English Teaching, Vol. 61, No. 3, Autumn 2006

Construct Validity of Listening Test Items: A Verbal Protocol Study

Sang-Keun Shin
(Ewha Womans University)

Shin, Sang-Keun. (2006). Construct validity of listening test items: A verbal protocol study. *English Teaching, 61*(3), 293-305.

The present study aimed to explore how emphasis items in MELAB function by investigating test-taking processes of eight Korean college students. More specifically, employing verbal report methodology, this paper investigated whether the emphasis items require prosodic competence of test takers, which they are designed to measure. Verbal protocol analysis showed that they indeed require test takers to pay attention to stress in the input. The fact that test takers should use stress information to infer implied meaning does support the construct validity of these items. However, the protocol data also suggest that test method itself, primarily decontextualized input and multiple coice item format, may introduce construct irrelevant variance. For some items, test takers did not have to process the whole input because they were able to answer some of the questions by simply processing the emphasized words. For other items, they were able to choose the correct answer even when they misunderstood the message. The results of the present study demonstrates how test takers' reports of test taking processes provide language testers with valuable information about the construct validity of language test items.

I. INTRODUCTION

Taxonomies of listening skills have long recognized the importance of second language learners' ability to recognize the use of stress in connected speech (Buck, 2001; Munby, 1978; Richards, 1983). According to Dirven and Oakeshott-Taylor (1984), stress and intonation are more important to word recognition than the actual sounds. Lynch (1998) also suggests that prosodic features have a direct impact on how listeners interpret discourse segments and can carry considerable meaning that supplements the literal meaning of the words. Yet, despite this recognition, we know little about how to assess such ability. If our tests do not measure this critical aspect of listening ability, our operational definition of listening ability will be inadequate and results in construct under-representation (Messick, 1989) problems. Emphasis type questions in the Michigan English Language Assessment Battery (MELAB) are one of the first attempts to measure this important component of listening ability in the large-scale standardized testing context. Since this task type has not been validated, not much is known about the constructs being measured by these items, and the primary purpose of the present study was to examine the construct validity of the emphasis items.

Since the inferences made on the basis of test scores and their uses are the object of validation rather than the test task itself (Bachman, 1990; Chapelle, 1999; Messick, 1989), validation is a process through which a variety of evidence about test interpretation and use is produced. Since the emphasis type items are designed to measure test takers' ability to process the meaning conveyed by supra-segmentals or prosodic aspects of a speaker's utterance, a critical validation question is to what extent test takers' prosodic competence is indeed responsible for their performance in the emphasis items. But how do we know if the emphasis type questions require test takers' prosodic competence?

Recently language testers have begun to utilize verbal reports procedures to better understand the processes involved in taking language tests (Anderson et al., 1991; Buck, 1991; Cohen, 1984; Jeong-Won Lee, 2002, 2004; Sang-Keun Shin, 2005; Storey, 1997; Wu, 1998). Analyzing verbal protocols provides an opportunity to compare the process test takers actually used with the one they were intended to use, thus enabling language testers to evaluate the appropriateness of their presumptions about what's being tested. In line with these studies, exploring the extent to which examinees respond in an appropriate manner to test tasks (Henning, 1987), this paper examined whether the emphasis type items engaged the particular language competence, that is, the prosodic competence, that they are designed to measure. It is hoped that this study provides useful insights into what the emphasis items measure for those who are to measure prosodic competence of test takers.

II. CONTEXT OF THE STUDY

The listening section of the MELAB is a tape-recorded segment containing 50 questions

TABLE 1
The Structure of MELAB Listening Section

Item Type	Number of Items
Short Question Choose the appropriate answer to a short question	8
Statement Identify the paraphrase of single utterances	7
Emphasis Choose the appropriate response to short 13 expressions articulated with emphasis on particular lexical items Identify how a speaker might continue after emphasizing a certain lexical item	10
Lecture Select appropriate answers to short questions based on a 3-4 minute mini-lecture	13
Conversation Select the appropriate answer to short questions based on an approximately 4-5 minute conversation	12

which last about 25 minutes. All listening items are multiple choice with three options to choose from. Table 1 presents the overall structure of the listening section.

The MELAB listening section has two types of emphasis questions (English Language Institute, 1996). As can be seen in Example 1, the first item type involves presenting candidates a question or a statement spoken in a certain way, with special emphasis on a particular structure. Test takers then choose the answer that tells what the speaker would probably say next.

Example 1:
Tom said he was going to drive to Chicago **next** week…

a. not last week.
b. not next month.
c. not fly.

The other item type presents test takers a question with an emphasized word and asks them to select the most appropriate answer to the question, as can be seen in Example 2.

Example 2:
Do you have John's **keys**?

a. No, but Jane does.
b. No, I have Jim's.
c. No, only his bags.

The MELAB technical manual states that the emphasis items focus specifically on the meaning conveyed by supra-segmentals or prosodic aspects of a speaker's utterance. As the examples show, the utterance stem is deliberately ambiguous, and this ambiguity can only be resolved by integrating the stress information contained in the pronunciation of the utterance. Thus, the emphasis items measure the combination of a particular and important performance skill in listening linked with the capacity to extract underlying meaning from language (English Language Institute, 1996).

III. REVIEW OF LITERATURE

One of the first tasks that language testers should perform when they design a language test is to define the construct of the test and then to develop test tasks, which they assume, tap into the particular aspects of language abilities. For example, we may ask test takers to make predictions on how a story would develop in order to measure their ability to make deductions, which follow logically from information in the text. If test takers successfully complete the task, test developers and users will make an 'interpretive claim' (Kane, 1992) that the test takers possess this particular reading skill. However, it is not easy to prove whether the item has succeeded in measuring it because the exercise of reading skills does not usually manifest themselves directly in overt behavior (Hughes, 2003).

Another challenge for language testers is that in addition to the abilities we want to measure, the test methods have an important effect on test performance (Bachman, 1990; Shohamy, 1984). When test performance is unduly affected by factors other than the ability being measured, the validity of score interpretation will be lessened. For this reason, it is important to understand the nature and extent of the effects of test methods.

Language testers have investigated test-taking processes to examine whether language test items works as test developers intended (Buck, 1991; Cohen, 1984; Jeong-Won Lee, 2002, 2004; Jeong-Won Lee & Mee Ja Ku, 2005; Nevo, 1989; Sang-Keun Shin, 2005; Wu, 1998). As discussed above, this is an important validity question in that a wrong answer is not necessarily due to a lack of understanding and the wrong answer “may come through an alternative, equally valid cognitive procedure” (Jeong-Won Lee, 2004, p. 146). Based on the assertion that the introspection methods are a valid means to obtain data on test takers’ thought processes in test performance (Ericsson & Simon, 1993; Green, 1998), these studies typically ask them to verbalize what they actually do when they respond to the items.

Buck (1991) examined how 6 Japanese learners of English processed 54 listening test items by asking them to describe their test-taking processes. The protocol data showed that short-answer question type resulted in a number of problems that lowered the reliability of the test. For example, the test takers were not able to answer some of the questions because they had run out of time. Some of the inference type questions did not function as expected. For example, one item was not passage dependent in that anyone who had not heard the text could have provided the correct answer based on his or her general knowledge.

Employing an immediate retrospective verbal report procedure, Wu (1998) looked into test-taking processes of Chinese college students in a multiple-choice (MC) listening comprehension test. The examination of the protocol showed that the question and optional answers exerted a constraining impact on the students’ listening processes, suggesting that listening comprehension was essentially a process of making sense of the linguistic input. However, the MC format allowed much uninformed guessing, which sometimes led to the students’ selection of the right answer for the wrong reasons.

Applying a strategy checklist based on verbal reports protocol, Jeong-Won Lee (2002) investigated test-taking strategies of seventy-eight Korean college students while they were taking the reading comprehension subtest of a college entrance exam, which consisted of 28 reading passages. The verbal reports showed that the single foremost popular strategy was deductive reasoning, followed by clues in the text, process of elimination, and returning to the passage. The chi-square analysis of the students’ strategy use with regard to the question types showed a statistically significant result. Another chi-square analysis showed that there was a statistically significant association between the reported strategies and question types proposed by the Pearson and Johnson question and answer relationships. In addition, there was a statistically significant relationship between strategy use and the two different reading ability levels. Specifically, good readers reported more frequently 8 out of 9 significant strategies than poor readers, suggesting that good readers engaged as many desirable strategies as possible to help them understand the text and arrived at the correct answers.

Sang-Keun Shin (2005) examined composing processes of six graduate ESL writers in timed essay situations and found that the participants were not able to go through composing steps they normally do in drafted writing situations. They were forced to go through a linear composing process and were not able to enjoy heuristic discovery. Based on the results, he concluded that essay exams do not necessarily produce a representative sample of test takers' writing ability in target language use domains.

As these previous studies have demonstrated, studies of test taking strategies help language testers examine the closeness-of-fit between testers' presumptions about what is being tested and the actual processes that test takers go though to produce acceptable answers (Bachman, 1990). In line with these previous studies, this study aims to obtain the construct validity evidence of the emphasis items to assess the extent to which components of prosodic competence are responsible for test takers' performance in the emphasis items. As Buck (1991) has demonstrated in his introspective study, test methods contribute to the total test score variance of listening tests. However, only a few studies have explored how test methods influence test takers' performance on listening tests, and the present study attempts to fill the gap in the literature. Even though the scope of the present study is limited to the validity of the emphasis items, it is hoped that the results of this study will provide useful insights into what tests of listening comprehension really test. Here are the research questions:

1. What process do the participants go through while performing the emphasis items?
2. What effects does the test method have on the test takers' performance?

IV. METHODS

1. Data

Each MELAB form has 10 emphasis items, and this study analyzed 20 emphasized items in two forms of the MELAB listening test: Forms B and C.

2. Participants

Eight English as a Second Language (ESL) students volunteered to provide immediate retrospection of the test-taking processes that they went through while responding to the items. Since a range of levels were sought in order to compare the listening process of test takers of different levels in proficiency, both high-intermediate- and advanced learners of English were selected. The participants' TOEIC scores ranged from 750 to 945. None of the participants had taken the MELAB before, and thus they were provided with the MELAB test manual so that they were able to familiarize themselves with the item types in the listening section.

3. Procedures

The participants were asked to recollect what their thinking was as they responded to the items. After each item, the tape was stopped and the students were asked to verbalize how they processed the item. Their verbal protocol data were audio-recorded and then transcribed and analyzed by a research assistant and the researcher. As in a real test, they listened to the input only once and then provided retrospection. To examine if there is any difference in their strategy-use across item types, they were also asked to answer the short question items, which require them to choose one option that is a reasonable answer to the question they heard. They were allowed to use their native language in case their L2 proficiency was insufficient to verbalize their test-taking processes.

V. RESULTS

This section reports test-taking processes the participants employed while tackling the emphasis items. As Buck (1991) points out, presenting the results of protocol study involves a number of challenges. Protocol data consist of long verbal descriptions which are not easy to summarize. In addition, since protocol study typically involves relatively small number of participants, it is quite possible that descriptive statistics and item statistics do not reflect how test and items work with a larger sample. Given these limitations, item statistics have not been reported, and test-taking processes are divided into two types: construct-relevant and construct irrelevant processes, which are and are not relevant to the theoretical definition of the construct of emphasis items, respectively.

1. Construct-relevant Test-taking Processes

The analysis of the retrospection data showed that the participants employed different strategies for different item types. When they were responding to the short question items, they processed the whole input. As can be seen in Protocol samples 3 and 4, the test takers processed the whole input and based on their comprehension, they got these items right.

Protocol sample 3: Jae /Item 5

He or she asked what is the appropriate price for the painting. So the answer is a, not more than 25,000 dollars

Protocol sample 4: Min /Item 7

Well, did people stop him when he became famous, the answer is no, but they followed him.

On the other hand, for the emphasis items, the participants focused on emphasized words, and they were able to solve emphasis items by concentrating on them. For example, Naoko got the Item 17 right by processing '*first*' as can be seen in Protocol sample 5. Please note that the emphasized words in the input are presented in bold.

Protocol sample 5: Naoko /Item 17

First *two paragraphs. Did he say to read first two paragraphs? I am not sure. But I knew that I had to find out which word was emphasized. So I just tried to catch the emphasized word, and did not pay attention to the rest of the sentence.* ***First*** *two paragraphs, so the answer is c, not the last two.*

As Protocol sample 6 shows, Jin also used the same strategy when responding to the Item 23.

Protocol sample 6: Jin /Item 23

Cotton *ribbon, he said something about typewriter. I didn't exactly understand what he said but Cotton was definitely stressed. So the answer is b, no a Nylon ribbon, right?*

These findings show that the emphasis items tap into different type of listening skills than the statement items and that the test takers had to process the prosodic information of the input materials to provide correct answers.

Since they employed different strategies for different item types, they answered different item types incorrectly for different reasons. For the short question items, they provided wrong answers primarily when they failed to understand either the whole or part of the input. This was because these items ask test takers to choose a choice that is an appropriate response to the question they heard. In the following Protocol sample, Myung had trouble with the first part of the input, that is, '*can't you be more specific*'. Since he managed to understand the second half of the input, he made a semi-educated guess. He selected y*es, he's doing now* instead of *no, that's all I know*.

Protocol sample 7: Myung /Item 2

He said something about what he does but I didn't get the first part.

As for the emphasis items, on the other hand, the students gave wrong answers primarily when they missed emphasized words. For example, in Protocol samples 8 and 9, Hyun had trouble locating emphasized words, so she provided wrong answers to the both items.

Protocol sample 8: Hyun /Item 18
I heard that she wanted her children to have double room on the second floor. I understood what she said but failed to decide upon the emphasized word. There were three or four words that were stressed.

Protocol sample 9: Hyun /Item 20
The question was about what ***graduate*** *students should do by next Monday. But was next emphasized or was it graduate? Graduate was prominent but I had to listen to the rest of the sentence to see if another word was more prominent than graduate. And then I got confused. I am not sure which one was more emphasized.*

These findings show that the participants missed the emphasis items when they failed to locate the emphasized words, thereby supporting the construct validity of the items. Even when they understood the input, they missed the items when they failed to locate emphasized words.

2. Construct-irrelevant Test-taking Processes

The examination of the protocols showed that even when they misunderstood the input or failed to comprehend the whole sentence, they were able to answer some items correctly as long as they managed to locate the emphasized word. Had they been asked to explain in a sentence what the speaker wanted to say rather than to choose one of the three options, they must have missed these items. Since the test method, which had nothing to do with the construct of the items, allowed them to provide correct answers, these processes were considered construct-irrelevant. In the following Protocol sample 10, for example, Yun misunderstood the first intonation unit (Gilbert, 1993).

Protocol sample 10: Yun /Item 20
In the first part, the girl asked Are you a ***graduate*** *student ? Since* ***graduate*** *was emphasized, the anwer is c, no undergraduate.*

The question was “Are graduate students supposed to register..?” However, Yun misunderstood the first phrase *are graduate student* as *are you graduate student.* However, since she recognized that graduate was emphasized, she selected “undergraduate” as an answer. In the Protocol sample 11 below, Jae misheard ‘meal’ as ‘mail.’ Since he caught the emphasized word ‘pink,’ he got this item correct.

Protocol sample 11 : Jae /Item 22
The woman said eat ***pink*** *something before mail*

The fact that the students were able to provide the correct answer even when they misunderstood or missed a part of the input indicates that the multiple-choice (MC) format enabled them to select the right answer for the wrong reason.

However, this strategy does not always work. Even though Jin located the emphasized words in Protocol sample 12, she provided a wrong answer because she misinterpreted the meaning of verb ‘call.’ This finding suggests that these items also require bottom-up aspect of listening skills.

Protocol sample 12: Jin /Item 19
The director asked the gird to ***call*** *the new students.* ***Call*** *the new students, which means call their names. There is no corresponding response among the alternatives. Call their names, not write them. That doesn’t make a sense. There is no appropriate choice to choose.*

In another Protocol sample, Naoko provided an incorrect response to the following item because she perceived ‘my dress’ as one chunk.

Protocol sample 13: Naoko /Item 21
Will ***my*** *dress be ready by? Mydress Mydress, What is mydress?*

It is true that the test takers were required to process the emphasized words to solve the items, thereby supporting the construct validity of the emphasis items. However, the fact that the students were able to provide correct answers even when they misunderstood the input suggests that the MC format creates contruct-irrelevant variance because test method rather than the test takers’ listening ability allows them to get the items right.

VI. CONCLUSION

The main purpose of the present paper has been to use the verbal report methodology to examine how the emphasis items in MELAB function. More specifically, this paper investigated the extent to which the emphasis items require prosodic competence of test takers which they intend to measure.

Verbal protocol analysis showed that the emphasis items indeed require the test takers to pay attention to stress in the input to answer the questions. They had to first locate emphasized words and then inferred either what the speaker was going to say next or what was appropriate answer to the question. Even when they understood the whole input, they sometimes gave wrong answers to some items because they failed to locate emphasized words. The critical question is whether these are construct-relevant or construct-irrelevant test taking strategies. The fact that test takers should use stress informaiton to infer implied meaning appear to support the construct validity of these items.

However, the participants got certain items right sometimes for wrong reasons. Specifically, they did not have to process the whole input because they were able to answer some of the questions by simply processing the emphasized word. They were also sometimes able to provide the correct answer even when they misunderstood the message as long as they identified the emphasized words. Again the question is whether these test taking strategies are part of the communicative strategies that test developers hope would affect performance in the emphasis items. The protocol data suggests that the answer may be no. As discussed earlier, verbal protocol data indicate that test method itself, primarily decontextualized input and MC item format, may introduce construct-irrelevant variance.

This validation study has not drawn a decision that the emphasis items are valid or invalid. This is because validity is not an all-or-nothing attribute and that validation is an on going process (Bachman, 1990; Chapelle, 1999). Since the same test can be valid for one purpose but invalid for another purpose, each test user, thus, has to evaluate construct validity evidence of the emphasis items for a given purpose and draw his or her own validation conclusion. As for the emphasis items, different types of evidence can be collected. For example, future validation efforts may examine the extent to which construct relevant variables of the input, such as prominence of emphasized words, account for the variance of the item facilities. The prediction of item difficulties will provide evidence for the substantive aspect of construct validity by revealing the extent to which hypothesized knowledge and processes are responsible for test taker performance. Another interesting validation study will be to gather content evidence by collecting the judgements of experts concerning the precise ability (or abilities) that they believe that the emphasis items measure. As such, content analysis will provide evidence for a hypothesized match between test items and the construct that the items are intended to measure (Chapelle, 1999). As Jeong-Won Lee (2004) has successfully demonstrated in her study, a detailed investigation of the test-taking processes across the proficiency levels would also enhance our understanding of how test takers' language ability and test method work in a given assessment context.

Prosody is involved at all levels of listening processing, even single word (Buck, 2001), and is one of the main clues used by listeners to process incoming speech (Celce-Murcia, Brinton, & Goodwin, 1996). It may take time and trouble to design items to assess test takers' prosodic competence, but we need to remember that they are too important to exclude. The results of this introspection study may suggest that the emphasis item types have the promise of making very simple measures for prosody and may be quite suitable for large-scale standardized listening comprehension tests. Please note, however, that protocol analysis answers only a limited question of how the items work. It should also be taken into account that "it is impossible to prove that verbalized information actually reflects that is heeded as a task is carried out" (Green, 1998, p. 11). Therefore, we need to collect other type of evidence that the items do indeed measure what they are designed to measure.

Finally, this study has demonstrated the usefulness of verbal reports methodology for providing language testers with valuable information about how language test items function. It should be stressed that the emphasis item type is only one sample of test methods for listening tests. Given that "we are still rather ignorant in the area of listening comprehension" (Wu, 1998, p. 27), it is hoped that more efforts will be made to utilize this research tool for collecting validity evidence for listening test items.

REFERENCES

Anderson, N. J., Bachman, L. F., Perkins, K., & Cohen, A. D. (1991). An exploratory study into the construct validity of a reading comprehension test: Triangulation of data sources. *Language Testing, 8*(1), 41- 66.

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

Buck, G. (1991). The testing of listening comprehension: An introspective study. *Language Testing, 8*(1), 67-91.

Buck, G. (2001). *Assessing listening.* Cambridge: Cambridge University Press.

Celce-Murcia, M., Brinton, D., & Goodwin, J. (1996). *Teaching pronunciation: A reference for teachers of English to speakers of other languages.* Cambridge: Cambridge University Press.

Chapelle, C. (1999). Validity in language assessment. *Annual Review of Applied Linguistics, 19,* 1-19.

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

Dirven, R., & Oakeshott-Taylor, J. (1984). Listening comprehension (Part I). *Language Teaching, 17*(4), 326-342.

English Language Institute. (1996). *MELAB technical manual.* Ann Arbor, MI: University of Michigan.

Ericsson, K. A., & Simon, H. (1993). *Protocol analysis*. Cambridge, MA: MIT Press.

Gilbert, J. (1993). *Clear speech: Pronunciation and listening comprehension in North American English*. Cambridge: Cambridge University Press.

Green, A. (1998). *Verbal protocol analysis in language testing research.* Cambridge: Cambridge University Press.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Cambridge, MA: Newberry House Publishers.

Hughes, A. (2003). *Testing for language teachers* (2nd ed.). Cambridge: Cambridge University Press.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook.* Cambridge: Cambridge University Press.

Kane, M. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Lee, Jeong-Won. (2002). An exploratory study on reading comprehension test-taking process and strategies in the EFL context. *English Teaching, 57*(4), 177-195.

Lee, Jeong-Won. (2004). A study on English reading test-taking strategies. *English Teaching, 59*(4), 145-195.

Lee, Jeong-Won; & Ku, Meeja. (2005). A case study on the use of English reading test-taking strategies. *Journal of English Language Teaching, 17*(2), 179-201.

Lynch, T. (1998). Theoretical perspective on listening. *Annual Review of Applied Linguistics, 18,* 3-19.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan.

Munby, J. (1978). *Communicative syllabus design.* Cambridge: Cambridge University Press.

Nevo, N. (1989). Test-taking strategies on a multiple-choice test of reading comprehension. *Language Testing, 6*(2), 199-215.

Richards, J. C. (1983). Listening comprehension: Approach, design, procedure. *TESOL Quarterly, 17*(2), 219-240.

Shin, Sang-Keun. (2005). A construct validation study of timed essay tests. *Foreign Languages Education, 12*(4), 1-22.

Shohamy, E. (1984). Does the testing method make a difference? The case of reading comprehension. *Language Testing, 1*(2), 147-176.

Storey, P. (1997). Examining the test-taking process: A cognitive perspective on the discourse cloze test. *Language Testing, 14*(2), 214-231.

Wu, Y. (1998). What do tests of listening comprehension test?: A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing, 15*(1), 21-44.

APPENDIX

Sample Items

2. Can’t you be more specific about what he does?
5. What is a reasonable price for John to pay for this house?
7. When he became famous, did people keep stopping him on the street?
17. Now look Jane, I told you to rewrite the **first** two paragraphs.
18. I wanted my **children** to have the double room on the second floor.
19. The director told Carol to **call** the new students by Thursday.

20. Excuse me, are **graduate** students supposed to register for classes next Monday?
21. Will **my** dress be ready by four o'clock?
22. Should Jim take three of the **pink** pills before meals?
23. Does that electric type writer in Susan's office use a **cotton** ribbon?

Applicable levels: university
Key words: listening tests, validity, verbal reports, test-taking processes

Sang-Keun Shin
Dept. of English Education
Ewha Womans University
11-1 Daehyun-dong , Seodeamun-gu
Seoul 120-750, Korea
Email: sangshin@ewha.ac.kr

Received in May, 2006
Reviewed in June, 2006
Revised version received in August, 2006