A Retrospective Study on Test-takers’ Cognitive and Metacognitive Processes in Taking a Compound Dictation Test

GU Xiangdong
Chongqing University

SHI Chunyan
Henan University of Urban Construction

Abstract

This study investigates the cognitive and metacognitive processes used by EFL test-takers in completing a compound dictation test through analyses of their verbal protocols obtained immediately after the test and through semi-structured retrospective interviews. The study explores relationships between the test-takers’ actual performance and the instructions for the compound dictation test, the language abilities measured, any major construct-irrelevant factors affecting the test-taking processes, and the performance patterns of performance across test-takers of different overall proficiency levels. Findings are as follows. (1) Test-takers’ actual performance does not apparently relate to the instructions given in the compound dictation test, which may raise doubts over the effectiveness of the instructions, thereby posing a potential threat to test validity. (2) Test-takers may pay more attention to the pronunciation of the words to be used to fill the blanks than to the actual meaning of the words, their difficulties in spelling and sentential expression possibly interfering with test performance even though they appear to use a variety of cognitive and metacognitive strategies throughout the test. (3) A number of construct-irrelevant factors posing a potential threat to test validity were found in the test-taking processes. Some of these factors are related to the test-takers, for example, memory capacity, attention failure and psychological factors, while others relate to the test, including speech rate and time constraints. (4) Test-takers in the study at all three proficiency levels appeared not to follow the instructions given in the compound dictation test. However, higher proficiency test-takers appeared to make more frequent use of cognitive and metacognitive strategies, and they appeared to be less influenced by construct-irrelevant factors.

Key words: College English Test Band 4; compound dictation; test-taking processes; retrospective approach; linguistic knowledge; strategy use; cognitive; metacognitive
1. Rationale of the study

Dictation is one of the most longstanding techniques for testing proficiency or progress in the learning of a foreign language and is still quite widely used in both language teaching and testing. However, the dictation test has been subject to sustained criticism (Roach, 1945; Stansfield, 1985), mainly on the grounds that it is not clear what, in terms of language learning skills or strategies, it is actually testing. Compound dictation is a variant of straight dictation involving three listenings to a text and the completion of blanks in the written version. Compound dictation was adopted as an alternative test format in the Listening Comprehension component of the College English Test (CET4 and CET6, the national English test at tertiary level in China). After the revision of the CET in 2006, compound dictation was employed as a compulsory element of the CET Listening Comprehension. The stated purpose of the Compound Dictation test is to “assess students’ listening ability at various levels (from vocabulary to passage)” (2006: 3). Research into the constructs the test actually measures and into how test-takers actually perform during the test would appear to be a matter of some importance.

2. Research questions

A retrospective approach was undertaken in this study to the investigation of test-takers’ cognitive and metacognitive processes when taking a CET compound dictation test in order to address the following four research questions:

1) Is test-takers’ actual performance consistent with the instructions given in the compound dictation test?
2) What does the CET compound dictation test actually measure?
3) Are there construct-irrelevant factors affecting test-taking processes?
4) Are there different patterns of compound dictation test-taking processes across candidates of different proficiency levels? If so, what are these differences?

3. Literature review

3.1 Listening comprehension

3.1.1 Listening ability constructs

Research indicates that second language listening comprehension is a complex, multidimensional cognitive process (Buck, 1991, 1992). Based on previous research, Buck (2001) proposes a comprehensive framework to describe listening ability constructs. The framework divides these into two general categories, i.e., language competence and strategic competence. Language competence consists of grammatical knowledge, discourse knowledge, pragmatic knowledge and sociolinguistic knowledge, while strategic competence includes cognitive and metacognitive strategies. Buck takes the basic division into language competence and strategic competence, as well as the sub-categories of...
language competence from Bachman and Palmer’s (1996) model of language ability. The description of strategic competence is based on the work of Purpura (1997, 1999), who suggests that strategic competence consists of both cognitive strategies, which are the conscious mental activities related to comprehending, storing and using linguistic knowledge, and metacognitive strategies, which consist of those conscious or unconscious mental activities that perform an executive function in the management of cognitive strategies.

3.1.2 CET Listening comprehension skills and requirements
The CET-4 testing syllabus (2006) states that Listening Comprehension should assess students’ ability to acquire oral information. The listening skills listed include:

- understanding the main idea
- grasping important points or relevant details
- judging the speaker’s opinion or attitude
- inferring implicit meaning
- judging the communicative functions of utterances
- recognizing phonetic features, for example by distinguishing sounds, understanding the significance of stress and intonation in continuous utterances
- understanding syntactic relationships: comparison, cause, result, degree, aim and so on.

The CET-4 Listening Comprehension requires test-takers to meet the basic College English Curriculum Requirements (2007: 24) as follows:

Students should be able to follow classroom instructions, everyday conversations, and lectures on general topics conducted in English. They should be able to understand English radio and TV programs spoken at a speed of about 130 to 150 words per minute (wpm), grasping the main ideas and key points. They are expected to be able to employ basic listening strategies to facilitate comprehension.

Based on listening ability constructs and CET listening comprehension skills requirements, we propose a framework for analyzing test-takers’ cognitive and metacognitive processes when taking the compound dictation test. This framework is both theory-based, i.e., taking account of language abilities as specified by Bachman and Palmer (1996); Buck (2001); Purpura (1997, 1999), and as empirically driven, i.e., through probing the consistency of relationships between test-takers’ performance and the instructions of the test, and probing construct-irrelevant factors according to the features of the test tasks and from the test-takers’ verbal protocols (see Figure 1).
3.2 Empirical studies on dictation tests
Dictation is traditionally defined as a technique used in language teaching and testing in which a passage is read aloud to students, with pauses during which students must try to write down what they have heard as accurately as possible (Richards, Platt & Weber, 1985). According to Taylor (1980: 88), dictation requires:

(i) reading a passage aloud, (ii) dividing the passage into phrases suitable for committal to short-term memory and re-reading phrases by phrases with gaps long enough for subjects to record the preceding phrase in writing, (iii) optionally re-reading each phrase as it was being written, and (iv) re-reading the whole passage as in (i).

Dictation has long been used in foreign language teaching and testing. Accordingly, various empirical studies have been conducted to investigate the validity of the dictation test. Oller (1971) studied the scores of 100 students who took the ESLPE (English as Second Language Placement Examination), which is composed of five parts: 1) vocabulary, 2) composition, 3) phonological discrimination, 4) grammar, 5) dictation. A correlation coefficient of .86 was found between the dictation score and the total score, which is only slightly less than a correlation of .88 between the composition score and the total score. Oller and Streiff (1975) presented a more complete description of the ESLPE format, re-evaluating Oller’s data (1971). They found that the correlations between dictation and other parts of the examination were significant at the .001 level. They conclude that dictation seems to be the best single measure of the totality of English-language skills tested. Oller and Streiff (1975) offered a comprehensive theoretical explanation to account for the use of dictation as a measure of language proficiency.
Pico (1990, cited from Wang, 2006) investigated the validity of the partial dictation test to measure foreign language listening comprehension. The findings suggest that short dictation tests can be constructed as useful measures of classroom second language proficiency. Kaga (1991) conducted two studies to investigate the reliability and validity of dictation for testing Japanese proficiency. Kaga concludes that “when some reasonable modifications are applied, dictation is an adequate and efficient measure of proficiency levels for learners of languages which have a good fit between pronunciation and orthography” (1991: 121). Ding (2001) investigated the reliability and validity of graduated dictation and selective dictation. He deduces that dictation is an integrative application of both receptive skills and productive skills. Factors including phonology, vocabulary, syntax, semantics, discourse, pragmatic rules and extra-linguistic knowledge affect test-takers’ successful comprehension. Deng and Zhang (2004) used a questionnaire survey to investigate students’ listening strategy use in dictation. The survey data show that dictation can accurately reflect learners’ overall language proficiency and can examine learners’ language knowledge and skills from various aspects.

While these studies seem to suggest that dictation measures the test-takers’ general language proficiency (including pronunciation, vocabulary, grammar, orthography, etc.), other studies indicate that dictation is a good indicator of test-takers’ listening ability.

In an attempt to introduce dictation as part of testing listening comprehension, Templeton (1977) conducted a listening cloze experiment by giving test-takers a transcript of a spoken text, with words deleted, then played a recording of the text and asked test-takers to fill in the blanks based on what they had heard. Templeton reports that the test results are highly correlated with those of other listening tests. Buck (1992) explored the technique of the listening summary which asked test-takers to fill in blanks on a summary of the listening passage, so as to prevent them filling in blanks without actually understanding the meaning. Buck found the technique reliable. It has good construct validity in that it is more closely related to other tests of listening than to tests of reading that use a similar format.

Oakeshott-Taylor (1977, cited from Buck, 2001) examined the errors test-takers made when taking dictation tests and observed that they were related to interpretation of the acoustic signal, phonemic identification, lexical recognition, morphology, syntactic analysis and semantic interpretation. Based on this, she argues that dictation tests assess performance at all stages of the “speech perception” process. Yoshida (1978, 1981) reports statistically significant positive correlations between dictation and the results of listening tests, and claims that dictation can be a good teaching device based on these findings. His findings indicate that dictation is a good predictor of learners’ listening ability.

Most of the research on dictation indicates that the dictation test is an effective language testing tool, but what it actually measures, whether it measures the general language proficiency or the ability of listening comprehension remains controversial. Besides, most of the validation studies of dictation tests discussed above adopt a statistical analysis approach based on test-takers’ performance (test scores) to reach the conclusion, which is product-oriented. The test-takers’ test-taking processes have been more or less neglected in research while it is possible that test-takers’ test scores may not directly
reflect their correct grasp of the given problems (Cohen, 1984; Grotjahn, 1986). As test-takers’ mental processes are complex, reliance on statistical techniques does not provide a complete picture of what the dictation test actually measures.

3.3 Studies on compound dictation
In China, compound dictation—a combination of spot dictation and dicto-comp (Wang, 2006), appeared first in the CET-4 in 1997 and two years later in the CET-6, as an alternative test format in Listening Comprehension. After the revision of the CET in 2006, compound dictation was employed as a compulsory test format in the part of Listening Comprehension.

However, research on the CET compound dictation is sparse. Among the few studies undertaken, some researchers have analyzed test-takers’ commonly made errors (Hu, 2008; Lin, 2007; Zhang, 2000), while others put forward suggestions on strategies or test-oriented techniques for taking compound dictation tests (Ma, 2002; Yang, 2007; Zhang, 2004). The few studies conducted appear to take researchers’ experiences as teachers as their focus and are replete with theoretical speculations.

3.4 Summary of the literature
A review of the research suggests a number of problems yet to be solved in the field of compound dictation validation studies:

1) Controversy over the construct validity of dictation test has never been settled;
2) Most previous studies on dictation used a statistical approach or product-oriented approach, which tells little about what is exactly measured in dictation tests;
3) Only a few studies are concerned about the compound dictation test and they are mainly based on researchers’ teaching experiences as teachers;
4) Empirical studies employing a retrospective approach to probe into compound dictation test are not available in the literature.

4. Methodology
This study employs a retrospective verbal report approach to examine the construct validity of the CET-4 compound dictation test. Retrospection is a data-collecting method in which participants are asked to report their mental processes some time after the mental events. Some studies (Alavi, 2005; Buck, 1991; Grotjahn, 1986; Wu, 1998) have used this method to investigate students’ test-taking processes. The studies attest to the reliability and validity of verbal protocol analysis (VPA). A number of observations concerning VPA have emerged from the literature:

1) The immediate retrospective verbal report procedure may produce incomplete verbalizations, so a retrospective interview following immediately after the verbal protocols retrospection is needed to provide additional information.
2) Selection of participants for retrospection is important. Given similar educational background and shared first language, participants should be motivated to talk as required,
and have little difficulty in expressing themselves. In addition, participants should be told clearly the purpose of the study and what they should do.

3) Before the formal data-collection procedure, rehearsal of the verbal report method is important.

4) Timing in research design is crucial for obtaining valid data. There should be a time limit for the test in order to imitate the circumstances of a real test. However, the time limit should not be set in the verbal report procedure as this may cause anxiety and impede participants’ thought processes.

5) Questions asked in the retrospective interview should consist mostly of “what” instead of “why” questions, which could help participants focus on the task and thoughts directly leading to their task performance.

4.1 Participants
Nine Chinese tertiary level non-English majors participated in the study. All of them had learned English for seven years, and shared a common background in education and culture. They were selected from 121 students enrolled in two English classes at Band 3 based on their scores on a listening pretest (which included a compound dictation). The participants were all taking the CET-4 at the end of the term. Pretest material was chosen from the Listening Comprehension part of CET-4 (December 2007), which was unfamiliar to them. The scores on the compound dictation tasks of the listening test were used as the main criteria for the selection of the participants with reference to their English teacher’s assessment of their listening comprehension ability. A range of levels (high, medium, low) was sought for comparing the test-taking processes of participants across different proficiency levels. Three participants at high proficiency level were selected from the students whose scores ranked in the top 20% of the total students; three at the intermediate proficiency level from the intermediate 20%; and three at the low proficiency level from the bottom 20%. The results of one-way ANOVA show that the dictation scores of participants at three levels were significantly different, with a post-hoc test indicating that high proficiency level participants’ dictation scores were significantly higher than those of the intermediate proficiency level and low proficiency level students ($p = .000 < 0.05$) and the scores for the intermediate proficiency level group were significantly higher than those for low proficiency level group ($p = .000 < 0.05$).

4.2 Instrument
The compound dictation used in the present study was chosen from the sample test paper of CET-4 (2006). It contains a reading passage of around 250 words, from which 8 words and 3 sentences have been deleted. The passage used in this study is a news report about Russia’s effort to join the WTO. The compound dictation requires the test-takers to fill in the original words for word blanks and the original words or main points in their own words for sentence blanks on the basis of their understanding of the passage.

4.3 Procedures
The data-elicitation procedures included five phases: the instructional phase, the pilot
phase, the immediate retrospection phase, the retrospective interview phase and the data transcription phase. The immediate retrospection phase constitutes the main phase of the study. The participants listened to the recording three times with some time for checking the answers as they would have in a live examination context. In this way, the test-taking time was controlled and limited by the recording. Following this, each participant reported immediately what he or she was thinking while performing the task, item by item. The participants were given prompts and encouraged to say whatever was in their mind when taking the test. There was no time limit for the retrospective report in order to create a non-threatening atmosphere, which was seen as essential in the experiment (Wu, 1998). The utterances were audio-recorded.

After the immediate retrospection phase, the participant answered the interviewers’ questions concerning his or her strategies of thoughts whilst performing the task. In order to meet the criterion of openness, the questions were semi-structured in advance and were mostly “what?” questions. For example: What did you do after the first hearing? What’s your understanding of the main idea of the listening passage? It was hoped that the retrospective interviews would enable the participants to recall more information relevant to the test-taking processes. This phase, which was intended to elicit additional information about the test-taking processes during the immediate retrospection phase, was also audio-recorded.

4.4 Coding reliability
Rather than coding the protocols in linguistically “small” clauses defined by the presence of one verb (e.g., Trabasso & Magliano, 1996), the protocols were parsed into thought units. Each unit contained a more or less complete idea, irrespective of the number of verbs present (Cote, Goldman & Saul 1998; Whitney & Budd, 1996). E.g., In one test-taker’s protocol, we coded the following as a unit: “For some words, if I could not understand them immediately when I heard them pronounced, I would write phonetic symbols beside them, and then try to figure them out at the next hearing.” This parsing conveys an almost complete idea of how the test-taker filled in the word blanks when taking the test.

To check the coding reliability, two researchers coded the verbal report data independently. The inter-rater agreement was encouragingly high (82%). Two weeks later, one of the researchers coded a portion of the data again, and the intra-rater agreement was calculated to be 87%, indicating that the coding of the verbal protocols was reliable. Agreement was reached between the two researchers for those anomalous items that were disputed.

4.5 Data analysis
Data analysis focused on the following aspects: 1) the extent of consistency of test-takers’ actual performance with the instructions of the compound dictation test; 2) the language abilities measured; 3) the major construct-irrelevant factors affecting test-taking processes; and 4) the performance patterns among test-takers of different proficiency levels in their test-taking processes.
5. Findings and discussions

5.1 Consistency between test-takers’ performance and the instructions

Table 1 sums up a comparison between what test-takers are instructed to do and what they actually did in this study for each listening event. The table reveals inconsistency between the performance and the instructions.

<table>
<thead>
<tr>
<th>Requirement of the CET-4 instructions</th>
<th>Test-takers’ actual performance</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>First hearing</strong></td>
<td></td>
</tr>
<tr>
<td>Listen carefully for its general idea</td>
<td>Mainly concentrated on the blanks instead of understanding the content of the passage, wrote the words or initial letters according to the pronunciation they heard; noted down some initial or ending words or some nouns of the missing sentences.</td>
</tr>
<tr>
<td><strong>Second hearing</strong></td>
<td></td>
</tr>
<tr>
<td>Fill in the blanks numbered from 36 to 43 with the exact words, for blanks 44 to 46, either use the exact words or write down the main points in your own words</td>
<td>Listened carefully for the words they didn’t complete, or checked and revised the words they dictated; mainly filled in the sentence blanks, tried best to note down what the sentence said, but couldn’t grasp all; paid attention to the preceding or following contexts using extra time, tried to understand the meaning.</td>
</tr>
<tr>
<td><strong>Third hearing</strong></td>
<td></td>
</tr>
<tr>
<td>Check what you have written</td>
<td>Word blanks almost filled; checked ambiguous items; continued to complement the missing sentences; made some patching; comprehended meaning of the words or text, and tried to seek clues for the unfinished sentences.</td>
</tr>
</tbody>
</table>

The contents in the column “test-takers’ actual performance” are summaries of the coded protocols. For the actual numbers of responses, evidence is provided in Table 4 below.

The analysis of the participants’ verbal protocols revealed a number of inconsistencies:

1) Five of the nine participants felt the speech rate of the recording was quite fast on first hearing. Therefore, they initially failed to understand the passage.

2) Four of them reported that the meaning of the passage was difficult to understand on first hearing.

3) All the participants felt more confident about their test performance if they noted down some words during first hearing.

4) All the participants were afraid of making mistakes when attempting to express meaning in their own words.

5) All the participants had difficulty in organizing the fragments they had heard into coherent sentences in their own words.

These five points help to explain why test-taker’s actual performance was inconsistent with the CET-4 instructions, which may raise doubts about the credibility of the instructions, thereby posing a potential threat to the validity of the compound dictation test.
5.2 Language abilities measured in the compound dictation test

5.2.1 Linguistic knowledge
This section presents the compound dictation’s measuring of linguistic knowledge at the word and sentence levels.

Pronunciation
According to the participants’ verbal protocols, most of them were concentrating on the pronunciation of the words for the blanks instead of their meaning. Some of them were unable to distinguish similar speech streams, such as “trip”, “trab”, “trap”; “effect”, “effort”, “officers”, “officials”, etc. (see Table 2).

Table 2. Examples of participants’ difficulty in distinguishing similar speech streams

<table>
<thead>
<tr>
<th>Key</th>
<th>These “greenhouse gases” trap heat in the atmosphere and are blamed for changing the world’s climate.</th>
</tr>
</thead>
<tbody>
<tr>
<td>H1</td>
<td>This greenhouse gases transport heat in the atmosphere and blame for changing the climate.</td>
</tr>
<tr>
<td>M1</td>
<td>This green house gases trip to hit the and for changing the world climate.</td>
</tr>
<tr>
<td>L1</td>
<td>These greenhouse gases trab hit the atmosphere that changing the world’s climate.</td>
</tr>
</tbody>
</table>

“H1” refers to a high proficiency participant. The same goes for the rest.

Generally, when the test-taker wrote a word correctly, it was because the word had been heard and its meaning was clear. When the test-taker heard the pronunciation of a word without knowing its meaning, the spelling of the word was almost always incorrect. Though most of the participants (7 out of 9) considered word pronunciation plays an important role in completing the task, when asked how they usually learned new words, only two of them (H1, H2) reported that they would remember words by their pronunciation; that is, they bear in mind how the word is pronounced and pay attention to both the phonetic rules and the word meaning.

Spelling
An analysis of the nine participants’ reports of test-taking processes, particularly when completing the word blanks, revealed that most of them heard the words clearly and correctly. However, as Table 3 shows, of the 72 words produced, 23 were misspelled and 6 were left blank, accounting for 40% of the words produced. These spelling mistakes and aborted attempts would not receive a score.

Table 3. Spelling mistakes made by the participants in word dictation

<table>
<thead>
<tr>
<th>Item No.</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1</td>
<td>√</td>
<td></td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H2</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>H3</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A Retrospective Study on Test-takers’ Cognitive and Metacognitive Processes in Taking a ...

<table>
<thead>
<tr>
<th>Item No.</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key</td>
<td>effort</td>
<td>officials</td>
<td>negotiate</td>
<td>balanced</td>
<td>competition</td>
<td>exchange</td>
<td>process</td>
<td>environmental</td>
<td></td>
</tr>
<tr>
<td>M1</td>
<td>effects</td>
<td>✓</td>
<td>meet</td>
<td>/</td>
<td>competition</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>M2</td>
<td>✓</td>
<td>✓</td>
<td>/</td>
<td>valenced</td>
<td>✓</td>
<td>✓</td>
<td>process</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>M3</td>
<td>✓</td>
<td>✓</td>
<td>discuss it</td>
<td>fast</td>
<td>competition</td>
<td>✓</td>
<td>✓</td>
<td>enviromental</td>
<td></td>
</tr>
<tr>
<td>L1</td>
<td>effort</td>
<td>✓</td>
<td>shaid</td>
<td>✓</td>
<td>competition</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td>✓</td>
<td>✓</td>
<td>/</td>
<td>✓</td>
<td>competition</td>
<td>✓</td>
<td>✓</td>
<td>enviromentally</td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td>✓</td>
<td>✓</td>
<td>/</td>
<td>voilence</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

✓correct response; / left blank.

Vocabulary
The protocols also revealed that the participants were able to complete the word blanks after hearing the recording a second time. Frequently used words could be directly recognized automatically, whereas the less frequent words were still parsed in syllables. They either incorrectly spelled the infrequent words or simply opted to skip them (see Table 4).

Table 4. Participants’ responses to word dictation on each hearing

<table>
<thead>
<tr>
<th>Item</th>
<th>36</th>
<th>37</th>
<th>38</th>
<th>39</th>
<th>40</th>
<th>41</th>
<th>42</th>
<th>43</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>First hearing</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>3</td>
<td>3</td>
<td>6</td>
<td>6</td>
<td>2</td>
<td>28</td>
</tr>
<tr>
<td>Second hearing</td>
<td>3</td>
<td>3</td>
<td>/</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>13</td>
</tr>
<tr>
<td>Third hearing</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>/</td>
<td>2</td>
</tr>
<tr>
<td>Total (correct)</td>
<td>7</td>
<td>6</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>43</td>
</tr>
<tr>
<td>Incorrect/blank</td>
<td>2</td>
<td>3</td>
<td>8</td>
<td>5</td>
<td>5</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>29</td>
</tr>
</tbody>
</table>

Table 4 shows that the words “exchange” (Item 41) and “process” (Item 42) were recorded accurately by two-thirds of the participants on first hearing. Other words required further hearings. Words such as “negotiate” (Item 38) were either misspelled or not attempted by most participants. Only one test-taker (H2) wrote the word “negotiate” correctly. This test-taker later reported that he used to listen to the Voice of America Programs (VOA) every day and was experienced with dictation exercises. As a high-frequency word in VOA, he might have often encountered “negotiate”.

Five participants considered the extent of their familiarity with such words as having a strong impact on their test performance. Some participants considered word length to be a factor in their successful reporting of the words. One participant (H2) suggested that words with several syllables such as “competition” proved less challenging than words with fewer syllables because shorter words give fewer clues to their identity on first hearing. By way of contrast, another participant (H3) was able to hear words with fewer syllables clearly.
**Sentential expression**

As expressed in their verbal protocols, the participants had difficulty in putting the fragments they had heard into complete sentences or expressing sentences in their own words (Table 5).

| Table 5. Sentential dictation: participants’ responses (excerpts) |
|---|---|---|---|
| Item | 44 | 45 | 46 |
| **H1** | This greenhouse gases transport heat in the atmosphere and blame for changing the climate but currently, nations producing only 44% of approve the protocol Russia produce 17% green gases | To join the WTO, a country make reach Kyoto Protocol that is also WTO members |
| **M1** | This green house gases trip to hit the and for changing the world climet But currently nations producing only 44% Russia produces about 17% of the world greenhouse gases | To join the WTO, a country must reach trade agreements with major trade countries and Russia will also make agreements with China India that WTO members |
| **L1** | These greenhouse gases trab hit the atmosphere that changing the world’s climate But currently, nation’s producers about 44% Russia produces about 19% the greenhouse gases | To join the WTO, a country must trade the agreement that also WTO members |

Most participants considered Item 46 easiest to understand and write, as words in the sentence were simple and familiar to them. Items 44 and 45 were difficult. Most participants did not hear the important phrase “trap heat” in Item 44 nor the term “approve the protocol” in Item 45. These words were not difficult to understand, but most of the participants reported that they had not seen these phrases used in such a way.

Test performance shows that test-takers at different language proficiency levels do not exhibit large differences in word dictation tasks, but differ in sentence dictation tasks (see Table 6). The high- and middle-level test-takers performed better than low-level test-takers in sentential dictation.

| Table 6. Participants’ scores of the word and sentence dictation tasks |
|---|---|---|---|---|---|---|---|---|---|
| | H1 | H2 | H3 | M1 | M2 | M3 | L1 | L2 | L3 |
| Word score | 2.5 | 3 | 2 | 2 | 2.5 | 2 | 2.5 | 2 | 3 |
| Sentence score | 4.5 | 3 | 3.5 | 4 | 4 | 3 | 2.5 | 2 | 1.5 |

**Semantics**

Besides recognition of vocabulary, the participants occasionally related the meaning of a word to the sentential context in order to confirm the accuracy of their dictation. For example,

(H2): When I heard “officials” I wrote it down, and I saw “Russian” in this sentence then I felt certain it wouldn’t be wrong, because it was coherent in this sentence.
Reports revealed that the test-takers tended to consider more semantic meaning when completing sentence blanks. For example,

(L2): When I was listening for word blanks, I mainly wrote based on its pronunciation and basically didn’t think of its meaning. For the missing sentences, I would draw its meaning from what I heard. You hear it and you know its meaning, then you can write it down, not necessarily the original sentence.

In general, the participants attempted to understand what a sentence or the text was attempting to convey when they encountered listening problems, that is, when they could not write down what they heard or when they were confronted with an unfamiliar word in a sentence. Determining the semantic meaning appeared to help facilitate sentential comprehension. For example,

(M1): The second time after the sentence was read, there was a long time interval to think and write, I could add something better than I did the first time. But to a later time, I probably wrote more than half of the sentence, I found the latter part of it I had heard was forgotten or very confusing, and then I tried to look at the following words in the passage. For example, I seemed to have heard this word kind of like “comments”, and it may be collocated with the “reach”, and then I saw there was “agreements” beneath, I just wrote it down directly, I thought it should be that.

In general, the participants of higher proficiency demonstrated more frequent use of semantic meaning than low level participants as revealed in their protocols. Participants of high-level and middle-level employed contextual clues ten and eight times respectively, which low proficiency participants used only five times. This may account for better performance of higher-level participants in their sentence dictation tasks.

5.2.2 Strategy use
Strategies are deliberate and represent cognitive steps that the learners can take to assist in acquiring, storing, and retrieving information (Paris, 1983, cited from Anderson, 1991). Because memory, especially short-time memory, has limited capacity, it is necessary to depend on various strategies to deal with a large amount of information. Many strategies are conducive to processing a piece of information more deeply and more elaborately.

Nine strategies were identified from the retrospective protocols of the participants. These strategies can be further categorized as cognitive strategies (i.e., inferencing, prediction, contextualization, reconstruction, and translation) and metacognitive strategies (i.e., pre-listening preparation, selective attention, comprehension monitoring, comprehension evaluation) (Purpura, 1997). They are presented in Table 7 with corresponding descriptions and examples listed.
Table 7. Strategies employed by participants when taking the compound dictation test

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Description</th>
<th>Protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferencing</td>
<td>Infer or guess missing information or meaning of words or sentences</td>
<td>(1) When listening to the sentence, I heard some words roughly, then I inferred the meaning using some of the nouns in the sentence. (H2)</td>
</tr>
<tr>
<td>Prediction</td>
<td>Anticipate general contents or details before and during listening</td>
<td>(2) Before the dictation started, I saw the first two sentences and I knew this was a news item, and I got a rough idea about its contents. (L3)</td>
</tr>
<tr>
<td>Contextualization</td>
<td>Relate new information to a wider familiar context</td>
<td>(3) I used the context of this utterance, because I heard “the others only…”, following “Russia has not yet approved it” in the previous segment, so I thought the others’ percentage should be under 44% …; and there would also be some kind of a transition before “Russia”. (H3)</td>
</tr>
<tr>
<td>Reconstruction</td>
<td>Recreate meaning using words having been heard</td>
<td>(4) At that time, I only got a rough understanding that currently the production rate of some countries was only 44% and that they had approved the protocol, but then I heard “17%” or I thought I heard “17%”, so I felt the meaning was not right after I wrote down, “but Russia produced 17%…”; there was a transitional word “but”. (H1)</td>
</tr>
<tr>
<td>Translation</td>
<td>Change listening input into L1 before interpretation</td>
<td>(5) The first time I heard “environment”, and I didn’t understand, as it should be an adjective here. And at that time I suddenly forgot how to spell it, so then I noted down the meaning in Chinese. The second time I heard it clearly and I made sure of it. (M1)</td>
</tr>
<tr>
<td>Pre-listening preparation</td>
<td>Preview contents in different forms; establish purpose for listening</td>
<td>(6) Before listening, I only had time to look at the first paragraphs briefly. (L3)</td>
</tr>
<tr>
<td>Selective attention</td>
<td>Notice specific aspects of input; listen selectively according to purpose</td>
<td>(7) I followed the recording, then I looked for where the blank was in the text and focused carefully on that place. (L1)</td>
</tr>
<tr>
<td>Comprehension monitoring</td>
<td>Monitor understanding while listening</td>
<td>(8) I heard the word “official” the second time, but I couldn’t react to it immediately, I thought I’d just leave it and continue to listen; the third time I noted down the phonetic symbols … (H2)</td>
</tr>
<tr>
<td>Comprehension evaluation</td>
<td>Check understanding for accuracy, completeness and acceptability</td>
<td>(9) The first time I heard “balanced”, but the second time I didn’t hear the syllable “d”, in it. Then I looked at its context, analyzed its syntax, and finally I felt there should be a “d”. As for the word “negotiate”, I thought about how to spell it, but it seemed I still wrote it wrongly. For other words, I didn’t do much revision, just checked whether there was a grammar mistake or not. (H1)</td>
</tr>
</tbody>
</table>

Use of cognitive strategies
Cognitive strategies are used to process utterances directly by transforming them into mental presentations that can be stored and recalled (Goh, 2002). They are essential to successful comprehension by directly manipulating and transforming the input information. Table 8 shows the frequency of the use of cognitive strategies by the participants, and it is clear that the higher the participants’ proficiency, the more
frequently they used cognitive strategies.

**Table 8. Frequency use of cognitive strategies**

<table>
<thead>
<tr>
<th>Strategy</th>
<th>High</th>
<th>Middle</th>
<th>Low</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Inferencing</td>
<td>9</td>
<td>7</td>
<td>7</td>
<td>23</td>
</tr>
<tr>
<td>Contextualization</td>
<td>6</td>
<td>4</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Reconstruction</td>
<td>4</td>
<td>3</td>
<td>1</td>
<td>8</td>
</tr>
<tr>
<td>Prediction</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>Translation</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>22</td>
<td>17</td>
<td>10</td>
<td>49</td>
</tr>
</tbody>
</table>

Of the five strategies, the inferencing strategy was the most frequently used one. Inferencing helps test-takers to infer the missing information or meaning of words or sentences. Zheng and Gu’s study (2008) provides evidence for the claim that constructing an interpretation of a listening text is largely an inferential process in the light of the actual input and test-takers’ background knowledge.

Contextualization is the second most frequently used cognitive strategy. When the participants applied the contextualization strategy, they were less preoccupied with getting the exact words. Instead, they were more interested in constructing the big picture in terms of cohesion within the text (see Protocol 3 in Table 7). By reconstruction strategy, the test-takers could reconstruct the meaning using the words they had heard. Although generally helpful, it was not useful when the test-takers did not hear or note down a sufficient number of key words for recreating the original message correctly as we can see from Protocol 4 in Table 7. The use of prediction strategy enabled the participants to anticipate content before and during listening, and this helped them process the input more quickly (see Protocol 2 in Table 7). Translation occurred only three times. Translation enabled the participants to better remember the words or understand their meaning. However, translation actually slowed down processing and often diverted the test-takers’ attention from the clues that might have assisted their comprehension (Zheng & Gu, 2008). Therefore, few participants translated what they had heard into Chinese during the test.

**Use of metacognitive strategies**

The term “metacognition” refers to a person’s understanding of any cognitive process (Carrell, 1989). Metacognitive strategies are used to manage complex cognitive processes before, during and after processing the information (Goh, 2002). They help learners regulate their own cognition and facilitate the effective linking of strategies, thus promoting efficient cognitive processing of the input information (Thompson & Rubin, 1996). Test-takers become aware of the processes underlying their own listening and begin consciously to oversee and regulate these processes by planning approaches to various tasks. In this study, higher proficiency test-takers tended to use metacognitive strategies more than less proficient test-takers (see Table 9).
Table 9. Frequency use of metacognitive strategies

<table>
<thead>
<tr>
<th></th>
<th>Comprehension evaluation</th>
<th>Comprehension monitoring</th>
<th>Selective attention</th>
<th>Pre-listening preparation</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>16</td>
<td>7</td>
<td>5</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>Middle</td>
<td>13</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>27</td>
</tr>
<tr>
<td>Low</td>
<td>8</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>21</td>
</tr>
<tr>
<td>Total</td>
<td>37</td>
<td>21</td>
<td>15</td>
<td>5</td>
<td>78</td>
</tr>
</tbody>
</table>

Comprehension evaluation is the most frequently employed metacognitive strategy by the participants. Many participants were aware of the effectiveness of their performance. They tended to evaluate the accuracy or acceptability of what they wrote down, as shown in Protocol 9 in Table 7. This would lead to a decision about the need (or otherwise) for further revision. Comprehension monitoring is the second most frequently used strategy. Comprehension evaluation and comprehension monitoring are clearly distinguished. Comprehension monitoring refers to the strategy that test-takers use when encountering listening difficulties, assessing the importance of problematic parts and deciding whether to ignore them and continue to listen in spite of difficulty, or actively seeking clarification, or determining the value of subsequent parts, and varying the intensity of attention accordingly (Goh, 2000), as shown in Protocol 8 in Table 7.

The participants reported that they would focus on the specific parts of the text or listen selectively according to purpose. Selective attention occurred 15 times in the test-taking processes. For the compound dictation test, most test-takers thought their task was to fill in the blanks and, as a consequence, they were only concerned about the content in the blanks. Where there was a blank, the test-takers would direct their attention to writing what they heard. In allocating attention selectively, the participants were not distracted by small problematic parts as they maintained information processing (Protocol 7 in Table 7).

Pre-listening preparation strategy refers to the strategy of mental preparation for a listening task, including previewing contents in different forms and establishing a purpose for listening. Such planning as scanning the content before listening, though the test-takers were unable to read the content completely, enabled them to anticipate the general idea of the passage, thereby engendering confidence during the test task, as some participants reported. Establishing a purpose for listening enabled the test-takers to focus their attention on the important parts of the text and perform the task effectively.

There is a positive relationship between strategy use and listening performance. It has been found that students who use a greater range of strategies are able to make more progress in listening comprehension (Bacon, 1992).
A one-way ANOVA analysis with a post-hoc test revealed that scores of the high-level participants were significantly higher than those of the low-level participants ($p = .026 < 0.05$). Figure 2 shows that the high-level participants display more use of strategies than the middle- and low-level participants, indicating that a positive relationship exists between test-takers’ proficiency levels and the frequency of strategy use. Test-takers show a clear preference for metacognitive strategies. Their frequent use of metacognitive strategies may be ascribed to the effects of the test method. Since compound dictation tests constantly involve filling in the blanks, test-takers may consciously direct their concentration towards the missing information instead of listening for the general idea, and they themselves are aware of the problems of insufficient comprehension during the test-taking processes. Metacognitive strategies, such as comprehension evaluation and comprehension monitoring, are frequently invoked when test-takers solve the items and fill in the blanks. In the meantime, cognitive strategies, as Goh (2000) suggests, act directly on the input to make sense of it. Examination of the participants’ protocols indicates that, though test-takers employ cognitive strategies such as inferencing and contextualization to facilitate their understanding and filling out the blanks, they do not use them enough to make sense of the message in the text.

5.3 Construct-irrelevant factors affecting test-taking processes
Examination of the protocols reveals that there are a number of factors which play an important role in the test-taking processes that are not directly related to test-takers’ language ability (see Table 10).

<table>
<thead>
<tr>
<th>Factors</th>
<th>Protocols</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory capacity</td>
<td>(1) First time, I clearly heard the second part of the sentence, but I couldn’t write it down immediately. I felt I caught most of the words but when I tried to write them down, I quickly forgot. (M3)</td>
</tr>
</tbody>
</table>

Figure 2. The number of strategies employed according to proficiency

Table 10. Construct-irrelevant factors affecting test-taking processes
All the participants indicated that memory capacity and attention influenced their test-taking efficiency. Speech rate, time factor and mental state were also influential as indicated by most participants (see Table 11). Unlike strategy use, the higher the proficiency levels of the test-takers, the less affected they were by construct-irrelevant factors.

Table 11. Frequency of each construct-irrelevant factor per different proficiency level

<table>
<thead>
<tr>
<th></th>
<th>Memory capacity</th>
<th>Attention failure</th>
<th>Speech rate</th>
<th>Time constraints</th>
<th>Psychological state</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>High</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>/</td>
<td>4</td>
<td>17</td>
</tr>
<tr>
<td>Middle</td>
<td>5</td>
<td>7</td>
<td>5</td>
<td>4</td>
<td>2</td>
<td>23</td>
</tr>
<tr>
<td>Low</td>
<td>8</td>
<td>5</td>
<td>6</td>
<td>5</td>
<td>1</td>
<td>25</td>
</tr>
<tr>
<td>Total</td>
<td>18</td>
<td>16</td>
<td>15</td>
<td>9</td>
<td>7</td>
<td>65</td>
</tr>
</tbody>
</table>

Among the construct-irrelevant factors, some are related to test-takers (memory capacity, attention failure and psychological state) and others are related to the test itself (speech rate and time constraint). Memory capacity is the most significant construct-irrelevant factor. The cognitive load placed on the test-takers’ memory was high and interfered with their “true” listening ability, thereby masking any inferences about how well they could listen.

6. Conclusions

This study has employed an approach that combines two elicitation techniques, immediate retrospection and retrospective interviews, in investigating how nine EFL test-takers took the CET-4 compound dictation test. With reference to the four research questions, the following conclusions can be drawn:

(1) The test-takers’ actual performance was not consistent with the instructions given in the compound dictation test, which may raise doubts about the credibility of the instructions thereby posing a potential threat to test validity.
(2) The test-takers appear to pay too much attention to the pronunciation of the words to be filled in rather than their meaning in the context. Difficulties in spelling and sentential expression interfered severely with test performance though the participants used a variety of cognitive and metacognitive strategies in their test-taking processes.

(3) A number of construct-irrelevant factors were identified in the test-taking processes. Some are related to the test-takers (psychological state, attention failure, and memory capacity) while others are related to the test itself (speech rate and time constraint). Construct-irrelevant factors also pose a potential threat to test validity.

(4) The test-takers at all three proficiency levels failed to follow the instructions given in the compound dictation test. However, the high proficiency test-takers made more frequent use of cognitive and metacognitive strategies and they appeared to be less influenced by construct-irrelevant factors than lower proficiency test-takers.

The generalizability of these conclusions should be treated with caution, especially given the limited number of the participants employed in the study. However, the study remains important for the following reasons: 1) it has proposed an analytical framework for the construct validation of the compound dictation test; 2) the research findings have practical implications for language testers and provide advice on how to reduce potential threats to construct validity in test development and test evaluation, and for language teachers and learners on how to improve language proficiency; and 3) the retrospective approach has proved an effective methodology for exploring issues of construct validity.

Acknowledgements

This paper is part of the achievements funded by the Fundamental Research Funds for the Central Universities in China (No. 0205005201030) and by the National Scholarship Council of China (No. 2010850066) granted to the first author to support a one-year visiting professorship and period of collaboration with University of Cambridge ESOL Examinations from October 2011 to September 2012.

The authors would also like to thank Dr Stuart Shaw of Cambridge International Examinations, Dr Roger Hawkey, consultant to Cambridge ESOL and the anonymous reviewers of the *Chinese Journal of Applied Linguistics* for their constructive feedback for the revision of the paper.

References


and Local Examiners.

*(Copy editing: Duncan SIDWELL)*