The Application of Verbal Protocol Analysis in Second/Foreign Language Testing Research

Sutthirak Sapsirin
Chulalongkorn University Language Instiute

Abstract

Verbal protocol analysis (VPA) is a research method that has been used quite extensively in second/foreign language (SL/FL) testing research. Its perceived value comes from its potential to reveal cognitive processes employed by test takers or raters, which can provide key insights into how one actually takes a test or rates test responses. This article aims to demonstrate how VPA has been applied in SL/FL testing research and propose other potential applications of the method. The article describes verbal protocol analysis in terms of its characteristics, use in language testing research, and procedures for data collection and analysis. Concerns about its validity are also presented. Finally, the article concludes with recommendations for further use of VPA in other areas.

Keywords: verbal protocol analysis, language test validation

การใช้การวิเคราะห์ **Verbal Protocol** ในงานวิจัยด้านการทดสอบภาษาที่สองหรือภาษาต่างประเทศ

สุทธิรักษ์ ทรัพย์สิรินทร์
สถาบันภาษา จุฬาลงกรณ์ มหาวิทยาลัย

บทคัดย่อ

การวิเคราะห์ Verbal Protocol เป็นวิธีวิจัยที่มีการใช้กันค่อนข้างกว้างขวางในงานวิจัยด้านการทดสอบภาษาที่สองหรือภาษาต่างประเทศ คุณค่าของวิธีการนี้ที่เป็นที่รับรู้กันคือวิธีการนี้สามารถใช้แสดงให้เห็นถึงกระบวนการทางปัญญาที่ผู้ทำแบบทดสอบหรือผู้ประเมินความสามารถของผู้เข้าสอบใช้ ซึ่งสามารถทำให้เกิดความเข้าใจอย่างถ่องแท้ถึงกระบวนการที่เกิดขึ้นจริงในการทำแบบทดสอบหรือการประเมินผู้เข้าสอบ บทความนี้มีวัตถุประสงค์เพื่อแสดงให้เห็นว่ามีการนำการวิเคราะห์ Verbal Protocol ไปใช้ในงานวิจัยด้านการทดสอบทางภาษาที่สองหรือภาษาต่างประเทศกันอย่างไร และเพื่อให้ข้อเสนอแนะเกี่ยวกับความเป็นไปได้ในการนำการวิเคราะห์ Verbal Protocol ไปใช้ในด้านอื่นๆ บทความนี้เริ่มต้นด้วยการอธิบายลักษณะของการวิเคราะห์ Verbal Protocol การใช้วิธีการนี้ในการวิจัยด้านการทดสอบทางภาษา และขั้นตอนการเก็บและการวิเคราะห์ข้อมูล นอกจากนี้ยังได้มีการนำเสนอปัญหาด้านความตรงของวิธีการ ในตอนท้ายบทความนี้สรุปด้วยการเสนอข้อแนะนำเกี่ยวกับความเป็นไปได้ในการใช้วิธีการนี้ในด้านอื่นๆ

คำสำคัญ: การวิเคราะห์ Verbal Protocol การศึกษาความตรงของการทดสอบทางภาษา

Introduction

Verbal protocol analysis (VPA) is a methodology that has recently received much attention in second/foreign language (SL/FL) testing research as it can offer insightful information which may not be available through other research methods. It has been used in SL/FL testing since the 1980s (e.g. Cohen, 1984a; Grotjahn, 1986) to explore the processes and strategies employed in test taking and rating.

The application of VPA in SL/FL testing research has been largely used in language test validation. The aims of the article are to demonstrate how the method has been used for this purpose and to suggest other potential uses of the method. First, the characteristics of VPA, its use in language testing research, and data collection and analysis procedures will be demonstrated. Then, concerns about its validity will be presented. Finally, recommendations for further applications of VPA in other areas will be discussed.

What is verbal protocol analysis?

Verbal protocol analysis is a qualitative methodology which asks participants to "think aloud" or "talk aloud" as they are performing a task (concurrent reports), or verbalize after they finish a task (retrospective reports) (Green, 1998). According to an information processing model proposed by Ericsson and Simon (1993), these verbal protocols (or verbal reports) are generated by "a subset of cognitive processes that generate any kind of recordable response or behavior" (p. 9). This model holds that the information that is stored in short-term memory (i.e. thoughts) while one is performing a task is the information that is reportable. In addition, information that is kept in long-term memory can also be reported after it has been retrieved. Based on this assumption, it is claimed that these types of verbal protocols, either concurrent or retrospective, are "the closest reflection of the cognitive processes" (Ericsson & Simon, 1993, p. 16), and that they can accurately reflect cognitive processes if appropriate techniques are used to elicit them (Ericsson & Simon, 1993).

Types of verbal protocols

Verbal protocols can be classified based on different criteria. As stated previously, verbal protocols comprise concurrent and retrospective reports (Ericsson & Simon, 1993). Concurrent reports are produced at the same time participants are carrying out a task. For example, a participant is asked to think aloud as s/he is reading a passage. Retrospective reports, on the other hand, are generated after participants finish a task. In the case of a reading task, a participant reads the passage first. After finishing reading, s/he will report their thoughts. Retrospective reports can be conducted with some stimuli to help participants retrieve their cognitive processes. This type of retrospective report, called stimulated recall (Gass & Mackey, 2000), can make use of such stimuli as the test taker's test booklet (Phakiti, 2003) and a video of the test taker performing a test task (Barkaoui, Brooks, Swain & Lapkin, 2013).

To elicit valid concurrent or retrospective reports, the researcher should ask participants to either talk aloud or think aloud, but not to explain or justify their thoughts (Ericsson & Simon, 1993). For talking aloud, participants are asked to say out loud everything that they say to themselves silently while they are doing a given task. Therefore, what is reported is already in verbal form. However, when doing some tasks, participants may also pay attention to non-verbal information such as that about a text (Green, 1998). When reporting their thoughts, participants then have to transform this type of information into a verbal form before verbalizing. This characterizes thinking aloud.

In addition to the categories described above, verbal protocols may differ in the way prompting or mediation is used (Green, 1998). In a non-mediated procedure, a participant is asked to talk aloud or think aloud and is prompted only when pausing for a period of time. The prompts will be non-intrusive; for example, the researcher may say "Keep talking" to remind the participant to continue thinking aloud. In a mediated procedure, in contrast, the researcher will ask participants to explain, justify, etc. their thinking processes in addition to talking or thinking aloud. Both non-mediated and mediated procedures may be used for concurrent and retrospective reports.

In SL/FL research, verbal reports can also be categorized in a somewhat different but overlapping way. That is, they can be classified as self-report, self-

observation or self-revelation (Cohen, 2000; Cohen & Hosenfeld, 1981). In the context of language testing, self-report is "learners' description of what they do, characterized by generalized statements" (Cohen, 2000, p. 127) about test-taking strategies. That is, participants describe the way they usually take a test. Self-observation is "the inspection of specific, not generalized, language behavior" (Cohen, 2000, p. 127) either introspectively (i.e. within 20 seconds of the cognitive event) or retrospectively (20 seconds or so after the cognitive event) (Cohen, 1984b). This type of data involves reference to some actual language testing event. Both self-report and self-observation can be elicited by asking participants to speak about the strategies they use or by other means such as questionnaires and diaries.

The last type of verbal report, self-revelation, or think-aloud, is defined as "stream-of-consciousness disclosure of thought processes while the information is being attended to" (Cohen, 2000, p. 128). Self-revelation differs from self-observation in that self-revelation data are participants' thoughts that are not analyzed; however, self-observation data are thoughts which are analyzed then reported by the participants. When comparing the three types of data, Cohen (2000) points out that self-observation and self-revelation data might be more valid than self-report, due to it being a description of generalized behavior and does not concern the description of what participants actually do during or after the task performance.

Use of VPA in language testing research

The literature indicates that VPA can be a useful tool for language research. Its value is derived from its ability to reveal information on cognitive processes underlying performance that cannot be obtained by other research techniques (Buck, 1991; Camps, 2003; Kormos, 1998; Weigle, 1999). The method makes it possible to investigate cognitive processes more directly (Cohen, 2000; Wigglesworth, 2005) such as processes in composing (Smagorinsky, 1989), reading (Crain-Thoreson, Lippman & McClendon-Magnuson, 1997; Hosenfeld, 1984; Pressley & Afflerbach, 1995), listening (Goh, 2002) and speaking (Cohen & Olshtain, 1993).

The literature also shows that VPA continues to play an increasingly significant role in language testing research. This is evident from the number of studies which have

used VPA independently (e.g. Buck, 1991; Orr, 2002; Sakyi, 2000), and with other qualitative or quantitative methods (e.g. Anderson, Bachman, Perkins & Cohen, 1991; Cohen, 1994; Milanovic, Saville & Shuhong, 1996; Phakiti, 2003; Sasaki, 2000; Weigle, 1999).

The increased use of VPA in research may be in response to a call for greater application of VPA as well as other qualitative methods in language test validation (e.g. Bachman, 2000; Banerjee & Luoma, 1997; Grotjahn, 1986; Lazaraton, 2008). The growing interest in VPA and other qualitative approaches may reflect "the introduction of the view of language as communication and the consequent rise of performance assessment; the increased importance of process in theories of learning and teaching; and more recently, the legitimacy of multiple perspectives and constructions" (Banerjee & Luoma, 1997, p. 275).

The current thinking on validity, which has changed the way validation research is carried out, also has led to increased use of VPA in language testing research. For example, Messick's (1989) unified validity framework has been greatly influential in educational and language assessment research (see Chapelle, 1999, for review on validity in language assessment). Using Messick's (1989) work as a foundation, Bachman (1990) describes types of evidence which can be used to support the interpretation of test scores and test use. To investigate construct validity, one of the several approaches that can be taken is analysis of processes underlying test performance, which includes verbal protocols among other methods. In language test validation processes, VPA can be used to answer such questions as:

> Does the test in question actually measure the set of skills it purports to measure?
> Do two different versions of the same test measure the same skills?
> Do the raters heed the marking criteria in assessing performance on the task in question?
>
> (Green, 1998, pp. 14-15)

The literature has revealed that VPA can be employed to investigate a variety of issues in SL/FL testing. The following section presents the topics that have been studied through VPA. However, it should be noted that several topics may be examined in the same study.

***Nature of constructs*:**

Constructs that have been examined are, for example, those of reading (e.g. Rupp, Ferne & Choi, 2006), listening (e.g. Buck, 1991), speaking (e.g. Sato, 2014) and strategic competence (e.g. Phakiti, 2003). For instance, Sato (2014) examined the construct of interactional oral fluency between peers by using VPA, correlation and regression analysis. VPA was employed to compare raters' perceptions of individual and interactional oral fluency, and the two quantitative analyses to examine the relationship between the rated scores and the temporal aspects of speech. The analysis of verbal protocols revealed that an important component of interactional oral fluency was scaffolding. In addition, another component, pauses, was viewed differently in the two types of performance. With regard to the quantitative analyses, it was found that individual oral fluency was a weak predictor of oral fluency in the interactional context. These findings indicate that individual and interactional oral fluency may be different constructs, and that the latter should be considered a joint performance between speakers.

Another study of constructs demonstrated the use of VPA along with another qualitative method rather than quantitative ones as used in Sato's (2014) study. Rupp et al. (2006) analyzed interview and concurrent verbal reports of 10 ESL learners while responding to a reading test with multiple choice (MC) questions. They found that the construct of reading comprehension in a testing context is shaped by item design and text selection, which makes it different from the construct of reading in non-testing situations. In a testing context, test takers relied on key word matching when responding to MC questions. Their response processes were also affected by the difficulty of the text or the questions and were not linear as the processes proposed in a model of reading comprehension in a non-testing context were.

How test takers approach a test:

A number of studies have looked into what test takers attend to when taking a test (e.g. Bax, 2013; Wagner, 2008; Xu & Wu, 2012). For instance, Wagner (2008) investigated how eight ESL learners attended to and used nonverbal information in a video listening test to process the video text and answer comprehension questions. The

participants gave concurrent verbal reports while watching a video text and while answering test questions. The results showed that the participants made a reference to nonverbal information in the video texts. However, they varied in their ability to process and use the nonverbal behaviors to understand the video texts and answer comprehension questions. Based on the findings, the researcher argued that nonverbal information is important in processing spoken language. Therefore, to test listening ability, a video listening test should be used rather than an audio-only test as the former allows the listener to use components of spoken language that are part of real-life listening tasks.

Another study employed a different type of VPA to examine test taking processes. In a study on reading tests conducted by Bax (2013), stimulated retrospective recall interviews were used to supplement eye-tracking data to investigate cognitive processing of test takers performing a reading test. The data for eye tracking were collected from 38 participants, 20 of which were randomly selected for a stimulated recall interview. The study found that proficient and less proficient test takers significantly differed in their ability to read expeditiously and in attention paid to some aspects of test items and reading texts.

In a study that explored test taking strategies for a high-stakes writing test with picture prompts, Xu and Wu (2012) combined two types of VPA with other research techniques. That is, they collected think aloud and retrospective interview protocols from 12 students, analyzed their writing and interviewed four of the students' teachers. It was found that students employed a variety of test-taking strategies as coached in their classrooms. Moreover, for fear of losing points, they avoided expressing their own ideas in one of the writing tasks, which contradicts what the test task aims to measure.

Processes test takers employ in integrated tests:

As integrated tests have become more widely used, studies of test taking processes in such tests have received more attention during the past several years (e.g. Barkaoui et al., 2013; Plakans, 2008, 2009; Plakans & Gebril, 2012; Weigle, Yang & Montee, 2013). For example, Barkaoui et al. (2013) employed stimulated recalls to 30 test takers to examine their strategic behaviors in performing integrated and independent speaking tasks in the TOEFL iBT and the relationship between these behaviors and their

test scores. After completing each task which was video-recorded, the test takers watched the video and reported what they were thinking while performing the test. The analyses showed that test takers used more strategies when taking the integrated tasks than the independent tasks. In addition, the strategies used in different integrated tasks were similar to each other and differed from the independent tasks. Finally, there were no significant relationships found between strategies and total test scores. The researchers concluded that the findings provide support for the inclusion of integrated tasks in a speaking test.

In a study of reading-into-writing tasks, Weigle, Yang & Montee (2013) explored the reading processes test takers used when they performed a reading test in which they responded to test questions by writing short answers. Similar to Barkaoui et al. (2013), this study used a variety of data, that is, they collected think-aloud data, retrospective interviews, semi-structured interviews and test scores. The results revealed that the test takers engaged in reading processes that appeared in the real world context and they needed to apply a high level of language proficiency to successfully understand the texts and respond to short answer questions. These findings provide evidence for the validity of the test.

Factors that can affect test-taking processes:

Several studies have examined factors that influence test-taking processes, for example, the effects of test method (e.g. Buck, 1991; Yi'an, 1998), test task difficulty (e.g. Babaii & Moghaddam, 2006), topic familiarity (e.g. Lee, 2015), and cultural schemata (e.g. Sasaki, 2000). Babaii and Moghaddam (2006), for instance, examined the effect of test task difficulty on test takers' macro-level processing when doing a *C*-test. Four test tasks were used; each was different in terms of text difficulty (low vs high level of syntactic complexity and abstraction), and the presence of clues about the number of missing letters (presence vs absence of clues). 119 students took the test and 36 of them gave retrospective think aloud protocols. Students' scores were analyzed with ANOVA and the frequency and percentage of protocols with chi-square analyses. The results showed that texts that had a high level of syntactic complexity and abstraction and had

no clues increased the difficulty of test tasks. This, in turn, elicited more macro-level processing.

Like Babaii and Moghaddam (2006), Lee (2015) used both quantitative and qualitative methods for a study which looked into the impact of topic familiarity on strategies used in reading tests. 36 EFL students took a reading test with familiar and unfamiliar topics, produced retrospective protocols and were interviewed. The analysis of verbal protocols showed that students used six categories of strategies: general approaches, discourse structure, vocabulary/sentence-in-context, multiple-choice test management strategies, test wiseness and background knowledge. In addition, results of ANOVA analysis showed that strategies used in taking the test with familiar and unfamiliar topics were not statistically significant.

Test development or revision:

Some studies have incorporated verbal protocol analysis in their test development or revision processes (e.g. Liu, 2007; Uiterwijk & Vallen, 2005). Liu (2007), for example, developed a pragmatic test for Chinese EFL learners with a focus on the speech act of apology. The test development consisted of several stages. First, a group of Chinese EFL learners were asked to provide examples of situations which involved apologies and state how likely they felt the situations were to occur. Then, a metapragmatic assessment was used to find out whether the variables in the situations the learners provided were perceived by Chinese university students and native speakers of English in the same manner. Next, the resulting situations were incorporated into a questionnaire and pilot tested. After that, multiple choice options were created for the items. Finally, the construct validity of the new questionnaire was investigated through Rasch analysis and VPA. The findings showed that the data from verbal protocol analysis supported those from the Rasch analysis, which indicated that the test was a useful instrument to measure pragmatic knowledge.

In another study, VPA was also employed along with other techniques to alert test developers to possible item bias against immigrant students in a Dutch achievement test (Uiterwijk & Vallen, 2005). The procedures included (1) statistical differential item functioning (DIF) detection methods, (2) an investigation of sources of DIF which

consisted of literature search, content analysis, expert judgements, and students' think aloud data, and (3) identification of biased item. It was found that 17.4% of test items may cause DIF and possible DIF sources were, for example, the use of idioms, low-frequency words, and subject matter that involved Dutch culture. However, only some of these items that contained elements that were not part of the construct to be measured were considered biased items, which constituted 4% of all items.

In addition to data obtained from test takers, verbal protocols can provide rich evidence about raters' performance. The following are topics that have been examined by verbal reports from raters.

Raters' decision making process:

Several studies have looked at how raters judge the quality of test takers' performance and determine scores, for example, of writing (e.g. Cumming, 1990; DeRemer, 1998; Gebril & Plakans, 2014; Wiseman, 2012), speaking (e.g. Ang-Aw & Goh, 2011; Brown, Iwashita & McNamara, 2005; Weigle, 1999) and of vocabulary (e.g. Li & Lorenzo-Dus, 2014). For instance, Li and Lorenzo-Dus (2014) used think aloud to investigate how raters assessed vocabulary in a speaking test. The analysis of verbal protocols of 25 raters showed that raters focused on both vocabulary and non-vocabulary features when assigning vocabulary scores. The vocabulary feature that was most frequently attended to was lexical sophistication and the test taker's use of advanced words had a direct impact on the vocabulary scores s/he received. In addition, the non-vocabulary features that raters paid attention to, for example, pronunciation, fluency and grammar had an impact on their vocabulary rating. These findings indicate that it may not be possible to rate vocabulary as a discrete construct in a speaking test.

Gebril and Plakans (2014) investigated raters' decision-making behaviors while they rated reading-to-write tasks, the way they approached source use, features that influenced their scoring, and the challenges they faced when scoring. The study did not only collect data from think aloud like the previous study but also from interviews of two raters. It was found that in terms of decision-making behaviors, raters reported more judgment behaviors (the processes of evaluating essay quality) than interpretation behaviors (strategies used to make sense of the essay). As for the way raters approached

source use, raters located source information, checked citation mechanics and judged the quality of source use. With regards to features that influenced scoring, linguistic features and citation mechanics were reported as critical when raters scored lower level essays. As score levels were higher, raters shifted their attention to organization and development issues and quality of source use. Finally, raters reported several challenges in rating: difficulties in identifying text from source materials and that produced by the test takers, difficulty in scoring texts that contain copied source materials or overuse of quotations, and difficulty in scoring borderline essays. It was concluded that integrated tasks are complex and require rater training and rubrics that address these challenges in order that scores derived from such tasks will be justifiable.

How raters interpret oral interaction:

As many speaking tests now use pair or group test tasks, research has been conducted to explore the rating processes of this task type. For instance, Ducasse and Brown (2009) investigated what raters focused on when they assessed paired interaction in a speaking test. The researchers asked 12 experienced teacher-raters to give both retrospective reports and stimulated recall after watching videos of test performance. The findings showed that the raters focused on three interactional features when rating a paired oral test: non-verbal interpersonal communication, interactive listening and interactional management. The researchers suggest that the results can be used to define what interaction is and to develop interaction-based rating scales for speaking.

Factors that can affect variability in rating processes:

The factors that have been investigated which can affect variability in rating processes include those such as test takers' first language (e.g. Winke & Gass, 2013), task characteristics (e.g. Weigle, 1999), rating scales types (e.g. Barkaoui, 2007, 2010; Li & He, 2015), rater experience (e.g Barkaoui, 2010; Connor-Linton, 1995; Isaacs & Thomson, 2013; Joe, Harmes & Hickerson, 2011; Weigle, 1999), rater training (e.g. Weigle, 1994) and raters' first language (e.g. Zhang & Elder, 2014). For instance, Winke and Gass (2013) examined the influence of raters' knowledge of test takers' L1 on rating their oral proficiency. In the study, 26 raters were videotaped while rating test

takers from three L1 backgrounds. Then, they watched the videos of themselves and reported what they were thinking at that time. The data from the stimulated recall revealed that a test taker's accent and L1 can affect the rating of some raters, which can lessen score reliability.

Unlike Winke and Gass (2013), Li and He (2015) incorporated VPA with other techniques to investigate the use of holistic and analytic rating scales by 9 raters assessing essays. That is, the study used concurrent think aloud, questionnaires and semi-structured interviews. The findings showed that when using the holistic scale, raters more frequently used self-monitoring-interpretation strategies, the strategy of considering local language features and some self-monitoring-judgment strategies. However, with the analytic scale, self-monitoring-judgement strategy, error-classifying strategy and quality-assessing strategy were more often used. In terms of text focus, with the holistic scales, the features that raters paid more attention to were the general quality of language use and non-scale-related language features. However, with the analytic scale, the features that received more attention were coherence and grammar. The study shows that scoring rubrics have an influence on rating processes and that raters interact with rubrics in different ways.

Another study looked into the effect of raters' first language employing not only VPA but also quantitative methods. Zhang and Elder (2014) compared native and non-native English speaking raters' behaviors when they judged oral performance of test takers using Many-facet Rasch measurement and content analysis of their stimulated recall protocols. The quantitative analysis showed that the scores assigned by both groups of raters were similar in terms of score consistency and rating severity. Similarly, the qualitative analysis revealed that raters were not different in terms of the features that they focused on when they applied the rating scale. These findings led to the conclusion that raters' L1 may not affect rating outcomes in oral assessment given that appropriate training in the use of rating scale has been provided. The finding also supports the claim that native and nonnative raters do not apply different standards in assessing oral language performance.

Development of a framework or model of scoring processes:

Some studies using VPA have aimed to construct a model which describes raters' behavior and criteria they use in essay rating (e.g. Cumming, Kantor & Powers, 2002; Sakyi, 2000). Cumming et al. (2002), for instance, conducted three coordinated studies that aimed to develop a framework to describe raters' decision making processes while holistically evaluating ESL/EFL compositions. All studies collected and analyzed concurrent verbal reports from experienced raters. The finalized framework consisted of 27 decision making behaviors which fell under self-monitoring focus, rhetorical and ideational focus, and language focus.

Development of scoring rubrics:

Some studies have used VPA to develop scoring rubrics (e.g. Zhao, 2012). For instance, Zhao (2012) aimed to develop and validate an analytic scale of voice in L2 argumentative writing by using both qualitative and quantitative analysis of rater performance. The qualitative data analysis which involved think aloud of four raters followed by interviews supported the quantitative data analysis which found that the construct of voice includes three subcomponents: the presence and clarity of ideas in the content, manner of idea presentation, and writer and reader presence. The qualitative data also yielded valuable information on what raters viewed as important in measuring voice, which was not present in the scoring rubric. The data from both analyses led to the revision and validation of the new rubric of voice. It was found that the revised rubric could be useful in measuring voice in L2 argumentative writing.

Data collection procedure for verbal protocol analysis

The following points should be considered when collecting verbal protocols (Bowles, 2010; Ericsson & Simon, 1993; Gass & Mackey, 2000; Green, 1998):

Determining the appropriateness of a task:

Before using VPA, the researcher should first determine whether or not the task proposed is suitable for the methodology (Green, 1998). Reading, listening, writing or speaking tasks are generally suitable for protocol studies. However, the following tasks

are not likely to yield useful information on thought processes: tasks that involve guessing, tasks that require Yes/No or True/False responses, tasks that are too simple for the participants, perpetual-motor tasks and visual encoding tasks, and speaking tasks where the participants are asked to give concurrent reports (Green, 1998).

Task analysis:

The next step involves analyzing the task to identify a set of possible strategies that participants may use to carry it out (Ericsson & Simon, 1993; Green, 1998). This can help the researcher to construct a coding scheme for data analysis.

Procedure selection:

In this step, the researcher chooses between talk aloud and think aloud methods, concurrent and retrospective reports as well as mediated and non-mediated procedures (Green, 1998). Green (1998) recommends concurrent reports, except for listening, speaking and simple reading test tasks. In cases where retrospective reports are chosen, the shorter the delay, the more likely the verbal reports can reflect the actual processing (Ericsson & Simon, 1993; Gass & Mackey, 2000; Pressley & Afflerbach, 1995). Ericsson and Simon (1993) also prefer concurrent protocols over retrospective ones; however, they recommend using both sets of data. This is because even though retrospective reports can be incomplete, they can provide the general structure of the thought processes; thus complementing data obtained from concurrent reports. Other researchers also agree with this idea as the data obtained from concurrent reports alone may indicate that some participants do not use the target processes. However, when retrospective reports are also collected, the data can reveal that more participants actually do use the processes, reflecting a more accurate number (Alavi, 2005; Camps, 2003). In addition to using more than one type of VPA in a study, the combination of VPA with other data collection method is also recommended (Barkaoui, 2011; Gebril & Plakans, 2014). For example, think aloud protocols can be collected with a follow-up interview to gain multiple perspectives about rating behaviors (Gebril & Plakans, 2014).

Instructions:

The researcher should prepare clear instructions and pilot test them before use (Ericsson & Simon, 1993; Gass & Mackey, 2000; Green, 1998). The instructions should be standardized (Gass & Mackey, 2000) and should specify clearly that the participants should focus solely on completing the task given as this can ensure that they use the same thinking processes as when they do the task silently (Ericsson & Simon, 1993). In addition, the instructions should emphasize that participants are to report thoughts as they occur without trying to make their reports more coherent. As for retrospective reports, the instructions should tell the participants to start their retrospective reports with "I first thought of…" to help them recall their thoughts.

Some tasks involve automatic processes which are not stored in short-term memory, and therefore are not reportable. In order to make the processes reportable, the researcher should design the instructions so that the processing is slowed down (Ericsson & Simon, 1993). For example, in a reading task, participants may be required to pause between sentences to verbalize their thoughts before reading the next sentence (Ericsson & Simon, 1993). Or they may be asked to read a passage that has been marked; whenever they see a mark, they will stop reading and start thinking aloud (Crain-Thoreson et al., 1997). Another way to facilitate the participants in the case of a listening or speaking test is to pause the VDO when discourse boundaries occur (Wagner, 2008) or to segment a spoken text after 20-25 seconds (Li & Lorenzo-Dus, 2014).

As for the language used for verbal reports, participants should be allowed to use their first language (Kormos, 1998). Reporting in the SL/FL may be problematic because it may interfere with task performance. Also, the reports may not reflect the thought processes accurately if the participants are not proficient in that language.

Generally researchers should not tell participants what thinking processes they are interested in as this can influence the way participants verbalize (Ericsson & Simon, 1993; Pressley & Afflerbach, 1995). However, if the research objective is to investigate whether participants use a particular process or not, or is to understand how particular processes are used, then researchers can state specifically what processes they are focusing on to elicit the processes of interest (Pressley & Afflerbach, 1995; Weigle et al.,

2013). Researchers may remind the participants to focus on specific processes by repeating or bolding the key processes in instructions (Li & Lorenzo-Dus, 2014).

Practice:

After the instructions are given, participants should do some practice in giving the report to ensure that they understand the procedure and can perform as instructed. Participants should practice an easy and general task such as multiplying numbers or solving an anagram (Ericsson & Simon, 1993; Green, 1998). After these tasks are completed, participants should practice the task and the technique that the researcher aims to use in the study. For instance, participants may practice thinking aloud while reading short paragraphs before reading longer ones in the main data collection procedure (Weigle et al., 2013).

Verbal reporting:

During verbal reporting, the researcher should clarify the instructions if participants do not give the verbal report as instructed (Green, 1998). When participants pause, researchers should remind them to continue talking by saying "Keep talking" (Ericsson & Simon, 1993). If the participants say they do not remember their thoughts, it is suggested that researchers accept that answer and move on. Also, the researcher should not sit opposite or beside the participant as this may create social interaction which may cause changes in the sequence of thoughts in task performance (Ericsson & Simon, 1993).

With regards to collecting retrospective reports, the researcher may use supplementary data such as notes the participants wrote while doing a reading test or a video recording of their speaking task, to help participants retrieve their thought processes (Gass & Mackey, 2000; Green, 1998).

Data analysis procedure for verbal protocol analysis

The data analysis procedure consists of data transcription, coding and analysis.

Data transcription:

There are several recommendations for data transcription (Green, 1998). For example, recorded protocols should be transcribed as they are without any modification even though they may be incomplete or contain grammatical errors. Time markers as well as prosodic and paralinguistic elements should also be indicated in the transcripts as they are informative. For instance, time markers can show the length of time spent on a particular cognitive activity and pauses can be used to segment protocols to a single process.

Coding:

The next step is to develop a coding scheme and assign a code to each segment of protocol (Green, 1998). There are several ways to develop a coding scheme; one or more methods may be used in a study. For example, one can do task analysis as mentioned earlier (Ericsson & Simon, 1993; Green, 1998). Another method is to develop a coding scheme based on the protocols collected in one's study. The coding scheme can then be piloted with samples of data and refined afterwards (Cumming, 1990). Finally, one can use a coding scheme that has been developed by other researchers with or without modifications (Anderson et al., 1991; Barkaoui et al., 2013; Gebril & Plakans, 2014; Goh, 2002; Pressley & Afflerbach, 1995).

After a coding scheme is developed, verbal protocols can next be segmented and coded; each segment represents a single cognitive process. Therefore, segments may vary in length ranging from a single word to a phrase or paragraph (Li & Lorenzo-Dus, 2014; Xu & Wu, 2012).

After coding, the researcher should establish inter-coder reliability and intra-coder reliability (Green, 1998). A small random sample of data is usually selected to be coded by a second coder (inter-coder) or the same coder after the first coding (intra-coder). Then, reliability coefficients are calculated, for example, through percentage of agreement and Cohen's kappa (Anderson et al., 1991; Barkaoui et al., 2013; Bowles,

2010; Cumming, 1990; Gebril & Plakans, 2014; Green, 1998; Weigle et al., 2013; Zhang & Elder, 2014).

Analysis:

After the data are coded, they may be reported qualitatively or quantitatively or using a combination of both depending on the research questions or hypotheses (Gass & Mackey, 2000; Green, 1998). For example, verbal reports of rating processes can be analyzed and presented qualitatively (e.g. Rupp et al., 2006; Orr, 2002). Or codings may be tallied, and percentages or frequencies presented (e.g. Gebril & Plakans, 2014; Wagner, 2008; Weigle et al., 2013). In addition, statistical analyses can be performed. For example, *t*-tests can be conducted to find out whether groups of participants differ in their cognitive processes (e.g. Cumming, 1990; Sasaki, 2000). Or chi-square can be conducted to examine relationships between strategy use and other factors (e.g. Anderson et al., 1991). Correlational analyses can also be performed to investigate the relationships between reported strategies and other variables such as test scores (Barkaoui et al., 2013).

Concerns about verbal protocol analysis

Although VPA has been well accepted by many researchers, it has also been criticized in a number of areas. The major concerns about using VPA are in regards to reactivity and veridicality (Barkaoui, 2011; Ellis, 2001; Polio, 2012). Reactivity happens when concurrent verbal protocols affect the process of doing a task or the product of a performance. For example, thinking aloud while rating was found to affect the rating processes, and severity as well as self-consistency in scoring by some raters (Barkaoui, 2011). Reactivity may also be found in a verbal reporting procedure that requires participants to explain or describe while verbalizing or interpret their task performance (Fox, Ericsson & Best, 2011).

Another problem, veridicality, concerns not only concurrent but also retrospective verbal protocols. That is, the information obtained from concurrent verbal protocols can be limited since it is not possible for participants to verbalize every thought (Nisbett & Wilson, 1977). Some participants are more articulate than others in reporting

their thoughts, and some people may find it difficult to report thoughts while performing a task (Weigle, 1994). Similarly, retrospective reports may be incomplete because there is a delay between task performance and verbal reports (Bowles, 2010). In addition, participants giving retrospective protocols are supposed to report only what they were thinking while doing the task. However, they may also report thoughts that occur after the task has been completed.

Another challenge of the methodology is the interactive and social nature of protocol data (Barkaoui, 2011; Pressley & Afflerbach, 1995; Sasaki, 2008; Smagorinsky, 1989, 2001). These features can be seen in think aloud data in which the participants address the researchers despite their absence during the data collection (Barkaoui, 2011; Sasaki, 2008). This suggests that the content and types of verbal protocols may be influenced by the awareness of the audience. Therefore, researchers should also take this into consideration when collecting, analyzing and interpreting verbal reports (Barkaoui, 2011; Sasaki, 2008). In addition, involving an interviewer in data collection may affect the quality of thinking (Norris, 1990), and task performance and sophistication of retrospective reports (Leighton, 2013). Apart from the involvement of interviewers, item difficulty was also found to have a significant effect on the consistency of response processing in concurrent and retrospective reports (Leighton, 2013).

Other criticisms on the methodology are that it is labor and time intensive (Wolfe, 1997). This also leads to low statistical power when statistical analysis is applied to think aloud data due to small sample sizes.

This section has reviewed concerns and empirical studies of the validity of VPA in several fields. As the number of studies on its validity conducted in the SL/FL testing is not large and they address a variety of different issues, it is difficult to draw any useful conclusions in regards to SL/FL testing. Despite some criticism, for example, in terms of incompleteness and interactive features, a significant number of researchers feel that this does not invalidate the verbal protocols obtained (e.g. Barkaoui, 2011; Goh, 2002). VPA is regarded as a useful tool as it can reveal data on cognitive processes and strategies used by test takers and raters (e.g. Barkaoui, 2011; Crisp, 2008; Green, 1998; Leighton 2013). Therefore, guidelines for the data collection and analysis procedures that have

been presented earlier should be strictly followed to address these challenges and maximize the quality of verbal protocols.

Recommendations for further use of VPA in SL/FL testing

The literature has shown that VPA has been extensively used for SL/FL test validation purposes. However, the present author would like to point out that the methodology has potential in other areas of SL/FL testing. For example, it can be applied to validation of other assessment instruments such as self-assessment, which is a valuable tool in improving learning as it can provide learners with an understanding of their current ability and the target performance (Fulcher, 2010; Oscarson, 2014). As can be seen in the case of self-assessment, its validity has been investigated widely through the comparison of students' self-assessment with their test scores, teachers' ratings or peer assessment (e.g. Brantmeier, Vanderplank & Strube, 2012; Matsuno, 2009; Saito & Fujita, 2004).

However, little research thus far has employed VPA to explore this issue. Since VPA can elicit cognitive processes in test performance as demonstrated in earlier sections, the method then can play an important role in addressing the validity of self-assessment as well. For example, researchers may ask students to assess their English writing abilities by responding to a self-assessment questionnaire and to give retrospective reports about their cognitive processes while completing the questionnaire. The researchers can then compare the processes they aim the questionnaire will elicit with those reported by the students. The analysis can provide evidence about the validity of the self-assessment instrument.

Another potential use of VPA is in rater training. As previously discussed, studies that involve raters' think aloud or retrospective reports have shown what raters attend to when evaluating and assigning scores. If VPA is implemented during rater training sessions, raters may benefit from the knowledge of their own decision-making processes. They can become more aware of how well the strategies they use correspond with those that are specified in rating scales. In addition, they can compare their strategy use with others and discuss ways to avoid bias and improve rating consistency and accuracy, which in turn can improve the validity of test score interpretation and use.

Conclusion

In conclusion, VPA has been applied quite extensively in studying language test validation; however, the validity of the method itself has been criticized, for example, in terms of its completeness and social nature. Nevertheless, it is still considered to be of value in studying test taking processes and rating processes as it may be the only tool that can directly reveal these cognitive processes and strategies (e.g. Barkaoui, 2011; Crisp, 2008; Green, 1998; Leighton 2013). Researchers, however, must be aware of its limitations and, as suggested by Ericsson and Simon (1993), Gass and Mackey (2000) and Green (1998) to name a few, care should be taken in data collection as well as analysis in order to maximize its value in language testing studies.

References

Alavi, S. M. (2005). On the adequacy of verbal protocols in examining an underlying construct of a test. **Studies in Educational Evaluation**, 31(1), 1-26.

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

Ang-Aw, H. T. & Goh, C. C. M. (2011). Understanding discrepancies in rater judgment on national-level oral examination tasks. **RELC Journal**, 42(1), 31-51.

Babaii, E. & Moghaddam, M. (2006). On the interplay between test task difficulty and macro-level processing in the C-test. **System**, 34(4), 586-600.

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

Bachman, L. F. (2000). Learner-directed assessment in ESL. In G. Ekbatani & H. Pierson (Eds.), **Learner-directed assessment in ESL** (pp. ix-xii). New Jersey: Lawrence Erlbaum Associates, Inc.

Banerjee, J. & Luoma, S. (1997) Qualitative approaches to test validation. In C. Clapham & D. Corson (Eds.), **Encyclopedia of language and education, Volume 7: Language testing and assessment** (pp. 275-287)**.** Amsterdam: Kluwer.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. **Assessing Writing**, 12(2), 86-107.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. **Language Assessment Quarterly**, 7(1), 54-74.

Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. **Language Testing**, 28(1), 51–75.

Barkaoui, K., Brooks, L., Swain, M. & Lapkin, S. (2013). Test-takers' strategic behaviors in independent and integrated speaking tasks. **Applied Linguistics**, 34, 304-324.

Bax, S. (2013). The cognitive processing of candidates during reading tests: Evidence from eye-tracking. **Language Testing**, 30, 441-465.

Bowles, M. A. (2010). **The think-aloud controversy in second language research**. New York, NY: Routledge.

Brantmeier, C., Vanderplank, R., & Strube, M. (2012). What about me? Individual self-assessment by skill and level of language instruction. **System**, 40(1), 144-160.

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 Series, MS-29). Princeton, NJ: Educational Testing Service.

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

Camps, J. (2003). Concurrent and retrospective verbal reports as tools to better understand the role of attention in second language tasks. **International Journal of Applied Linguistics**, 13(2), 201–221.

Chapelle, C. A. (1999). Validity in language assessment. **Annual Review of Applied Linguistics**, 19, 254-272.

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

Cohen, A. D. (1984b). Studying second-language learning strategies: How do we get the information? **Applied linguistics**, 5(2), 101-112.

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

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

Cohen, A. D. & Hosenfeld, C. (1981). Some uses of mentalistic data in second-language research. **Language Learning**, 31(2), 285-313.

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

Connor-Linton, J. (1995). Looking behind the curtain: What do L2 composition ratings really mean? **TESOL Quarterly**, 29(4), 762-65.

Crain-Thoreson, C.; Lippman, M. Z. & McClendon-Magnuson, D. (1997). Windows on comprehension: Reading comprehension processes as revealed by two think-aloud procedures. **Journal of Educational Psychology**, 89(4), 579-591.

Crisp, V. (2008). The validity of using verbal protocol analysis to investigate the processes involved in examination marking. **Research in Education**, 79(1), 1-12.

Cumming, A. (1990) Expertise in evaluating second language compositions. **Language Testing**, 7, 31-51.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. **Modern Language Journal**, 86, 67–96.

DeRemer, M. L. (1998). Writing assessment: Raters' elaboration of the rating task. **Assessing writing**, 5(1), 7-29.

Ducasse, A. M., & Brown, A. (2009). Assessing paired orals: Raters' orientation to interaction. **Language Testing**, 26(3), 423-443.

Ellis, R. (2001). Introduction: Investigating forum-focused instructions. **Language Learning**, 51(Suppl. 1), 1-46.

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

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. **Psychological bulletin**, 137(2), 316-344.

Fulcher, G. (2010). **Practical language testing**. London: Hodder Education.

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

Gebril, A. & Plakans, L. (2014). Assembling validity evidence for assessing academic writing: Rater reactions to integrated tasks. **Assessing Writing**, 21, 56-73.

Goh, C. C.M. (2002). Exploring listening comprehension tactics and their interaction patterns. **System**, 30, 185-206.

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

Grotjahn, R. (1986). Test validation and cognitive psychology: Some methodological considerations. **Language Testing**, 3, 159-186.

Hosenfeld, C. (1984). Case studies of ninth grade readers. In J. C. Alderson & A. H. Urquhart (Eds.), **Reading in a foreign language** (pp. 231-249). London: Longman.

Isaacs, T. & Thomson, R. I. (2013). Rater experience, rating scale length, and judgments of L2 pronunciation: Revisiting research conventions. **Language Assessment Quarterly**, 10, 135-159.

Joe, J. N., Harmes, J. C., & Hickerson, C. A. (2011). Using verbal reports to explore rater perceptual processes in scoring: A mixed methods application to oral communication assessment. **Assessment in Education: Principles, Policy & Practice**, 18(3), 239-258.

Kormos, J. (1998). Verbal reports in L2 speech production research. **TESOL Quarterly**, 32(2), 353-358.

Lazaraton, A. (2008). Utilizing qualitative methods for assessment. In E. Shohamy & N. H. Hornberger (Eds.), **Encyclopedia of language and education, Volume 7: Language testing and assessment** (2nd ed.) (pp. 197-209). New York: Springer.

Lee, J. (2015). Language learner strategy by Chinese-speaking EFL readers when comprehending familiar and unfamiliar texts. **Reading in a Foreign Language**, 27(1), 71-95.

Leighton, J. P. (2013). Item difficulty and interviewer knowledge effects on the accuracy and consistency of examinee response processes in verbal reports. **Applied Measurement in Education**, 26(2), 136-157.

Li, H., & He, L. (2015). A comparison of EFL raters' essay-rating processes across two types of rating scales. **Language Assessment Quarterly**, 12(2), 178-212.

Li, H. & Lorenzo-Dus, N. (2014). Investigating how vocabulary is assessed in a narrative task through raters' verbal protocols. **System**, 46, 1-13.

Liu, J. (2007). Developing a pragmatics test for Chinese EFL learners. **Language Testing**, 24(3), 391-415.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. **Language Testing**, 26(1), 75–100.

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

Milanovic, M., Saville, N. & Shuhong, S. (1996). A study of the decision-making behavior of composition markers. In M. Milanovic & N. Saville, (Eds.), **Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem (Vol. 3)** (pp. 92-114). Cambridge: Cambridge University Press.

Nisbett, R. E. & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. **Psychological Review**, 84(3), 231-259.

Norris, S. P. (1990). Effects of eliciting verbal reports of thinking on critical thinking test performance. **Journal of Educational Measurement**, 27, 41-58.

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. **System,** 30(2), 143-154.

Oscarson, M. (2014). Self-assessment in the classroom. In A. Kunnan (Ed.), **The companion to language assessment, Volume. II: Approaches and development** (pp. 712–729). New York: Wiley-Blackwell.

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

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. **Assessing Writing**, 13, 111-129.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. **Journal of English for Academic Purposes**, 8(4), 252-266.

Plakans, L. & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. **Assessing Writing**, 17, 18-34.

Polio, C. (2012). How to research second language writing. In A. Mackey & S. Gass (Eds.), **Research methodologies in second language acquisition: A practical guide** (pp. 139-157). London: Blackwell.

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

Rupp, 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), 441-474.

Saito, H., & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. **Language Teaching Research**, 8(1), 31-54.

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In A. J. Kunnan (Ed.), **Fairness and validation in language assessment: Selected papers from the 19th Language Testing Research Colloquium, Orlando, Florida** (pp. 129-152). Cambridge: University of Cambridge Local Examinations Syndicate.

Sasaki, M. (2000). Effects of cultural schemata on students' test-taking processes for cloze tests: A multiple data source approach. **Language Testing**, 17(1), 85-114.

Sasaki, T. (2008). Concurrent think-aloud protocol as a socially situated construct. **IRAL**, 46, 349-374.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second language acquisition and language testing approaches. **System**, 45, 79-91.

Smagorinsky, P. (1989). The reliability and validity of protocol analysis. **Written Communication**, 6(4), 463-479.

Smagorinsky, P. (2001). Rethinking protocol analysis from a cultural perspective. **Annual Review of Applied Linguistics**, 21, 233-245.

Uiterwijk, H., & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. **Language Testing**, 22(2), 211-234.

Wagner, E. (2008). Video listening tests: What are they measuring? **Language Assessment Quarterly**, 5, 218-243.

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. **Language Testing**, 11(2), 197-223.

Weigle, S. C. (1999). Investigating rater/prompt interactions in writing assessment: Quantitative and qualitative approaches. **Assessing Writing**, 6(2), 145-178.

Weigle, S. C., Yang, W. & Montee, M. (2013). Exploring reading processes in an academic reading test using short-answer questions. **Language Assessment Quarterly**, 10(1), 28-48.

Wigglesworth, G. (2005). Current approaches to researching second language learner processes. **Annual Review of Applied Linguistics**, 25, 98–111.

Winke, P., & Gass, S. (2013). The influence of second language experience and accent familiarity on oral proficiency rating: A qualitative investigation. **TESOL Quarterly**, 47(4), 762-789.

Wiseman, C. S. (2012). Rater effects: Ego engagement in rater decision-making. **Assessing Writing**, 17(3), 150-173.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. **Assessing Writing**, 4(1), 83-106.

Xu, Y. & Wu, Z. (2012). Test-taking strategies for a high-stakes writing test: An exploratory study of 12 Chinese EFL learners. **Assessing Writing**, 17(3), 174-190.

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(1), 21-44.

Zhang, Y., & Elder, C. (2014). Investigating native and non-native English-speaking teacher raters' judgements of oral proficiency in the College English Test-Spoken English Test (CET-SET). **Assessment in Education: Principles, Policy & Practice**, 21(3), 306-325.

Zhao, C. G. (2012). Measuring authorial voice strength in L2 argumentative writing: The development and validation of an analytic rubric. **Language Testing**, 30(2), 201-230.

Biodata
Sutthirak Sapsirin received her B.A. (English) (second-class honors) from Chulalongkorn University, M.A. (Education) from the University of Kansas, USA, and Ph.D. (Language Assessment and Evaluation) from Chulalongkorn University. She is currently an instructor at Chulalongkorn University Language Institute. Her research interests include language assessment and test-taking strategies.