FROM THE PERSPECTIVE OF TEST-TAKERS: THE IMPACT OF TWO MAJOR THREATS TO CONSTRUCT VALIDITY ON TEST PERFORMANCE

Chun Hua Mei
(957992897@qq.com)
Wuhan University, China

Abstract

Little research has been done on how to specifically detect Messick’s (1996) proposal of the two major potential threats to construct validity, namely, construct under-representation and construct irrelevant variance in a test. In order to bridge this gap, a qualitative study was conducted to examine and analyze the college freshmen’s accounts of their responses to invalid items when they were finishing the multiple-choice items in listening section of the College English Test Band 4 (CET-4). Those freshmen were chosen in Wuhan University according to their English scores in College Entrance Examination and they all already achieved the high level of English proficiency in high schools. The participants were required to think aloud their thought processes in taking the test items and talk about their experiences after finishing the items. All their accounts were recorded and analyzed to detect the trace of the two major threats to construct validity. Thus, this study aims to explore a way to investigate the test-takers’ responses to the potentially invalid test items with the threats of construct under-representation and construct irrelevant variance and the impact of the two major threats on their test performance.

1 Introduction

Test validation is widely deemed to be empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied settings that might erode and promote the validity of score interpretation and use (Messick, 1996). Therefore, the examination of validity has been traditionally focused on the collection and discussion of the types of evidence that support a particular meaning or use of test scores. Researchers tend to conduct large-scale quantitative studies to involve great range of samples so that they can gather enough factual evidence to support a given interpretation or use. However, from Carroll’s (1968) definition of a test, “a psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual”, and Messick (1989, p. 55) argues that a test’s construct interpretation might need to vary from one type of person to another ... and is a major mystery in educational and psychological measurement. Then, we are clear that a test is a process in which different groups of individuals with various characteristics and attributes interact with the test’s content and test method facets. Therefore, speaking of a given test’s validity, in addition to the test’s content and test method, test validation must consider how test-takers perform as well (Bachman, 1990), for the performance we observe and measure in a language test is actually a sample of an individual’s total
performance in that language tested, so the testees’ responses and thinking modes hidden in test-taking process contain much more information that can hardly be detected in the investigation of a given interpretation or use of test scores. Thus, it is meaningful for us to conduct a qualitative study to investigate test items from the perspective of test-takers by examining and analyzing the testees’ accounts of their responses to invalid items when they are completing the test. The College English Test Band 4 (CET-4) is one of the most important English proficiency tests which aim, on one hand, to evaluate the undergraduates’ mastery of the required English course content in colleges and universities in China and, on the other hand, to examine the students’ actual and practical English capability concerned with listening, reading, writing and translating, and meanwhile, little attempt has been made to substantiate Messick’s (1996) proposal of two main threats to construct validity, that is, construct under-representation and construct irrelevant variance, in the validation research of language tests. Thereafter, in this study, we make a tentative effort to investigate test candidates’ reactions to the two major threats to construct validity in CET-4 as well as the impact of the two threats on test-takers’ actual test performance during the test-taking process and set out to explore a qualitative way to examine the construct of test items from the perspective of test candidates.

1.1 Construct validity

According to Bachman (1990, p. 254), construct validity concerns the extent to which performance on test is consistent with predictions that we make on the basis of a theory of abilities or constructs, and Embretson (1983) claims that construct validation should involve two separate stages: construct representation, which concerns the identification of the theoretical constructs underlying performance on a test’s items; and nomothetic span, which relates to the ability of a test to measure individual differences. We get to know that the first step we should follow in construct validation is to specify the constructs and then we need examine the correlation between construct definitions and predictions made from test performance. As for construct definition, Stenner, Smith and Burdick (1983) defined it as the process whereby the meaning of a construct ... is specified, in which we impart meaning to a construct by testing hypotheses suggested by construct theories. From this definition, it can be concluded that the meaning of a construct is concealed in the hypotheses in a given test’s construct theory. Therefore, the examination of construct validity should first clarify the construct of the test we aim to examine.

The chosen material in this study is the listening subtest of CET-4 in December, 2013. Based on the constructs of required skills in the CET-4 listening subtest summarized by Cao (2010) and the latest CET-4 syllabus issued by the CET committee in 2013, the listening comprehension subtest examines the test-takers’ ability to obtain verbal message and oral information, and the specific skills tested in listening comprehension are outlined:

A. Understanding the main ideas and essential details

1) Understanding the main ideas
2) Grasping the essential and specific details
3) Judging the talkers’ opinions and attitudes
B. Understanding the connotative meaning

4) Deducing the connotative meaning
5) Estimating the communicative functions of the utterances

C. Understanding the listening materials by referring to the language features

6) Discerning phonetic features such as pronunciation, stress and intonation in a connected speech
7) Apprehending sentential relationships, such as comparison, reason and result, degree and extent, range and scope, intention and purpose etc.

This specification of skills which the listening subtest intends to examine is used as a basis of our later analysis on the data collected. As this study aims to detect the subtle trace of the main two threats to construct validity in test candidates’ verbal reports on their thought processes in test-taking, it is necessary to learn the concept of construct under-representation and construct irrelevant variance.

1.2 Two threats to construct validity

According to Messick (1996), the validity of a test should be conscious of two main threats to construct validity. The first type is what he calls construct under-representation caused by an inadequate or incomplete sampling of the construct, whereby the test items in the assessment fail to include important dimensions or facets of the construct and therefore are unlikely to reveal what the construct is supposed to measure on test-takers’ true abilities. The other type is what he names construct irrelevant variance, that is, the measurement of factors that are simply not relevant to the construct, probably resulting in construct-irrelevant easiness or construct-irrelevant difficulty in the test. Davis et al. (2002, p. 221) believe that the validity of a language test is established by the extent to which it succeeds in providing an accurate and concrete representation of an abstract concept. The inappropriate representation and the inaccurate involvement of irrelevant factors substantially threaten the validity of a given test’s construct and therefore, should be cautioned and avoided at the beginning.

In Xi’s (2010) investigation of test fairness, it is noted that construct irrelevant variance and construct under-representation may have systematic and appreciable effects on test scores and test interpretation as a threat to test fairness. The hazard of these two threats to construct validity is widely accepted and many researchers (e.g. Stansfield, 2011; Wagner, 2010; Xi, 2010, 2005; Gebril, 2009; Ockey, 2007; Elder, 2001) have mentioned a few words about this concept of construct irrelevant variance and construct under-representation, but most of them just offer a few insightful clues but without any further interpretation. For instance, Xi’s studies on performance on line graph description task (2010) and visual chunks and planning’s impact on oral graph description (2005) have found that task characteristics introduce variance that is considered irrelevant to the construct, and task design principles should be followed to minimize potential construct-irrelevant variance. In his both studies, the introduction of another specific task trait is
in want of caution of potential construct-irrelevant variance but no need to detect what variance may threaten construct validity. Thereafter, we still have no idea how to probe the two threats of construct validity as well as the test-takers’ responses to the two major threats to construct validity hidden in test items. Only a few studies have been made on the detection of construct irrelevant variance (e.g. Dávid, 2007) but it is never enough for a further comprehensive understanding of the cause, performance and impact of this threat to construct validity.

However, the other major threat to construct validity, construct under-representation, seldom appears in the researchers’ vision except several countable references to this concept without further details and explanations (Stansfield, 2011; Wagner, 2010; Chen, 2009; Gebril, 2009; Ockey, 2007; East, 2007; Elder, 2001). In Wagner’s (2010) research on the effect of the use of video texts on ESL listening test-taker performance, he asserts that it is necessary to investigate the extent to which the involvement of a new task characteristic introduces construct relevant or construct irrelevant variance into a given listening test and also suggests that the specification of the construct of listening tests should not only exclude the threat of construct irrelevant variance but also the threat of construct under-representation, depending on the TLU domain/context with which the listening test is associated. East (2007) has conducted a comparative within-subjects study to examine whether there are significant threats to the construct validity of the writing test with access to dictionaries based on the examination of test scores and the explanation of the scores meanings according to test construct. However, among these studies, no further information but the research significance and caution against them is expressed on Messick’s (1996) proposal of the two major threats to construct validity. Thus, in order to bridge this gap, we set out to detect the subtle trace of construct under-representation and construct irrelevant variance by conducting a qualitative study on testees’ responses or reactions to a CET-4 listening subtest.

1.3 Test-takers’ responses or reactions

Based on Stenner, Smith and Burdick’s (1983) construct definition, we know that the validity of any given construct theory is a matter of degree, partially depending on how well it predicts variation in item scale values and personal scores and the range of these predictions. Therefore, the validation of a construct theory is, in part, dependent on the design of a test and the writing of test items, the consistency of rating and score interpretation and the variation of the test contexts. It can be explained that researchers are devoted to validation from the investigation of test scores and their meanings and the logical analysis of intended variances and factors in test design. Anderson, Bachman, Perkins and Cohen (1991) have concluded that traditional approaches to construct validation have focused on the content of the tests themselves and the products—test items or test scores and the relationships among these, largely ignoring the processes that test-takers employ in taking tests.

However, Shohamy’s (1983, 1984) findings in the investigation of the stability of oral proficiency assessment and the impact of test method on reading comprehension confirm that the product-oriented research methodology falls short of elaborating how the test methods produce an effect on test performance and therefore, contributes little to test design and the interpretation of test scores. Bachman (1990, p. 271) claims that, two of these types of evidence, correlational
and experimental, are particularly powerful in validation, but what he believes a third way, that is, the analysis of the process of test-taking, holds perhaps the most promise for providing new sights into the factors that affect test performance. Furthermore, Dávid (2007) has evaluated the research of his earlier work and summarized that those psychometric analyses based on quantitative analytical method like the Rash Model cannot reveal the mental processes behind the answers or either, distinguish between construct-relevant and construct-irrelevant sources of difficulty because the scores investigated are only the results of those mental processes, so he argues that the think-aloud protocols may throw some light upon how students make choices between the options and the notion of construct validation must involve probes into how candidates actually respond to a test. Bax (2013) has conducted a research on the cognitive process of the test candidates by gathering evidence from eye-tracking to evaluate the cognitive validity of reading test items (Glaser, 1991; Field, 2012) and verified that test-takers’ responses to test items can be taken as a type of evidence to evaluate a test’s construct validity.

Therefore, we are aware that a qualitative study on testees’ verbal reports of their thought processes in test-taking can provide much more information that can hardly be detected in the investigation of a given interpretation or use of test scores. Alderson (2000, 2011) argues that what matters in test construct validation is not what the test constructors believe an item to be testing, but which responses the test-takers consider correct, and what thought processes underlie their choices. In other words, understanding test traits or items being measured requires our digging into the cognitive processing beneath test candidates’ choices of their assumed right answers. Then, it is inferred that the investigation of test-taking thought processes from the perspective of test-takers may offer us a new meaningful understanding of construct validity, especially the detection of construct under-representation and construct irrelevant variance. This study, thus, proceeds as a tentative trial to locate the cues of the two major threats to construct validity and to figure out what impacts the two threats have on their test performance through probes into the test candidates’ responses to test items and their thought processes from the perspective of test-takers.

2 Research questions

In this study, we aim to investigate the test-takers’ responses to the invalid test items with the threats of construct under-representation and construct irrelevant variance and the impact of the two main threats on the testees’ performance. Thereafter, this study is designed to answer the following questions:

1. What responses do the test-takers hold to the invalid items? What differences are between test-takers’ responses to valid items and their responses to invalid items?

2. What impact do the testees’ reactions to the two main threats to construct validity reveal on their test performance?

3 Research method

Verbal reports or think-aloud protocols have been widely used to verbalize testees’ thought
processes in many studies. Buck (1994) has used verbal protocols with a structured interview to clarify the hypothesis that L2 listening test will not meet the assumption of unidimensionality made by norm-referenced measurement model and also urged the future researchers to replicate his research in other subjects and tests. Dávid (2007) has taken advantage of the participants’ verbal reports on their completion of a 45-item grammar test to investigate the response validity. Kong (2011) is one of the few researchers in China who attempt to examine the construct validity of test items or tasks from the test-taking process or the test-takers’ responding process, but he has actually made a quantitative study on the test-takers’ think-aloud protocols and still fails to avoid the common weaknesses of quantitative studies, that is, the failure to detect subtle changes in the examinees’ thought processes. Therefore, we set out to conduct a small-scale qualitative study on test-takers’ thought processes during their responding to the test items by digging out much hidden information in their think-aloud protocols.

3.1 Participants

The participants selected in this study were 11 freshmen in Wuhan University, among whom nine students were female and the rest two were male. According to the information collected by questionnaires, the participants were chosen by their performances in the College Entrance Examination (they all scored above 130 with the total score of 150) as well as their regular performance in high schools. The participants are found to have already achieved a high English proficiency in high schools and all volunteered to take part into this research. They had never taken or practiced any test of College English Test Band 4 (CET-4) so that the interference of test preparation or test familiarity to CET-4 can be avoided in their taking the test items.

3.2 Test material

The listening subtest sample in this study was taken from the first set of the standard CET-4 in December, 2013. The selected sample test items lasted 24 minutes in total and only the previous 25 MC items in this sample were tested in order to avoid subjectivity and uncertainty in judging the right answers. In addition, the students could spare more time and were more inclined to verbalize their test-taking reasons and experiences by taking the MC items. The 25 items consisted of 8 short conversations, two long conversations and 3 passages, and the questions were separated from the options on paper together with the listening material audio.

3.3 Procedure

The participants were required to think aloud their thought processes and record the accounts of their responses to each item. The play of the listening material audio stopped at each question interval to leave the students enough time to finish their think-aloud protocols. I made an exemplary verbal report to show each student how to verbalize their thoughts. In order to make their verbal reporting authentic and smooth, Chinese with code-switching was the preferential reporting language and their verbal protocols were later translated into English.

Before the formal test, two pilot studies were respectively carried out on a male and a female student. According to their performance, I found that it was inappropriate to encourage the
students who were not good at expression to talk more about their item-taking thoughts in their verbal reporting, for they might get nervous and their test performance might be interrupted, but it was also necessary to ask the students to give more essential information during their reporting, for they might ignore some important ideas and easily forget them after finishing the items. Therefore, the way that the participants verbalize should be adjusted to fit their different characteristics. This experience helped me collect much more useful information in the formal test.

Within the 11 students’ verbal protocols in this study, only 7 female students’ accounts were translated into use and the other four students’ verbal reports were discarded for their indistinguishable voice, limited expression and other reasons.

4 Results

This study was designed to examine the students’ responses to different items, so before our investigation of the test-takers’ responses to the items, we need understand the constructs of the selected listening subtest. Based on the CET specification of abilities tested in the listening subtest, an experienced test expert was asked to define the constructs or item focuses of each MC item: only item 5, item 8 and item 15 were designed to examine test-takers’ understanding the connotative meaning, more specifically, deducing the implied meaning, and the rest 22 items were constructed to examine test candidates’ grasping and understanding specific details, in which most details were not essential at all. It seemed that the constructs of the MC items in this listening subtest only involved two required sub-skills outlined in the CET official test specification and therefore, might be under the threat of construct under-representation.

According to the data of students choosing each option in each item in Appendix 2, we found that the seven students all chose the intended right answer in 11 items: item 4, item 5, item 11, item 12, item 15, item 16, item 17, item 18, item 19, item 20, item 23. Among those items, item 16 to 19 are the four questions beneath passage one in section B of the chosen listening material. A passage has a specific context and covers considerable possible test focuses, so items of a passage should be designed to test a variety of listening skills in the specified construct. However, item 16 to 19 were all defined as to test students’ ability to understand specific details. Thus, the students’ think-aloud protocols of the four items below should be analyzed to testify whether they take the items by only referring to the under-representative and specific details.

<table>
<thead>
<tr>
<th>Table 1. Verbal reports on item 16 to 19</th>
</tr>
</thead>
<tbody>
<tr>
<td>Think-aloud protocols</td>
</tr>
<tr>
<td><strong>Student A:</strong> I think the answer should be A. It asks why the girl (Donna) joined in this organization (the Peace Corps). She was willing to help others. I have heard the word “help”.</td>
</tr>
<tr>
<td><strong>Student B:</strong> My answer is A, for she said that she hoped to help others.</td>
</tr>
<tr>
<td><strong>Student C:</strong> I think the answer should be A. She explained that she just wanted to help others, so I just chose A.</td>
</tr>
<tr>
<td><strong>Student D:</strong> Actually, I choose A, for I have heard the, the lady went there (in Brazil) in order to help others.</td>
</tr>
<tr>
<td>Item</td>
</tr>
<tr>
<td>------</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
</tbody>
</table>
Student F: I choose B. It seemed to have mentioned that she wrote stories for newspapers.
Student G: I choose B. In the passage, it clearly said that she wrote stories because she earned little as a teacher, not enough to make ends meet.

In Appendix 2, we also noticed three items in which less than half of the students chose the right answers. They were respectively item 8, item 14 and item 25. In view of the item constructs located by that test expert, item 8 was designed to examine the testees’ understanding the connotative meaning by reference, and item 14 and 25 were written to examine the students’ understanding specific details. According to Appendix 2, in item 8, three students chose the right answer D, another three chose A, and one chose C; in item 14, two students chose the right option D, four chose B, and one chose A; especially in item 25, only one student chose the right answer A, three chose D, two chose B, and one chose C. Then, it is necessary to check their responses to the three items individually to testify whether their wrong choices are caused by the intended variance-relevant difficulty.

Table 2. Verbal reports on item 8, item 14 and item 25

<table>
<thead>
<tr>
<th>Item</th>
<th>Think-aloud protocols</th>
</tr>
</thead>
</table>
| Item 8 | **Student A**: I think the answer should be A since I have heard the *key word* “trip”.  
**Student B**: My answer is D, for she said that she booked the ticket, and then the man said that she had not *taken a vacation* for two years. Then, she was going to go on a holiday.  
**Student C**: The answer should be B. She said that she went to inquire *the ticket to Miami* yesterday. As for the other options, nothing is given in the conversation.  
**Student D**: My answer is D, for I have heard that she wanted to *travel somewhere*, but the man said it seemed that she had no money, or maybe no time.  
**Student E**: En...I have only heard *the word “trip” and also “Miami”*, so I can only choose A between option A and B.  
**Student F**: En, I choose D, for the lady said that she was going to *travel in Miami and already made reservation*.  
**Student G**: I choose A. It seemed to say that she *went on a business trip*. |
| Item 14 | **Student A**: I think the answer should be B, for I have *only heard* clearly the phrase of option B and as for all the other three options, I *didn’t hear clearly*.  
**Student B**: For item 14…err…I want to choose B. Based on my note-taking, I haven’t heard any specific information relevant to this question. By referring to my notes, I *exclude* option A and C. Option A is what has been mentioned at the first, C is what has been said at the third, and I *guess* its options should be sequenced to the time. What this question asks about happens between option A and C, so I choose B.  
**Student C**: The answer to item 14, I think, should be D. Actually, *I am not sure*, but…I seem to have heard that he was going to see the…err…the Gavin or something, and something would happen at 13:00…having lunch.  
**Student D**: I choose D. I only heard that his arrangement in the afternoon seemed to include D. Then, *I guess* that D should involve having a meal in the morning or at noon. That’s all the information. I have clearly heard B and C, but I’m sure that B is the arrangement in the afternoon. Thus, I guess D may happen at the lunch time. |
seems that A and B are the arrangements in the afternoon, maybe meeting or something, but B seems to a possible answer.

Student E: Err…I heard something like “gatement (engagement)”, so I think the answer to this item should be…A, for at the beginning of this article, he seemed to meet someone named Mr. Gartner…There are a lot of persons’ names and many companies. When the play started, I even didn’t begin to read the options of this item and just finished reading option A of the first item in this passage.

Student F: En, en, I choose B, for, for, option C and D happened in the afternoon. Between option A and B, I guess B.

Student G: Err…My answer to this item is B, for I have heard that he will meet this guy (Gianni Riva at Megastar) at 11:00 a.m. and it should be having lunch.

### Item 25

Student A: I think the answer should be B, for I heard that it mentioned compensation at the end, and the word “charge” also appeared in the passage.

Student B: I think the answer should be D. Passage 3 has repeatedly stressed that his (Mclaughlin’s) driver license was problematic so that he caused so many damages.

Student C: I think the answer should be B. He seemed to have caused many car accidents as well as lots of damages, so I think I should be to pay damages.

Student D: For item 25, I choose…I am not sure, either C or D. I guess, after several car accidents, he was likely to retake a driver license, but I also heard something like a fine. If he got a fine, option B is similar to option C, and then it is difficult to exclude any option, so I choose D.

Student E: Err…it said that if he (Mclaughlin) caused three accidents, he would be discharged. And he had been asked to attend retraining by the police in his first accident, so I think I should choose D, en, discharging, discharge. I forget what this word means.

Student F: en, I choose A. En, I have heard the word “discharge” and it seems that he will be fired.

Student G: My answer to item 25 is C. Err…I have heard “discharge”, “Discharge” should mean to pay for something, so it should be a fine.

### 4.1 Discussion

Before we discuss those findings, it should be that this is only an exploratory qualitative study on probing evidence of construct under-representation and construct irrelevant variance in the test-takers’ think-aloud protocols of their responses to the listening items. In this study, the sample size is very small and all the targeted participants are at high level of English proficiency, so we are unlikely to discuss the difficulty and discrimination of each item from the perspective of statistics. Referring to the think-aloud protocols listed in the results section, we start our discussion with the first research question.

**4.1.1 What responses do the test-takers hold to the potentially invalid items? What differences are between test-takers’ responses to valid items and their responses to invalid items?**

In the results part, items in passage one are suspected to under-represent the required skill constructs in the test specification of listening comprehension issued by CET committee. The
items seems only to examine test-takers’ grasping of specific details in this passage. From italicized words in the seven students’ narration of their test-taking thought processes of item 16 to 19 in Sheet 1, we found all the students easily got the answer to the four items directly from the specific words or details without any inference. This means that the students’ correct responses to the items depend on their merely grasping of the specific details without any reference to the required skills in the CET test specification. This can be further proved by the record of Student B’s interview and she speaks of her ideas of the easiness and difficulty in the items:

…I think most items in this test are not very difficult. If you take notes well, you are basically able to get the points. Err…the premise is that it doesn’t ask for too specific details, and then you can choose the right answer by referring to the options while listening. Actually, the keys to the questions are very direct. You don’t need much inference. You only need to catch the specific details.

From Student B’s response, it can be concluded that the under-representation of skill constructs in a test is likely to cause test easiness and testing too much on specific details may ignore examining other required skills, such as test candidates’ understanding of main ideas and application of their various inferential skills. However, to substantiate the threat of construct under-representation of the listening subtest, further supporting proofs still need be obtained from the quantitative validation research on the factor analyses and the large-scale sample statistics on the test difficulty and discrimination of these items.

Three items 8, 14 and 25 are discussed individually, for there is much different information in the students’ think-aloud protocols in Sheet 2 and additional information from the private interviews to each student. Item 8 needs the students’ inference from the essential meaning of the long conversation. According to the think-aloud protocols in Sheet 1, those students who chose the right answer actually understood the essential meaning of this conversation and knew that the woman were going to take a vocation. Those who chose the wrong one were successfully distracted by the plausible option A and D, for they misunderstood “trip” or “Miami” as key words. The real reason why they have made the wrong choice is their failure to catch the main idea of the conversation in the listening material (see Appendix 1) so that they cannot make a reasonable inference. This can be further testified by Student E’s narration. An interview was given to Student E’s hesitation about this item, for she still chose option A after my intended second play of this audio to her:

I still did not hear it clearly. She seemed to have mentioned that she had not enjoyed a rest for two years, but I do not want to change my answer. I still heard the key words “trip”, “Miami” and also “reservation”, but I don’t know what she has booked. I choose A. I don’t know the reason and I have just made a casual choice…I heard that she had not rested for two years. Maybe, she was always complaining that she was so busy that she had to go on another business trip. Maybe, she might have worked for two years and finally got a vocation. I am not sure which one I should choose…I don’t want to change my answer, still A.

Student E’s words represent a typical Chinese style of thinking mode of taking an item, that is, making a choice with rich associations of much irrelevant stuff. In fact, from the italicized
sentence, we found that she should have nearly approached the right answer, but because of her failure to catch the essential conversational meaning, her hesitation ends with the original wrong choice A. Judging from the students’ responses, item 8 is a well-constructed one which can examine not only test candidates’ understanding of essential ideas but also their ability to make a reasonable reference.

Test-takers can easily get the right answers to the under-representative items, which may result in their uniform behavior and performance, but they need take advantage of more information and skills to work out a well-designed valid item. Therefore, test-takers’ responses to a valid item never be identical and may reflect what they are able to do as well as what they are weak in.

4.1.2 What impact do the testees’ reactions to the two main threats to construct validity reveal on their test performance?

In view of the students’ verbal reports on item 14 and 25 in Sheet 2, numerous expressions of uncertainty like “I guess”, “I am not sure”, “not hear clearly”, “difficult to exclude” (the italicized words and sentences in Sheet 2) are widespread in the students’ responses. Testwise strategies like exclusion, guessing, and associations based on fragmentary words and linguistic features are widely applied, for what they have heard is not enough to work out the items. Thus, the students’ responses should be specifically analyzed to detect whether there is any trace of the two main threats to construct validity.

First, item 14 focuses on much too specific detail and the required information in taking this item is much too overloaded. Even the two students (Student C and Student D) who have made the right choice actually are not sure about their final answers. In the protocols in Sheet 2, we found that Student C happened to hear the key words and Student D made a logical guessing by referring to the fragmentary information listened. Student D’s narration of her test-taking experience indicates that heavily loaded information makes it really difficult to get the right answer to item 14:

I think information needed for item 14 is too much overloaded for me. It seems…I have probably guessed its meaning, but I am not sure, for the information is too much in their conversation so that it is easy to confuse. For instance, I have heard two different times in this conversation—in the morning and in the afternoon, but it finally turns out to question about the lunch. The information load is too heavy and I even did not have enough time to glance the options at the beginning.

Therefore, item 14 is supposed to be under the threat of construct irrelevant variance, for it put much burden on testees’ memory and turns out to mainly test their memorization, not their listening skills any more. This information is detected by examining the responses of the test-takers who chose the right answer and cannot be found by any statistics and quantitative studies, which confirms the significance of a qualitative study on test-takers’ responses to the detection the two main threats to construct validity.

Second, Item 25 test in passage three examinees’ understanding of a specific word’s meaning
other than the skill constructs listed in the CET specification. This item narrows down its construct to test examinees’ understanding of the word “discharge” which appears in the last line and has no contextual information to make a reasonable inference in the previous part. What’s more, the only student, Student F, chose the right answer just because she knew the exact meaning of “discharge”. From the italicized and boldfaced words in Sheet 2, it is certain that another three students has also recognized the exact word “discharge” but still made a wrong choice because they are unaware of its meaning. Item 25 may be a good vocabulary item, for its distracters really have featured their utmost intended interference, which is shown by the students’ responses in Sheet 2. Whereas, it is an item under a passage of the listening comprehension subtest which should never focuses on the meaning of a specific word. At least, some contextual cues should exist in the passage to help test candidates understand this word. Thereafter, both item 14 and item 25 are likely to be under the threat of construct irrelevant variance, which drives test candidates to wisely use their testwise strategies, so their test performance depends on their use of testwise strategies or other chance factors other than their listening abilities and skills.

5 Conclusion

This study only functions as a tentative qualitative research on the detection of the two major threats to construct validity, construct irrelevant variances and construct under-representation and their impact on test-takers’ performance through the analysis of their test-taking responses and thought processes. It is found that construct under-representation probably leads to test easiness and test-takers’ uniform test performance and behavior and construct irrelevant variance causes examinees’ test performance to depend much on their use of testwise strategies or other chance factors other than their listening abilities and skills.

In this study, all the selected participants have archived a high English proficiency in high schools. In Wu’s (1998) study on MC format’s effect on test-takers’ test performance in listening comprehension, the MC method in listening comprehension has been found to pose threats to the construct validity of the test, for it favors the more advanced listeners but disadvantage the less able listeners. Therefore, we choose the students of high English proficiency as participants to get more accounts of their test-taking thought processes and experiences. Maybe, a qualitative study need be conducted on a larger sample involving different groups of listeners with varied English proficiency.

This study may inspire further research on tests using other test methods or tests of test-takers’ other abilities, but some points are in need of further improvement or revision and should also be considered in the future studies: the sample size is too small to involve any statistical consideration; the think-aloud protocols are carried out at the interval of each item, so this may affect participants’ performance and thinking continuity during the long conversations and passages, for many items are based on one listening material. Since the participants in this study are at high level of English proficiency, this concern matters little.
Acknowledgement

I would like to thank my tutor for her guidance and advice. I’d also like to extend my gratitude to all the students taking part in my study.

Reference


**Appendix 1**

*Short conversation 8*

W: Yesterday I made reservations for my trip to Miami next month.
M: You must really be looking forward to it. You haven’t had any time off for at least two years.

Question 8: What is the woman going to do?

*Long Conversation 2*

M: Err... Sandra, I've finished with Mr. Gartner now. Do you think you could pop through in bringing me up-to-date on their arrangements for the Italian trip?
W: Certainly, Mr. Wilkinson. I'll bring everything with me.
M: Right, take a seat. Now my first meeting is when?
W: Your first meeting is on Monday the 21st at 9a.m. with Dr. Gucci of Bancos en Piedra in Milan.
M: OK, so can I fly out early Monday morning?
W: Well, there is a flight to Lenarty Airport which leaves at six thirty London time and gets in at eight thirty Italian time.
M: Yeah, but that only leaves me thirty minutes to clear customs in getting to the city center and it means I have to check in by five thirty, which means leaving home at about four fifteen.
W: I'm afraid so.
M: Hmm... not so keen on that. What's the program for the rest of that day?
W: It's quite full, I'm afraid. At eleven, you're seeing Gianni Riva at Megastar and then you'll have a lunch engagement with Gavin from the Chamber of Commerce at one.
M: Where's that?
W: You're meeting him at his office and then he's taking you somewhere.
M: Good, that sounds fine. What about the afternoon?
W: Well, at three thirty, you're seeing our sales representative there and then you're free till evening.
M: I see. I seem to remember that I'm having a dinner with someone from Bergamo.
W: That's right. And Mr. Betty from SAP Industries at eight.

Question 14: Who is Mr. Wilkinson going to have a lunch with on Monday?

*Passage one*

Donna Fredrick’s served with the Peace Corps for two years in Brazil. She joined the Peace Corps after she graduated from the college because she wanted to do something to help other people. She had been brought up on a farm, so the Peace Corps assigned her to an agricultural project. Before she went to Brazil, she studied Portuguese for three months. She also learnt a great deal about its history and culture. During her two years with the Peace Corps, Donna lived in a village in northeast Brazil. That part of Brazil is very dry and farming is often difficult there. Donna helped the people of the village to organise an irrigation project, and she also advised them on planting crops that didn’t require much water. When Donna returned to the States, she couldn’t settle down. She tried several jobs, but they seemed very boring to her. She couldn’t get Brazil out of her mind. Finally, one day she got on a plane and went back to Brazil. She wasn’t sure what she’s going to do. She just wanted to be there. After a few weeks, Donna found a job as an English teacher, teaching five classes a day. Like most of the teachers, she doesn’t make
much money. She shares a small apartment with another teacher. And she makes a little extra money by sending stories to newspapers in the States. Eventually she wants to quit teaching and work as a full-time journalist.

Question 16: Why did Donna join the Peace Corps after she graduated from college?
Question 17: What was Donna assigned to do in Brazil?
Question 18: Why did Donna go back to Brazil once again?
Question 19: How did Donna make extra money to support herself?

Passage three
On March 13th, while on duty Charles Mclaughlin, a very careless driver employed by the company Lummis was involved in another accident. The accident occurred in Riverside California. Not paying attention to his driving, Mclaughlin turned right on main street and 33rd street and hit Volkswagen rabbit. This caused minor damage to his truck and serious damage to the car. On the basis of the police report, the Lummis accident committee correctly determined that Mclaughlin had been quite careless. As a result of the committee's conclusion, the branch manager Mr. David Rossi reported that he had talked with Mclaughlin about his extremely poor driving record. Further evidence of Mclaughlin's irresponsibility occurred on May 6th when he was returning from his shift. That day he ran into a roll-up door at the Lummis facility in Valero, causing significant damage to the door. Damage to the truck, however, was minor. Finally, on June 7th, Mclaughlin once again demonstrated his carelessness by knocking down several mail boxes near the edge of the company's parking lot. There was damage to the mailboxes and minor damage to the truck. Mr. David Rossi stated that he had spoken with Mclaughlin on several occasions about his driving record. He added that he had warned Mclaughlin that three preventable accidents in one year could lead to his discharge, as indeed it should.

Question 25: What is most probably going to happen to Mclaughlin?

8. A) Go on a business trip.
   B) Look for a job in Miami.
   C) Make a ticket reservation.
   D) Take a vacation.

14. A) Mr. Gartner from Milan.
    B) Gianni Riva at Megastar.
    C) The company's sales representative.
    D) Gavin from the Chamber of Commerce.

16. A) She had a desire to help others.
    B) She wanted to find out more about it.
    C) She needed some overseas experience.
    D) She was interested in farming.

17. A) Carry out a cultural exchange program.
    B) Work on an agricultural project.
    C) Learn Portuguese.
    D) Teach English.

18. A) She found it difficult to secure a job in her own country.
    B) She wanted to renew her contact with the Peace Corps.
C) She was invited to work as an English teacher.
D) She could not get the country out of her mind.

19. A) By teaching additional English classes.
   B) By writing stories for American newspapers.
   C) By working part time for the Peace Corps.
   D) By doing odd jobs for local institutions.

25. A) He will lose his job.
    B) He will have to pay damages.
    C) He will be fined heavily.
    D) He will receive retraining.

Appendix 2

<table>
<thead>
<tr>
<th>Item</th>
<th>Option A</th>
<th>Option B</th>
<th>Option C</th>
<th>Option D</th>
<th>Key</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item 1</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>5</td>
<td>D</td>
</tr>
<tr>
<td>Item 2</td>
<td>1</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>Item 3</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>Item 4</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>Item 5</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>Item 6</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>C</td>
</tr>
<tr>
<td>Item 7</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>Item 8</td>
<td>3</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>D</td>
</tr>
<tr>
<td>Item 9</td>
<td>0</td>
<td>5</td>
<td>2</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>Item 10</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>Item 11</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>Item 12</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>Item 13</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>D</td>
</tr>
<tr>
<td>Item 14</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>D</td>
</tr>
<tr>
<td>Item 15</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>Item 16</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>Item 17</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>Item 18</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>D</td>
</tr>
<tr>
<td>Item 19</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>Item 20</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>0</td>
<td>C</td>
</tr>
<tr>
<td>Item 21</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>Item 22</td>
<td>0</td>
<td>0</td>
<td>6</td>
<td>1</td>
<td>C</td>
</tr>
<tr>
<td>Item 23</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>7</td>
<td>D</td>
</tr>
<tr>
<td>Item 24</td>
<td>0</td>
<td>5</td>
<td>1</td>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>Item 25</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>A</td>
</tr>
</tbody>
</table>

Number of students choosing each option in each item