Verbal Report in Language Testing

VERBAL REPORT IN LANGUAGE TESTING

Siwon Park

INTRODUCTION

Verbal protocol analysis (VPA) has become popular as a methodology to uncover psychological processes that a person goes through to perform a task (Faerch and Kasper, 1987; Ericsson and Simon, 1984, 1993). VPA is based on a strong assumption that subjects have "privileged access to their experiences" (Ericsson & Simon, 1993: xii), and that the information in their verbal reports is trustworthy. VPA is a different research technique from others that involve verbal reports since they are to be used to make direct inferences about the cognitive processes of interest (Green, 1998).

Since Ericsson and Simon (1984), numerous book volumes (e.g., Faerch & Kasper, 1987; Gass & Mackey, 2000; Green 1998) have been published for second language (L2) research to mainly introduce VPA. In addition, studies continue to appear which have adopted VPA as the primary research method in the field. In particular, the use of verbal reports has gained an increasing popularity as a viable research methodology in language testing because there have been frequent calls for the use of a process-oriented approach to test validation (Embretson, 1983; Messick, 1995; Ross, 1997). In L2 testing, verbal reports have been used to investigate mainly test-taking strategies and processes. Understanding such strategies and processes has been deemed crucial in drawing inferences about test-takers' abilities which are responsible for their performance.

In this paper on verbal protocols, I will first briefly discuss the characteristics of VPA . Secondly, I will review how verbal reports have been used to investigate differences of test-takers' strategies and processes for test-taking in language testing, and what the general findings were of prior studies. Finally, I will consider pedagogical implications of the use of verbal reports.

VERBAL PROTOCOL ANALYSIS

VPA has been developed as a methodology for examining thought and action (Pressley and Afflerbach, 1995). Under the information processing theory of memory, Ericsson and Simon (1993: xiii) assume that thought processes can be seen as a sequence of states of heeded information or thoughts. Therefore, the information or thoughts are relatively stable and can be verbalized. In think-aloud, for instance, subjects are instructed to "verbalize new thoughts and generate intermediate products as these enter attention" (Ericsson & Simon, *ibid*.: xiii). It is also assumed that when the subjects verbalize their thoughts with their attention focused on task performance, the sequence of thoughts is not altered by the very act of verbalization. Elaborating the temporal separation between processing and reporting, Ericsson and Simon (1994) draw a distinction between three levels of verbalization and argue that they are decreasingly reliable in order.

Levels of verbalization Level 1 verbalization is *talk-aloud,* which involves no intermediate processes and no additional oral encodings. Level 2 verbalization is *think-aloud,* which concerns descriptions or explications of the thought content. Verbalization at this level may take longer time than that of Level 1 since transformation of information may be required (e.g., transformation of images into words before they can be verbalized). Level 3 verbalization, (prolonged, as opposed to immediate) *retrospection* induces additional cognitive processing, therefore, changes one's thoughts or ideas. For that reason, the use of Level 3 verbalization is not recommended by Ericsson and Simon. They suggest that for valid elicitation of thought processes, the interpretive descriptions and explanations of cognitive processes must be left to the researcher, and instead he/she must encourage the subject only to focus on thoughts while performing the given task.

Categories of verbal reports Verbal reports have been subcategorized depending on how they are generated by the subject. For instance, subjects can generate verbal reports of their thoughts as they concurrently perform a task. Alternatively, verbal reports can be collected immediately after they complete the task (i.e., introspectively), or some time later on (i.e., retrospectively). With respect to the use of verbal reports for language testing research, Cohen[1] (1998, 2000, pp. 127-128; also cited in Gass & Mackey, 2000) classifies verbal reports into the following three subcategories:

- Self-report: learners' *general* description of what they usually do when they respond to a test item or take a test (e.g., questionnaires and interviews on general test-taking behaviors)
- Self-observation: the examination of specific language behavior either introspectively (within 20 seconds; e.g., *stimulated recall* in Gass & Mackey, 2000 or *immediate retrospection* in Yi'an, 1998) or retrospectively (e.g., questionnaires, journal entries, and interviews on a specific test-taking

[1] Note that Cohen's classification is not in line with Ericsson and Simon's, as Ericsson and Simon assume retrospection to be what Cohen proposes as introspection. In their work, as discussed earlier, Ericsson and Simon don't recommend use of what Cohen refers to as Self-report and Retrospection under Self-observation. They are labeled as Level 3 verbalization by Ericsson and Simon. Also, Ericsson and Simon specify 2-10 seconds for this retrospection (introspection in Cohen's classification) procedure.

instance)

- Self-revelation: concurrent think-aloud, i.e., stream-of-consciousness disclosure of thought processes while the information is being attended to

What distinguishes concurrent think-aloud from introspective self-observation is that concurrent think-aloud is only to reveal thoughts without attempts to analyze them. Once such thoughts are analyzed immediately after test-taking behavior, the data will become introspective self-observation. Also, self-report has been questioned for its validity due to the lag between the cognitive event of interest and the data collected. That is, learners may state their belief on what they usually do rather than what they actually did on the cognitive event (Cohen, 2000, p. 128).

In addition to concurrent think-aloud under self-revelation, immediate retrospective verbal reports may be preferred for certain types of tasks. Ericsson and Simon (1994, p. xvi) suggest that a subset of the sequence of thoughts occurring during task performance is stored in long-term memory. If stimulated immediately after the task is completed and using the cues in short-term memory, the sequence of thought would be retrieved. Therefore, so long as tasks can be completed in less than 10 seconds, subjects may be able to recall the actual sequence of their thoughts with high accuracy and completeness.

Limitations and criticism of verbal reports The popularity of verbal protocols is due to its methodological merit that looks directly into the “cognitive processes and learner responses that otherwise would have to be investigated only indirectly” (Cohen, 2000). However, the reliability and validity of verbal reports has been questioned. For instance, Nisbett and Wilson (1977, cited in Brown & Rogers, 2002) argue that “people often cannot report accurately on the effects of particular stimuli on higher order, inference-based responses…” (p. 233). In particular, once a cognitive skill becomes highly automatized, its underlying cognitive process may not be available for introspection. Therefore, it is recommended not to use a high-order cognitive process as the target of verbalization.

Another criticism of note is that the procedure of think-aloud may have an effect on task performance (Stoery, 1997; Stratman & Hamp-Lyons, 1994). Ericsson and Simon, however, argue based on the findings of their review study that verbal protocols are not reactive, that is, the act of verbalization does not affect performance or alter the sequences of thoughts. Although verbalization may slow down the task performance, it should not affect the task performance itself.

Moreover, Ross (1997) notes that the subjective nature of verbal report analysis may be a potential problem with introspective verbal protocol analysis (p. 236). In this regard, Ericsson and Simon (1993) contend that verbal reports collected through Level 3 verbalization must not be used. On the same account, they add that verbal reports may become unreliable if the interpretation of the reports is done by the subject(s) which again defeats the use of Level 3 verbalization in VPA. Therefore, the subject must report only on the content of working memory and not explain or evaluate their thinking in the verbalizing of their thoughts.

Thus far, it is not clear if there is a subsequent difference in quality between verbal reports produced through concurrent think-aloud and those obtained using self-observational procedures (e.g., introspective and retrospective methods). Cohen (1994a) argues that “it is possible to collect introspective and retrospective data from students just after they have answered each item on a multiple-choice reading comprehension instrument” (p. 127), citing Anderson (1991). In addition, the following suggestions are often made, in order to improve the quality of the verbal reports:

- Maximize the recency of verbal report of cognition and response to their actual occurrence.
- Use clear instructions that can help the subject to better access the information from their short-term memory.
- Train the subject enough to conform to the protocol instructions (especially., in case of concurrent think-aloud).

Other than the suggestions mentioned above, users of verbal protocols must be aware that VPA is to be used at best in connection with theories and theory building. That is, VPA must be used in light of relevant theories. Finally, recollections of recent episodes can possibly be valid; yet, dependent on retrieval cues. Typically, ‘why’ questions, questions about their motives for their behaviors, cannot be answered.

VERBAL REPORTS IN LANGUAGE TESTING RESEARCH

Proposing introspection as a method to investigate second language listening strategies, Ross (1997) considers verbal protocol analysis as a method of test construct validation and argues that it is still rare in language testing for logistical reasons (p. 218)[2] . He goes on to say that:

[2] Interestingly, the edited book on “validation in language assessment” by Kunnan (1998) did not include a single research paper using verbal protocol analysis, although the major theme of the book was to be on test-taking processes and test-taker characteristics and feedback. This shows the research tendency that focus mostly on quantitative approaches to the study of this topic. However, it has become more common for researchers to adopt statistical (i.e., product-oriented) methods as well as verbal report (process-oriented) analysis to draw more informed conclusions from their studies (e.g., Henning, 1992; Ross, 1997; Stoery, 1997)

> Introspection has considerable potential as a tool for investigating the psycholinguistic validity of item response patterns and can offer detailed qualitative data to supplement traditional and probabilistic approaches to test analysis, which have been limited to providing information about who should get items correct, but not why such items were correctly answered. (p. 219).

Accounting for the cognitive processes responsible for the observed performance behavior is proposed necessary for construct validation (Embretson, 1983; Messick, 1989, 1995). Understanding psychological as well as psychometric aspects of assessment is also considered necessary for generating validity arguments (Henning, 1992; Jourdenais, 2002; Snow, 1993; Storey, 1997). In that respect, there has been a line of language testing research that has adopted verbal protocol analysis in order to investigate if a test method or a test task helps elicit the right type of language samples intended for measurement, that is, the investigation of test method effect. Another line of research using verbal reports has concerned test-taking processes and strategies for the purpose of test improvement (extensively done by Cohen and his colleagues).

However, there has been notable confusion between strategies and processes, i.e., test taking strategies and processes. Process, as it includes strategies, is conceptually broader. Strategies are viewed as conscious process at least to some degree, but not all processes are conscious activities. In most cases, when certain cognitive activities become proceduralized, it is likely that they fall under the subconscious domain and become difficult to observe.

A distinction also needs to be made between two categories of strategies used by test-takers in language testing contexts (Rubb et al., 2006). On one hand, there are strategies that are employed by test-takers for successful completion of the skills engendered by the test (e.g., reading, listening, or speaking skills). These strategies must be viewed as construct relevant for score interpretation. On the other hand, there are strategies selected to deal with the cognitive demands introduced by test/task (i.e., method) characteristics. These strategies may be categorized as construct irrelevant and must be interpreted accordingly in making inferences about test-takers' ability for the language skill of interest. Studies (e.g., Nevo, 1989) have not proved the possibility of these two groups of strategies applied to the performance on the given test/task.

In language testing, the ability to use strategies was once categorized as strategic competence, which basically constitutes compensatory strategies to remediate lack of knowledge and skills to respond to a given task. However, under Bachman (1990) and Bachman and Palmer (1996), strategies are not considered compensatory any longer; rather, they are viewed as part of active cognitive processes adopted to complete the given task. This approach clearly subsumes the possibility of strategies as construct relevant. However, strategies (e.g., test-taking strategies) that Cohen and others have considered must be treated construct irrelevant that are tied mostly to the test method effect. That is, a distinction could be made between test/task (i.e., method) specific processes and processes that underlie the ability (construct) of interest.

PRIOR STUDIES USING VERBAL REPORTS

As mentioned earlier, in language testing, verbal reports have commonly been adopted to investigate test-takers' strategy uses on given tests/tasks. There are studies using VPA that have examined how test-takers responded to test items that measure language skills such as reading, writing, listening and speaking. Other studies have examined rater behavior, that is, how scores are assigned by raters. In this section, I will review prior testing studies that used verbal reports as at least part of the data sources according to the skills examined.

Verbal reports for reading and cloze tests Studies have been conducted to examine L2 learners' test-taking processes of reading comprehension and cloze tests (e.g., Feldmann & Stemmer, 1987; Rubb, Ferne, & Choi, 2006; Sasaki, 2000; Stoery, 1997; Yamashita, 2003). Feldmann and Stemmer (1987) appear to be the first study that used verbal protocols in L2 testing. They recognized the potential of verbal protocols as a methodology to enhance understanding of the processes that take place in learners working on an L2 test (p. 251). Their research goal was to investigate the construct of what the C-test is to measure. They attempted to identify and describe specific problem-solving behavior on the basis of strategies that were observed while test-takers were doing the C-test. For that purpose, they used think-aloud and retrospective interview. Their construct validation approach using VPA, although still premature at that time, opened up a possibility of VPA to be used for test validation.

One of the early test-taking strategy studies is Nevo (1989) in which she conducted a study on test-taking strategies on a reading comprehension test. In order to help test-takers' processing of response, she used a checklist of fifteen strategies each with a brief description. She found that there was a transfer of strategies from L1 to L2, although in the L2, students used more strategies that did not lead to the selection of a correct response than in their L1. One possible confusion notable in her study, however, was that strategies for reading comprehension and for test-taking were not distinguished which made it difficult to interpret her findings.

Stoery (1997), using concurrent think-aloud and immediate retrospection, investigated L2 learners' test-taking processes of a discourse cloze test designed to generate the discourse processing strategies. Stoery argues that the analysis of the test-taking strategies provide the evidence of the validity of test items and testing techniques because that type of analysis will help reveal if test-takers did engage in the processes supposed by the theory of ability. In his examination of the verbal reports, Stoery found that different items entailed varying degrees of construct validity. Also, a mismatch was noted between the theoretically assumed reading processes and the actual processes applied by some test-takers. He also commented on the problems he noticed with the verbal protocol techniques. Stoery found that the use of L2, Cantonese in thinking-aloud was not appropriate, although the subject was highly proficient in Cantonese. Moreover, he realized that the process of introspection itself may have affected the performance of the task of interest so that additional or different processes were employed.

Sasaki (2000), using retrospective (i.e., recall) protocols as part of the data sources, examined the effects of cultural schemata on the test-takers' processes of taking cloze tests. Her participants in the study completed either a culturally familiar or an unfamiliar version of a cloze test and were asked to give verbal reports of their test-taking processes. Her findings suggest that test-takers' cultural familiarity with the text content has an impact on EFL learners' performance on the cloze test and hence may pose a threat to adequate test score interpretation. Also, she found that the cloze tests could measure higher-order processing skills; yet, the position of blanks in the cloze test might introduce additional difficulty to test-takers' processes of test-taking.

Hudson and Park (2002) investigated how examinees react to various characteristics of web-based language testing for L2 reading and listening, especially as opposed to the paper-and-pencil (P&P) format test. A curriculum based low stakes Korean WBLT was developed and statistical as well as introspective methods were used to address their research questions. Using the think aloud protocols and the follow-up interviews, they revealed that the features of the WBLT were neither non-intrusive nor intimidating; therefore, did not increase test anxiety significantly.

Yamashita (2003) compares skilled and less skilled EFL learners in their processes of taking a gap-filling test. She adopted verbal protocols as part of her data sources to examine the construct validity of a gap-filling test as a measure of reading comprehension. She argues that her verbal reports revealed that generally text-level information was utilized by both skilled and less skilled readers, but their use of such information was not consistent between the two reader groups. In addition, she claims that gap-filling items used in her cloze test helped elicit either sentence-level or global-level reading ability, which supports the argument of the gap-filling test as a reading comprehension measure. However, her verbal reports also revealed that the gap-filling test generated processes not relevant to reading comprehension (e.g., grammar knowledge).

Both Sasaki (2000) and Yamashita (2003) recognize that studies conducted using verbal protocols suggest great potential towards the construct validation of tests and strongly recommend the researchers to adopt both product- and process-oriented methodology in testing studies. For instance, when verbal reports are used together with statistical data sources, one could gain insights that could have otherwise been missed in the absence of one or the other. That is, using the product- as well as process-oriented data, one could draw a stronger argument as to the validation of his/her test use (Messick, 1988).

Rubb, Ferne, and Choi (2006), using a semi-structured interview and concurrent think-aloud protocol, examined test-takers' use of strategies on multiple-choice (M-C) reading comprehension questions. The primary purpose of their study was to investigate the equivalence of reading processes and strategy uses in testing and non-testing reading conditions. Indeed, they found that different characteristics of M-C reading questions led test-takers to select different response strategies. Also, different M-C questions contributed to create different comprehension and response processes which strongly imply that reading comprehension for test-taking may not be the same as that in non-testing situations. Their findings of this method effect suggest important considerations as to how inferences of reading ability could be made based on the scores obtained using M-C reading questions.

Verbal reports for listening tests Buck (1990, 1991, 1994) reports a series of studies on Japanese college students on EFL listening tasks. These studies, originally from his dissertation work (1990), used verbal reports as part of the data sources. Students in Buck (1990) were asked to think-aloud as they performed EFL listening tasks on a narrative text. He analyzed the verbal reports to examine the types of knowledge, skills, and abilities that influenced item performance. Based on the results, he contended that each test response was a unique event, an interaction between a number of variables, many of which were personal and different from one test-taker to another. Buck concludes that language comprehension is by nature multidimensional, and testing it only increases the number of processing dimensions (1990, p. 424). Buck (1994) continues to argue that "it is difficult to conceive of listening tests measuring one unidimensional trait on which all test-takers can be placed in a linear progression from low ability to high ability" (p. 164). In that regard, it is not possible to say what each item measures.

Ross (1997) conducted an introspective analysis of listener inferencing on an L2 LC test. He asked his participants to provide an account of what words or phrases were heard in each test item and examined what item selection strategies were used. The use of recall data was to achieve further understanding about if high and low proficiency listeners may have applied misused selection strategies differently in relation to what he found from the comparison between item difficulty and ten strategies of interest.

Yi'an (1998), using retrospective verbal reports with two research questions, investigated the role of linguistic and non-linguistic knowledge in performing an M-C listening test, and if the M-C format of her listening comprehension test posed any method effect. She also examined if immediate retrospection would be found as a dependable research means to uncover listening processing. She found that listening comprehension is a process of making sense of the linguistic input in light of relevant non-linguistic knowledge and the purpose of listening. Yi'an's study also revealed that the M-C format differentially affected test-takers' performance on the listening test depending on their levels. In addition, guessing was found to be the factor that affected score interpretation. In her investigation of immediate retrospection for listening processing, data elicitation depended much on probing procedures; yet, properly employed, it helped reveal the processes of listening comprehension in test-taking.

Verbal Reports for speaking tests For performance testing such as speaking tests, it is not possible to use concurrent think-aloud. As a consequence, commonly adopted methods are retrospective method or interview. In addition, as Fulcher (2003) notes, "the focus is always on test-taking processes rather than test takers' cognitive processes, and the method really counts as an 'interview' rather than 'verbal protocol analysis" (p. 223; emphasis in original).

Cohen and Olshtain (1993) examined a role play using verbal reports in order to see what strategies test-takers use in achieving the test goal. Cohen, Weaver, and Li (1996) examined the effect of strategy instruction on L2 learners' performance on speaking tasks. The learners were asked to answer to the strategy checklists developed by the researchers beforehand, during, and after they completed the speaking task. A sub-group of learners were also asked to give their reasons for the frequency-of-use ratings that they had assigned to each strategy on the checklist by providing a verbal report while completing the checklist. Cohen, Weaver, and Li found that strategy instruction had a positive effect on the performance on specific tasks, but not all. Also, the increase in the use of certain strategies was related to the improvement in task performance, but not within a particular group, either comparison or experimental group or both. That is, certain strategies were more linked only to speaking performance improvement by either of the groups or both. Their strategy checklist also revealed that strategies were linked to specific tasks.

Swain (2001) took an interesting approach to assessing L2 speaking. In her study, Swain explored the potential of a dialog as a type of verbal report data. She suggested that the data could be used to promote understanding of how cognitive and strategic processes are constructed for performance on a task and how such information could be used in validating inferences drawn from test scores. Based on the findings, Swain argues that “the recording and examination of the dialogue of individuals jointly doing a task provides test-developers and test-researchers with additional insights to aid in the interpretation of test scores and to make recommendations about appropriate uses” (p. 297). Green (1998) also commented on the potential uses of verbal reports generated by two or more individuals working on a task together to understand the effect of tasks on performance conditions.

In the investigation of the comparability of direct and semi-direct speaking tests, O’ Loughlin (2001) adopted both qualitative and quantitative data collection techniques. In examining the process differences of the two test formats, O’Loughlin employed observations, interviews, and questionnaires, as relevant to one of three research phases. For the first phase of test design processes, non-intrusive observation and interview techniques were used. For the second test taking phase, he used questionnaires as well as interviews. For the last rating phase, immediate interviews as well as detailed questionnaires were carried out. The qualitative data collected throughout the three phases of process-oriented investigations confirmed the findings from the production data that the two tests of different formats tapped distinctly different components of oral proficiency. O'Loughlin's study serves a unique example of research that both qualitative and quantitative methods were employed to uncover some aspects of oral proficiency tests, of which the qualitative technique is rarely adopted.

Verbal reports for rater behavior

Orr (2002) used a verbal protocol analysis to investigate rater behavior. He asked 32 raters to assign grades based on the rating scales while watching videotapes of two FCE-type paired performance. He found that raters paid attention to aspects that were not present in the rating scales.

O'Donnell, Thompson, and Park (2006) conducted a verbal protocol study to understand rater behavior for second language oral assessment. They asked six raters to rate three testing sessions that involved four students in group discussion, and during and immediately after rating, they asked the raters to verbalize their rating processes. O'Donnell et al. found that raters have their own internal criteria for oral rating and pay attention to those features even though they are not described in the rating bands. Yet, they were mostly successful to negotiate their internal criteria with the institutionalized criteria described in the rating bands.

Brown, Iwashita, and McNamara (2005) conducted a rater orientation study using verbal reports, as part of a larger project. Unlike most other rater behavior studies, the purpose of Brown et al.'s study was to identify appropriate criteria for the assessment of test performance, rather than to determine how well raters were able to apply the specified criteria. Raters were told to provide the immediate verbal reports soon after they heard the first performance. Also, two task types were subjected to the rating processes. Brown et al. found from the verbal reports that all raters focused on the same general categories and tended to discuss the components of these categories in essentially similar ways (p. 101). They also found that among the categories, the content of test-takers' responses received a greater focus.

Fucher (2003) argues that this type of rater behavior study using verbal protocols reveals important information about how valid the inferences we made of learners' speaking ability as expressed as rating scores are, that is, how valid the rating processes are in assigning grades. This procedure suggests valuable information otherwise not obtainable for rater training and rating scale development/revision.

PEDAGOGICAL IMPLICATIONS

I believe the potential of verbal reports for testing studies have become thoroughly revealed throughout this paper. Another potential of verbal reports concerns the instructional use. Through thinking-aloud, instructors can help to make overt, to the students, the strategies they use to comprehend text and in turn that will help facilitate text understanding (Kucan & Beck, 1997).

Verbal reports collected from a learner(s) may be able to inform his/her teacher of where he/she experiences difficulty learning specific linguistic aspects. For instance, Cohen and Olshtain (1993) recommend teachers to devise a means for finding out more about the learning processes and strategies that their learners employ and to use the resulting information for advising.

In relation to L2 speaking, Cohen, Weaver, and Li (1996) suggest that if instructors systematically introduce and reinforce strategies that can help students speak the target language more effectively, their students may well improve their performance on language tasks. They also suggest that explicitly describing, discussing, and reinforcing strategies in the classroom can have a direct payoff on student outcomes. Such suggestions also call for training the teachers to learn how to deliver strategies-based instructions effectively in their classes.

Finally, the most common recommendation for the pedagogic use of verbal reports is made about strategy training. Studies that examine learner strategies using verbal report frequently recommend strategy training through verbalization.

REFERENCES

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

Bachman, L.F., & Palmer, A.S. (1996). *Language Testing in Practice.* Oxford: Oxford University Press.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-Academic-Purposes speaking tasks.* (TOEFL Monograph No. MS-29). Princeton, NJ: ETS.

Brown, J. D., & Rogers, T. S. (2002). *Doing second language research.* Oxford: Oxford University Press

Buck, G. (1990). *The testing of second language listening comprehension.* Unpublished doctoral dissertation, Department of Linguistics and Modern English Language, University of Lancaster.

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

Buck, G. (1994). The appropriacy of psychometric measurement models for testing second language listening comprehension. *Language Testing, 11*(2), 145-170.

Cohen, A. D. (1994a). *Assessing language ability in the classroom (2nd edition).* Boston, MA: Heinle & Heinle Publishers.

Cohen, A. D. (1994b). English for academic purposes in Brazil: The use of summary task. In C. Hills and K. Parry (Eds.), *From testing to assessment: English as an international language* (pp. 174-204). London: Longman.

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

Cohen, A. D. (1998b). Strategies and processes in test taking and SLA. In L. F. Bachman and A. C. Cohen (Eds.), *Interfaces between second language acquisition and language testing research.* Cambridge: Cambridge University Press.

Cohen, A. D. (2000). Exploring strategies in test-taking: Fine-tuning verbal reports from respondents. In G. Ekbatani and H. Pierson (Eds), *Learner-directed assessment in ESL* (pp. 127-150). Mahwah, NJ: Lawrence Erlbaum.

Cohen, A. D., & Olshtain, C. (1993). The production of speech acts by EFL learners. *TESOL Quarterly, 27*(1), 33-56.

Cohen, A. D., Weaver, S. J., & Li, T-Y (1996). *The impact of strategies-based instruction on speaking a foreign language.* Minneapolis: Center for Advanced Research on Language Acquisition, University of Minnesota (CARLA Working Paper Series #4).

Embretson, S. (1983). Construct validity: construct representation versus nomothetic span. *Psychological Bullentin, 93,* 179-197.

Ericsson, K. A., & Simon, H. A. (1984). *Protocol analysis. Verbal reports as data.* Cambridge, MA: The MIT Press.

Ericsson, K. A., & Simon, H. (1993). *Protocol analysis: verbal reports as data.* Cambridge: MIT Press.

Faerch, C., & Kasper, G. (1987). From product to process – introspective methods in second language research. In F. Faerch and G. Kasper (eds.), *Introspection in* *second language research* (pp. 5-23). Clevedon: Multilingual Matters Ltd.

Feldmann, U., & Stemmer, B. (1987). Thin- aloud a- retrospective da- in C-Te-taking: Diffe- languages diff- learners - sa- approaches? In Faerch, C. and G. Kasper (Eds.). (1987). *Introspection in second language research* (pp. 251-267). USA: Clevedon.

Fulcher, G. (2003). *Testing second language speaking.* Harlow: Pearson Education.

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

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

Henning, G. (1992). Dimensionality and construct validity of language tests. *Language Testing,* 9, 1-11.

Hudson, T., & Park, S. (2002) *Validity issues for selected versus constructed response Internet-based language tests.* Paper presented at AAAL, Arlington, Virginia

Jourdenais, R. (2001). Cognition, instruction and protocol analysis. In P. Robinson (ed.), *Cognition and second language instruction* (pp. 354-375). New York, NY: Cambridge University Press.

Kucan, L., & Beck, I. L (1997). Thinking aloud and reading comprehension research: inquiry, instruction, and social interaction. *Review of Educational Research, 67*(3), 271-299.

Messick, S. (1988). The once and future issue of validity: assessing meaning and consequences of measurement. In H. Wainer and H. I. Braun (Eds.), *Test validity* (pp. 33-45). Hillsdale, NJ: Lawrence Erlbaum Associates.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13-103). New York: American Council on Education, Macmillian

Publishing Company.

Messick, S. (1994). The interplay of evidence and consequence in the validation of performance assessments. *Educational Researcher,* 23(2), 13-23.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5-8.

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

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84,* 231-259.

O'Donnell, D., Thompson, G., & Park, S. (2006). *Revisiting assessmemt criteria in a speaking test.* Paper Presented at JALT2006 Annual Conference, Kitakyushu, Japan.

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. System, 30, 143-154.

Phakiti, A. (2003). A closer look at the relationship of cognitive and metacognitive strategy use to EFL reading achievement test performance. *Language Testing, 20*(1), 26-56.

Pressley, M., & Afflerbach, P. (1995). *Verbal protocols of reading.* New Jersey: Lawrence Erlbaum.

Riley, G. L., & Lee, J. F. (1996). A comparison of recall and summary protocols as measures of second language reading comprehension. *Language Testing,* 13(2), 173-189.

Ross, S. (1997). An introspective analysis of listener inferencing on a second language listening test. In G. Kasper and E. Kellerman (Eds.), *Communication strategies* (pp. 216-237), Rubb, A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: a cognitive processing perspective. *Language Testing, 23*(4), 414-474.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: a multiple data source approach. *Language Testing, 17,* 85-114.

Snow, R. E. (1993). Construct validity and constructed-response tests. In R. E. Bennett and W. C. Ward (Eds.), *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment* (pp. 45-60). Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Stratman, J. F., & Hamp-Lyons, L. (1994). Reactivity in concurrent think-aloud protocols: issues for research. In P. Smagorinsky (Ed.), *Speaking about writing: Reflections on research methodology* (pp. 89-112). Thousands Oaks, CA: Sage.

Swain, M. (2001). Examining dialogue: another approach to content specification and to validating inferences drawn from test scores. *Language Testing,* 18(3), 275-302.

Yi'an, W. (1998). What do tests of listening comprehension test? – A retrospection study of EFL test-takers performing a multiple-choice task. *Language Testing,* 15(10), 21-44.

Yamashita, J. (2003). Processes of taking a gap-filling test: comparison of skilled and less skilled EFL readers. *Language Testing, 20*(3), 267-293.