues)

Table (continued)

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range
Metacognitive												
Setting goals	.00	16.67	.00	33.33	3.87	20.00	3.24	14.29	.00	30.77	4.01	15.38
Identifying the purpose of the task	.00	14.29	.00	20.00	.00	7.14	.00	14.29	5.26	11.11	.00	12.50
Planning	6.51	60.00	4.67	36.36	5.56	25.00	4.65	14.29	5.41	20.00	.00	13.64
Monitoring	8.39	25.00	6.46	33.33	4.46	11.76	3.39	15.38	5.01	18.18	4.45	20.00
Self-correcting	.00	14.29	.00	10.00	.00	9.38	.00	7.69	.00	9.09	.00	10.00
Evaluating previous performance	.00	7.69	.00	18.18	.00	11.76	.00	11.76	.00	13.33	.00	9.09
Evaluating the content of what was read/heard	.00	14.29	.00	16.67	4.17	29.41	7.85	28.00	2.08	33.33	5.76	35.71
Evaluating performance	.00	18.18	5.90	40.00	.00	19.05	.00	16.00	.00	13.33	.00	15.38
Evaluating language production	.00	18.18	.00	33.33	.00	10.00	.00	30.77	5.57	15.79	4.45	15.38

(Table continues)

Table (continued)

	Task 1		Task 2		Task 3		Task 4		Task 5		Task 6	
	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range	Median	Range
Affective												
Lowering anxiety	.00	14.29	.00	.00	.00	6.67	.00	8.70	.00	12.50	.00	10.00
Encouraging self	.00	14.29	.00	.00	.00	14.29	.00	14.29	.00	9.09	.00	20.00
Justifying performance	.00	14.29	.00	15.38	.00	11.76	.00	6.67	.00	7.69	.00	8.33

Appendix K

Excerpts Illustrating Impact of Stimulated Recalls on Test-Takers' Behavior

Excerpt 13: 就是说，相当于，就这样说吧，当这个题型，后面这个题型和前面这个题型是一样的时候，就是有很大的帮助。(GJL) (Translation: that is … it's like … let me put it this way … it [stimulated recall] is very helpful when the next task's task type is the same as the previous one.)

Excerpt 14: … 就是，前面有一个是，就是我记 notes 嘛 [在纸上比划] 准备的时候，后来说的时候，其实根本都没用，当你问我的时候，我觉得根本都没用上，没有用。后来我做下面一道题的时候，我就只是看，只是脑子里，根本没有- 我没有<像上一题那样写>，所以觉得说的 时候好像好了一点点…。 (UVQ) (Translation: ... That is, I took notes [gesturing writing on paper] when I was preparing the previous task. Then I realized that [the notes] in fact were useless. When you asked me [during the stimulated recall], I felt that I did not use them at all. So when I did the next question, I just read, just thinking in my head. I didn't write notes <as I did in the previous task>. As a result, it went a bit better when I was responding …)

Excerpt 15: 有益只是说它第一，给了我一些时间，重新思考一下我当时做的那个过程，然后，我自己再想一遍以后，就会，可以发现一些问题，然后，在后面做题的时候，可以尽量避免。比如说，就像我刚才，发现这个（......）第二题嘛，我复述了它那个题目嘛，后来，我觉得这个是没有必要的。我在后面就 尽量不去这样。(GLW) (Translation: The benefits are, first, [it (stimulated recall)] gave me some time to reconsider the process of doing the task. After I thought it through, I discovered some problems. Then, when I did the next task, I was able to avoid them. For example, like earlier … I discovered … the second task … I repeated the question. Then I realized that it was not necessary. Then I tried my best to avoid doing that in the next task.)

Excerpt 16: 嗯，（…）对于我做，有点影响，就是说，我可能在，我在分析完我自己刚做 test 的行为之后，我就看到，知道，我的问题在哪儿，比如说，我的 notes 做的不足，然后，说话时候，很紧张，不流利，然后想的不够周到，就是因为我想到了，然后可能我在做下一个 test 的时候，就会把那些东西 correct 过来。按理说，是应该有帮助的。(UDZ) (Translation: um … for me, there is some effect. That is, I probably … After I analyzed my test-taking behaviors, I see … know my problematic areas, such as, not taking enough notes, then being very nervous when speaking, not fluent, didn't think through. Because I considered those, when I performed the next task, I would try to correct them. Reasonably speaking, it [stimulated recall] should be helpful.)

Excerpt 17: 我可能会, 因为一方面我说说我刚才的思维, 有点 help me, 就是帮我总结一下我刚刚的得与失那种感觉.所以说就像我前面说，哎呀我前面，说的时候我才意识到，我那个怎么组织上有问题,下面的时候我就想，如果时间有多,我就想，多写点东西，就可以有点益。对我来说，只有正面的 …。 (GSX) (Translation: I probably … because I talked about my thinking process, it helped me a bit. That is, it helped me to summarize on what I succeeded and failed. So, like what I said earlier, when I said it, I realized that I had an organization problem. The next time I would think that if there was enough time, I was thinking that I would write more, and that would be helpful. For me, there are only benefits …)

Test of English as a Foreign Language
PO Box 6155
Princeton, NJ 08541-6155
USA

To obtain more information about TOEFL programs and services, use one of the following:

Phone: 1-877-863-3546
(US, US Territories*, and Canada)

1-609-771-7100
(all other locations)

E-mail: toefl@ets.org
Web site: www.ets.org/toefl

*America Samoa, Guam, Puerto Rico, and US Virgin Islands