VIETNAM NATIONAL UNIVERSITY, HANOI
UNIVERSITY OF LANGUAGES AND INTERNATIONAL STUDIES
FACULTY OF POST-GRADUATE STUDIES

NGUYỄN THỊ MAI HỮU

AN INVESTIGATION INTO THE COGNITIVE VALIDITY OF THE SPEAKING SECTION OF THE VIETNAMESE STANDARDIZED TEST OF ENGLISH PROFICIENCY (VSTEP.3-5)

Nghiên cứu giá trị xác thực đối với quá trình tư duy của thí sinh khi thi phần thi nói bài thi chuẩn hóa đánh giá năng lực tiếng Anh từ bậc 3 đến bậc 5 theo Khung năng lực ngoại ngữ 6 bậc dùng cho Việt Nam (VSTEP.3-5)

Major: English Language Teaching Methodology

Code: 62220201

Supervisors: Professor Hoa Nguyen

Professor Fred Davidson

SUMMARY OF DOCTORAL DISSERTATION

HANOI, 2021

CHAPTER I: INTRODUCTION

1.1 Objectives of the study

The study is aimed at:

- outlining cognitive process underlying the speaking construct of the VSTEP.3-5, which can serve as a framework for establishing the cognitive validity of the VSTE.3-5 speaking test; and
- establishing the cognitive validity of the speaking section of the VSTEP.3-5.

For such aims, the research questions of the study are designed to establish the speaking cognitive validity of the test as a predictor of real-life performance basing on the central issues that a language test must deal with in terms of its cognitive validity (Field, 2013):

- **RQ1:** Does the VSTEP.3-5 speaking section actually cover the cognitive processes that it is supposed to represent?
- **RQ2:** To what degree are the cognitive demands imposed in the VSTEP.3-5 speaking section appropriately calibrated to reflect the levels of language competences of the test-takers?
- **RQ3:** How closely do the cognitive processes that the VSTEP.3-5 speaking section elicits from a candidate resemble the processes that he/she would employ in non-test conditions?

1.2. Organization of the study

The study is divided into 8 chapters as below:

Chapter 1: Introduction

Chapter 2: Literature Review

Chapter 3: The Vietnam Standardized Test of English Proficiency

Chapter 4: Methodology

Chapter 5: The cognitive processes supposedly represented in the VSTEP.3-5 speaking section

Chapter 6: The calibration of cognitive demands in the VSTEP.3-5 speaking rating scale

Chapter 7: The speaking cognitive processes in VSTEP.3-5 test and non-test conditions

Chapter 8: Conclusion

CHAPTER II: LITERATURE REVIEW

2.1. The concept of validity

The term *validity* is of concern of not only test users but researchers as well because it relates much to the quality of a test. The term *validity* has been coined for a long time.

According to Messick (1989), validity is "an overall evaluative judgement of the degree to which evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores and other modes of assessment" (S Messick, 1989, p. 13). The key concept is 'score meaning', which is defined as a "construction that makes theoretical sense out of both the performance regularities summarized by the score and its pattern of relationships with other variables, the psychometric literature views the fundamental issue as construct validity" (Samuel Messick, 1996, p. 245). Messick's view was important because it provided new understanding of validity as compared to the traditional definitions of validity.

Messick's views in language testing have been developed by Bachman (1990), who claims that the validity of a given use of test scores is the outcome of a complex process that must include "the analysis of the evidence supporting that interpretation or use, the ethical values which are the basis for the interpretation or use but also the test takers' performance" (Bachman, 1990, p. 237). Starting from Messick's "progressive matrix", Bachman focused on construct validity, and on the "value implications of interpreting the score in a particular way" by considering the theories of language and the relevant educational and social ideologies we attach to the score interpretation (Bachman 1990, p. 243). Bachman drew on Messick's theories and started from the analysis of the evidential "basis of validity", which he refers to as the gathering of complementary types of evidence into the process of validation to support the relationship between test score and interpretation and use. As far as the consequential basis of validity is concerned, Bachman argued that tests are not designed and used in a "value-free psychometric test-tube" but that they meet the needs of an educational system or of the whole society for

which we must assume the potential consequences of testing.

2.2. Validation in language testing

In the second edition of *Education Measurement*, Cronbach (1971) defined "validation is the process of examining the accuracy of a specific prediction or inference made from a test score… More broadly, validation examines the soundness of all interpretations of a test – descriptive and explanatory interpretations as well as situation-bound predictions" (Cronbach, 1971, p. 443).

In 1996, Messick stated that "test validation is empirical evaluation of the meaning and consequences of measurement, taking into account extraneous factors in the applied setting that might erode or promote the validity of local score interpretation and use" (Samuel Messick, 1996, p. 245).

Validation processes are developed and introduced by different scholars including Cronbach and Meehl (1955), Cronbach (1971), Messick (1989) in their publications quoted in many validation studies (Cronbach, 1971) (Cronbach & Meehl, 1955) (Loevinger, 1957) (S Messick, 1989).

In the 2000s, the socio-cognitive framework was introduced by Weir (2005) for test development and validation and gained much attention of scholars. The socio-cognitive framework helps researchers develop a transparent and coherent validity argument with systematic evidence, while at the same time addressing the interaction between different types of validity evidence. The framework comprises a number of components each of which must be attended to by the test developer at one or more points of the test development, implementation and validation cycle, namely test-takers' characteristics, cognitive validity, context validity, content validity, criterion-related validity, and consequential validity.

Among the validation models, the criterion model is more applicable for tests of which a criterion is available, the content model is often applied for achievement tests, the construct validity, unified model of validity and argument-based approach to validity are recently popular in test validation and applied by different scholars, the socio-cognitive framework offers a strong theoretical background for establishing the cognitive validity of language tests for the socio-cognitive framework aims to provide a comprehensive

validity argument for validation including different aspects of validity, the socio-cognitive validity framework provides detailed description of the cognitive processes that test-takers are supposed to experience in the test conditions of all four skill tests, one important aspect of the socio-cognitive framework is test-takers' characteristics. With all those features, the socio-cognitive framework is considered the first systematic attempt at a coherent approach to test development and validation, combining social, cognitive and evaluative dimensions of language use and linking these facets to the context and consequences of test use, specifically when it comes to the cognitive aspects relating to language tests. Also, with those features, the Weir's socio-cognitive framework is adopted as the validation model of this study.

2.3. The concept of cognitive validity

In language testing and assessment, the concept of cognitive validity was developed by Weir (2005) in his socio-cognitive approach to test validation. Cognitive validity, as described by Weir (2005), is similar to instructiveness as conceived by Bachman and Palmer (1996). This concerns the extent and type of involvement of a test taker's language ability, topical knowledge, and affective schemata in performing a language task (Bachman & Palmer, 1996). What is important is that the cognitive processing involved in real life language use should be reflected as far as possible in language test situations if claims for validity are to be supported. Cognitive validity is, according to Field (Field, 2013), of particular concern in the case of tests whose scores are employed predictively to indicate the test taker's suitability for a future university place, for a job in a particular domain as instances.

The evidences of cognitive validity are often collected through studying test-takers' behaviors using various types of verbal reporting (e.g., introspective, immediate retrospective, and delayed retrospective) to elicit their comments on what they actually do in a speaking (Field, 2011), listening (Field, 2013), reading (Khalifa & Weir, 2009), or writing test (Shaw, Shaw, & Weir, 2007).).

In general, according to Field (2013), cognitive validity also studies "how finely the relevant processes are graded across the levels of the suite in terms of the cognitive demands that they impose upon the candidate". The term "cognitive demand" mentioned here is a typical term in cognitive science, which refers to the demand which is placed on

cognitive abilities, through the dimensions of complexity, openness, implicitness, and level of abstraction (Edwards & Dall'Alba, 1981) or as stated by Stein (2009) as "the kind and level of thinking required of students in order to successfully engage with and solve a task" (Stein, Smith, Henningsen, & Silver, 2009). Also mentioned by Field (2013), cognitive validity considers the cognitive load that the test takers may encounter when taking a test. In this context, the term cognitive load refers to the levels of cognitive demands of processing information.

2.4 Cognitive validity and testing speaking

2.4.1. Assessing speaking

According to Louma, when assessing speaking, the elements of speaking ability should be studied in details. They are the sound of speech, grammar and spoken structures, vocabulary and spoken words, features of speech production, and functions of speaking (Luoma, 2004). The testing of pronunciation (both segmentals and suprasegmentals), spoken grammar, spoken vocabulary, and even sociolinguistic applications of speech all fall into the construct of speaking. These features are fundamental when designing and developing tests, which form the construct of the test.

Lado wrote "the ability to speak a foreign language is without doubt the most highly prized skill, and rightly so… Yet testing the ability to speak a foreign language is perhaps the least developed and the leased practiced in the language testing field." (Lado, 1961, p. 239)

Speaking is the verbal use of language to communicate with others. According to Fulcher, the construct of speaking includes sound features (pronunciation and intonation, accuracy and fluency), psychological aspect, speaking strategies (achievement and avoidance), structural aspects of speaking including opening and closing conversation and turn-taking, pragmatic aspects of speaking, vocabulary, and the co-construction of discourse (Fulcher, 2003).

Field (2004) developed the model of speech production following Levelt (1989, 1999) (Levelt, 1989; W. Levelt, 1999; W. J. Levelt, 1999) makes clear that any model of speech production, whether in L1 or L2 needs to incorporate a number of stages:

- A conceptual stage, where the proposition that is to be expressed first enters

the mind of the speaker

- A systematic stage, where the speaker chooses an appropriate frame into which words are to be inserted, and marks parts of it for plural, verb agreement, etc.
- A lexical stage, where a meaning-driven search of speaker's lexicon or vocabulary store takes place, supported by cues as to the form of the word (i.e. its first syllable)
- A phonological stage, where the abstract information assembled so far is converted into a speech-like form
- A phonetic stage, where features such as assimilation are introduced
- An articulation stage, in which the message is uttered

2.4.2. Cognitive validity in assessing speaking

Among the facets of validity mentioned in Weir's socio-cognitive framework is cognitive validity, according to Weir (2005), cognitive validity is established by a priori evidence on the cognitive processing activated by the test task before the live test event, as well as through the more traditional a posteriori evidence on constructs measured involving statistical analysis of scores following test administration. "Language test constructors need to be aware of the established theory relating to the cognitive processing that underpins equivalent operations in real-life language use" (Taylor, 2011). Based on the Weir's study of cognitive validity, Field (2011) adapted the Levelt's model of speech production (1999) to develop a cognitive validity framework for speaking.

The cognitive stages considered in the model are:

- Conceptualization: generating an idea or set of ideas for expression
- Grammatical encoding: constructing a syntactic frame and locating the lexical items that will be needed
- Phonological encoding: converting the abstract output of the previous stage into a string of words which are realized phonologically
- Phonetic encoding: adjusting the phonological sequence to make articulation easier; linking each of the syllables to a set of neutral instructions to the articulators; storing the instructions in a buffer while the clause is being articulated
- Articulation: producing the utterance

- Self-monitoring: focusing attention on the message immediately before and shortly after it is uttered in order to check for accuracy, clarity and appropriacy

CHAPTER III: THE VIETNAM STANDARDIZED TEST OF ENGLISH PROFICIENCY VSTEP.3-5

The VSTEP.3-5 was developed strictly following the 4 stages suggested in the Manual for language test development and examining introduced by the Europe Council in 2011 (de l'Europe, 2011) including planning, design, try-out and informing stakeholders. Products of the development stages are the VSTEP.3-5 format, VSTEP.3-5 specifications, VSTEP.3-5 item writer guidelines, VSTEP.3-5 raters' training manuals.

The general format of the VSTEP.3-5 is described in the below table:

Table 1: The VSTEP.3-5 test format[1]

Paper	Time allocation	Number of items/tasks	Item/task types
Listening comprehension	40 minutes including the time to transfer the answers to the answer sheet	3 parts, 35 MCQs	Test-takers listen to short conversations, instructions, notices, longer conversations, talks, and answer MCQs.
Reading comprehension	60 minutes including the time to transfer the answers to the answer sheet	4 passages, 40 MCQs	Test-takers read 4 passages about different topics with the difficulty levels of the passages varying from level 3 (B1) to level 5 (C1) and with the number of words ranging from 1900 to 2050 words and answer corresponding MCQs.
Writing	60 minutes	2 tasks	Task 1: Write an email of about 120 words, which account for one third of the total score of the Writing paper. Task 2: Write an essay of about 250 words about a given topic, developing the topic using specific arguments and examples. This task accounts for two third of the total score of the Writing paper.
Speaking	12 minutes	3 parts	Part 1: *Social Interaction* Test-takers answer from 3 to 6 questions about two different topics.

[1] Decision No. 729/QD-BGDDT dated March 11th 2015 issued by the Ministry of Education and Training of Vietnam

	Part 2: *Solution Discussion* Test-takers are provided with a situation and three options to deal with the issue raised in the situation. Test-takers gives arguments to support the option that they think is the best choice to deal with the issue and counter-arguments about the two other options. Part 3: *Topic development* Test-takers talk about a given topic using the suggested supporting ideas and/or their own ideas. Part three ends with some further questions to discuss the given topic.

With regards to speaking skill, the VSTEP.3-5 speaking section is described as below:

- Time length: 12 minutes (including 2 minutes of preparation: 1 minute in part 2 and 1 minute in part 3)
- General description: The speaking paper consist of three parts: (1) Social Interaction, (2) Solution Discussion, (3) Topic Development
- Output language: oral conversation (social interaction, discussion, questions and answers) and extended talk.
- Overall description: one-to-one speaking assessment model with

  - Part 1: Social Interaction (The examiner asks three to five questions, the test-taker answers the questions.)

    Topic 1: three questions

    Topic 2: three questions

  - Part 2: Solution Discussion (The examiner and test-taker discuss three options and select the best alternative.)

  - Part 3: Topic Development (The test-taker develops a topic, using a given outline.)

- Total number of tasks: 3
- Total score: 10 bands

  Cut-off scores:

  4.0 – 5.5: Level 3

  6.0 – 8.0: Level 4

  8.5 – 10: Level 5

The VSTEP.-3-5 specifications include two components: detailed descriptions of test items and tasks and sample items, tasks.

Regarding the detailed descriptions of the speaking tasks, three components mentioned are the language input and the test tasks. The language input includes the levels of difficulty of the input (vocabulary and structure difficulties), the content (the familiarity of speaking situations and topics). The test tasks are described in detail in terms of the language functions expected to be produced per each task, the task shells which describe the length (number of words) and ideas of sentences and questions, how the questions are formed and the scripts that the examiners should follow.

The other part of the VSTEP.3-5 specifications is the sample test. Below is an example of the speaking test section for test takers:

Part 1: Social Interaction (3')

Talk about the climate in your area.

- *What is the weather like in your area at this time of the year?*
- *Which season do you like the best? Why?*
- *Do you prefer to live in a cold region or hot region? Why?*

Talk about how your travelling experience.

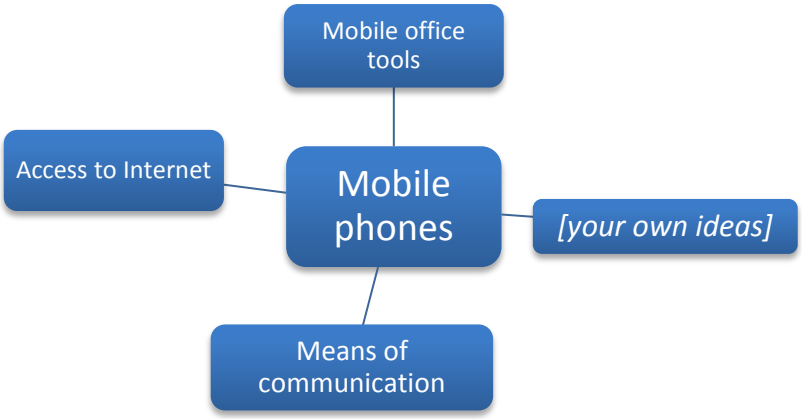
- *What was the last place you traveled to?*
- *Have you ever travelled alone?*
- *Which city in Vietnam do you like the most?*

Part 2: Solution Discussion (4')

Situation: You are having a birthday party and many of your friends are invited. Three locations are suggested: at home, in a restaurant, and in a karaoke bar. Which do you think is the best place for the party?

Part 3: Topic Development (5')

Topic: Mobile phones are useful tools at schools.



- *Do you think people will continue using mobile phones in the future?*

- *What negative effects do you think mobile phones have on young children?*
- *Do young people use mobile phones differently from old people in your country? How?*

CHAPTER IV: RESEARCH METHODOLOGY

4.1. Research questions

- **RQ1:** Does the VSTEP.3-5 speaking section actually cover the cognitive processes that it is supposed to represent?
- **RQ2:** To what degree are the cognitive demands imposed in the VSTEP.3-5 speaking section appropriately calibrated to reflect the levels of speaking competences of the test takers?
- **RQ3:** How closely do the cognitive processes that the VSTEP.3-5 speaking section elicits from a candidate resemble the processes that he/she would employ in non-test conditions?

4.2. Research methodology

4.2.1. Research methods

Mixed methods research, an emergent methodology increasingly used in linguistic studies in recent year, is the use of quantitative and qualitative methods in a single study or series of studies. To define, according to Creswell & Clark (2017), "mixed methods research is a research design with philosophical assumptions as well as methods of inquiry. As a methodology it involves philosophical assumptions that guide the direction of the collection and analysis of data and the mixture of qualitative and quantitative approaches in many phases of the research process. As a method, it focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone" (Creswell & Clark, 2017).

The mixed model of triangulation and embedded designs was applied for this study to capture the evidence of cognitive validity at the various stages of a testing cycle.

Firstly, via focus groups, which was conducted on almost 4 test developers, 5 item/task writers, 5 oral examiners, the cognitive demands imposed by the VSTEP.3-5 were

analyzed following the priori stages of the VSTEP testing cycle. Then, the scores and the levels of test fulfillment of 288 test-takers were analyzed to decide on the levels of cognitive demands calibrated in different language proficiency levels of the VSTEP.3-5 speaking rating scale. Questionnaires on those 288 test takers and stimulated-recall interviews on 30 test takers among the 288 ones were taken to study the cognitive processes elicited in the test and non-test conditions.

4.3. Data collection instruments

In order that the data collected were relevant for the study, the data collection instruments used are the VSTEP.3-5 test format and related documents, focus group interview questions, survey questionnaires, and stimulated recall interview questions.

Research questions	**Data collection instruments**
RQ1: Does the VSTEP.3-5 speaking section actually cover the cognitive processes that it is supposed to represent?	VSTEP.3-5 test development report, VSTEP.3-5 test format, VSTEP.3-5 test specifications and other related documents Focus groups
RQ2: To what degree are the cognitive demands imposed in the VSTEP.3-5 speaking section appropriately calibrated to reflect the levels of speaking competences of the test takers?	VSTEP.3-5 test scores
RQ3: How closely do the cognitive processes that the VSTEP.3-5 speaking section elicits from a candidate resemble the processes that he/she would employ in non-test conditions?	Survey questionnaires Stimulated recall sessions

4.4. Data analysis

The most important, and perhaps most difficult, aspect of mixed methods research is integrating the qualitative and quantitative data. According to Creswell & Clark (2017), one approach is to analyze the two data types separately and to then undertake a second stage of analysis where the data and findings from both studies are compared, contrasted and combined. The quantitative and qualitative data are kept analytically distinct and are

analyzed using techniques usually associated with that type of data; for example, statistical techniques could be used to analyze survey data whilst thematic analysis may be used to analyze interview data. In this approach, the integrity of each data is preserved whilst also capitalizing on the potential for enhanced understanding from combining the two data and sets of findings. Another approach to mixed methods data analysis is the integrative strategy. Rather than keeping the datasets separate, one type of data may be transformed into another type. That is to say that qualitative data may be turned into quantitative data or quantitative data may be converted into qualitative data. The former is probably the most common method of this type of integrated analysis. Quantitative transformation is achieved by the numerical coding of qualitative data to create variables that may relate to themes or constructs. These data can then be combined with the quantitative dataset and analyzed together.

The integrative strategy is the data analysis approach used for looking into the different types of data of the study hereof. Firstly, the qualitative data of focus group discussion were analyzed to cater for the cognitive validity evidence of the priori stages of the VSTEP.3-5 testing cycle. The results were later be triangulated with the empirical evidence of the results of 288 test takers taking the VSTEP.3-5 speaking section. The results of survey questionnaires and stimulated recall interviews were analyzed in connection to each other to compare and contrast the cognitive processes that the test takers experienced in the test contexts to what they did in real life situations. The results of all different types of data collection were then integrative for the findings of the study.

When analyzing the quantitative data including the test scores of 250 test takers and the survey data of these 288 persons, tools of EXCEL, SPSS, and FACETS were exploited to look into the data from descriptive statistics, correlations, t-test, ANOVA, Cronbach's alpha, factor analysis and Rasch and many-facet Rasch analysis. As for the qualitative data, the conversation analysis approach was applied with transcription of all the focus group discussions and stimulated recall interviews. With conversation analysis approach, the researcher shall adopt a radically emic approach to research, which avoids the use of secondary data, analyze prototypical examples of a particular phenomenon using different kinds of text-internal, convergent evidence to establish the credibility of an analysis, and seek to demonstrate that potential counterexamples have been anticipated

and encourage other researchers to replicate initial findings with different sources of data (Markee, 2000).

CHAPTER VIII: CONCLUSION

Chapters V, VI, VII present the data analysis, discussion and findings of the research, which are summarized in Chapter 8. Therefore, the contents of these three chapters are not presented in this summary of dissertation.

8.1. Summary of the findings

With the application of the Weir's socio cognitive framework for language test development and validation, the cognitive processes that the VSTEP.3-5 test takers are expected to experience, actually experienced in test and non-test conditions were investigated to provide validity evidence for the VSTEP.3-5 speaking section. The validity evidence covers three headings including the cognitive processes that the test is supposed to cover, the calibration of the cognitive demands across the different levels of the test, and the similarity between the processes in the test and in non-test conditions.

8.1.1. Cognitive processes that the VSTEP.3-5 test takers are supposed to experience

The cognitive processes that the test takers may encounter when taking the VSTEP.3-5 have been addressed at both the development and administration stages. Cognitive validity evidence was established for VSTEP.3-5 speaking section based on the model of speech production developed by Levelt (1989, 1999) and applied by Weir (2005) in the socio-cognitive model of test development and validation. The model falls into six major phases of processing: conceptualization, grammatical encoding, phonological encoding, phonetic encoding, articulation and self-monitoring.

8.1.1.1. Conceptualization

The conceptualization phase of cognitive processing in speech production in the VSTEP.3-5 speaking section shall be provided in two main headings of provision of ideas and Integrating utterances into a discourse framework.

Provision of ideas

The availability of information varies from level 3 to level 5 of the VSTEP.3-5 speaking test with the level of familiarity decreases from lower level of proficiency to higher level

of proficiency for a particular test taker, which ensures the level of cognitive load that the test taker of different levels of English proficiency may not affect much the production of the test taker per any particular target proficiency level.

The second aspect which may affect the provision of information to the VSTEP.3-5 test takers is how much support is provided by the test. It can be easily seen that a simple way of increasing the level of difficulty of a test is to reduce the support to the test takers as the level of the test increases. However, that is not the policy adopted in the speaking section of the VSTEP.3-5. The support to the VSTEP.3-5 test takers is provided to ensure comparability between the performances of candidates at a given level since the concepts and the areas of lexis upon which they draw are similar. Besides, the support is to make sure the cognitive load with regards to conceptualization does not affect the linguistic performance of the test takers.

The support to the VSTEP.3-5 test takers is observed in different ways. The first is the prompts, which range from questions of familiar topics (part 1) to more abstract topics (part 3), situations (part 2), topic and a mind map (part 3). The second factor which plays an important part in assisting conceptualization is whether the speaker is given time to preplan what to say (in terms of general ideas, of the links between those ideas or of the actual form of words to be used) or not. Considering the format of the VSTEP.3-5, the test takers have three minutes to prepare for their talk, the first is one minute to prepare for task 2 and the other is two minutes to prepare for part 3 of the test. Another factor that may affect the level of conceptualization is the level of difficulty of the language used to compose the three tasks of the VSTEP.3-5 speaking section. The level of difficulty of structures and vocabulary is of level 3 and lower to make sure that the test takers of proficiency level 3, level 4 or level 5 could understand the prompts and then can conceptualize the ideas. Or to express in other words, the prompts of the test are not hindering the test takers to produce utterances; only the vocabulary and words of the two additional questions of level 4 and level 5 could be of level 4 and level 5.

Integrating utterances into a discourse framework

Amongst the factors which Levelt identifies as affecting both macro- and microplanning are: awareness of the ongoing topic, thematization of new information, recognition of information shared and not shared with the listener, accommodation to the point- of- view

and even form of words of the interlocutor and certain basic principles which determine how information is ordered. The VSTEP.3-5 speaking section takes into account all those factors, which are presented in the interlocutor frame, and the rating scale with the criterium of "discourse management".

The VSTEP.3-5 model of speaking section of one to one with one assessor/interlocutor and one examinee. Though mentioned clearly in the construct of the test, interaction is observed only between the interlocutor and the examinee, which to some extent is not close to actual use of interactive language. One advantage of this type of testing model is that the test takers performance is not affected by the performance of his/her counterpart. Their performance is more affected by the co-operativeness of the interlocutor. Co-operativeness of interlocutor(s) refer to that a sympathetic interlocutor will facilitate successful communication by ceding a degree of control over the interaction to the user/learner, e.g. in negotiating and accepting modification of goals, and in facilitating comprehension, for example by responding positively to requests to speak more slowly, to repeat, to clarify. Features of speech of interlocutors are the characteristics of interlocutors' voice, e.g. rate, accent, clarity, coherence. Visibility of interlocutors refer to the accessibility of paralinguistic features in face to face communication facilitates communication. General and communicative competences of interlocutors, including behaviour refer to the degree of familiarity with norms in a particular speech community, and knowledge of the subject matter. Considering such issues that may cause different levels of cognitive demand imposing on the test takers, the VSTEP.3-5 speaking specifications were designed so as that the interlocutors use the preplanned language to different test-takers taking a same set of speaking tasks. Also mentioned in the VSTEP.3-5 speaking specifications, the interlocutors are not allowed to use the language other than what are introduced to them in the speaking interviewing scripts. The examiners' failure to apply the scripts may lead to different levels of cognitive demand imposing on the test takers.

8.1.1.2. Grammatical encoding

The linguistic content of speaking tests is often specified not in terms of grammatical structures but in terms of the language functions which test takers are required to perform. It can be treated as a question of mapping from the function that the test taker

wishes to perform to the pattern that best expresses that function. The issue under discussion when considering cognitive validity is not the complexity of the language that has to be retrieved but how easily the test taker is able to perform the mapping exercise. There are two ways in which the demands of mapping can be reduced in order to lighten the cognitive load upon lower level test takers with limited linguistic resources. One lies in restricting the number of functions that a test taker is expected to perform (particularly in relation to a single task). The other takes account of how accessible a given function is likely to be. Functions that are familiar, frequent and concrete will clearly be mapped more rapidly and reliably than those which are not.

Considering the VSTEP.3-5, the level of cognitive demand varies among the three tasks. The variety shown in the number of functions required and the accessibility of the functions mentioned in the test specifications. The number of functions vary among the three tasks. Among which, task 3 accounts for the most number of functions. The progression of cognitive demand can be clearly seen in the three tasks of the speaking paper. It can be seen that the level of cognitive demand increases from task 1 to task 3 of the VSTEP speaking paper. There is also a well-designed gradient of functions from task 1 to task 3 in terms of the demand imposed upon the test taker: moving from giving descriptions of social interactions to presenting a topic of higher level of complicacy and abstract with the highest level of abstract of the follow-up question number 3.

8.1.1.3. Phonological encoding

A part of the cognitive processing in speaking is the phonological encoding, which is identified to include the following characteristics which are contributing to the second language speaking fluency: use of preassembled chunking, length of uninterrupted speech, duration of planning pauses at syntactic boundaries, frequency of hesitation pauses, and ease of retrieving and assembling words.

Regarding preassembled chunks, chunking mentioned in the VSTEP.3-5 in the form of idiomatic expressions and colloquialisms; and is mentioned in the descriptors of level 5 of the rating scale (Vocabulary criterion). The VSTEP.3-5 test taker has progressed to certain level of language proficiency where they have established a repertoire of words which can be retrieved with minimal effort. However, the idiomatic expressions and colloquialisms are mentioned in the VSTEP speaking rating scale as a part of Vocabulary

criterion, not as a part of Fluency criterion. According to Field (2011), the chunking not just includes idiomatic expressions and colloquialisms, but includes other types of formulaic language as well.

Referring to the construct of the test, it can be seen that colloquialisms are not mentioned in the descriptions of different levels of proficiency. The use of colloquialisms is only mentioned at C2 level of proficiency of the CEFR. With regards to the other levels of proficiency mentioned in the construct of the VSTEP.3-5 speaking, the formulaic expressions are mentioned from level 1 to level 5. With level 1, it is the "very short, isolated, mainly pre-packaged utterances"; with level 2, it is "basic sentence patterns with memorized phrases, groups of few words and formulae"; with level 3, it is "a repertoire of frequently used routines". Though mentioned in the construct of the test, the above descriptors are not part of the VSTEP.3-5 speaking scale. The matter is perhaps partly that formulaic language can be both a negative and a positive indicator so far as an assessor is concerned. On the one hand, it can indicate dependence upon a limited range of highly conventionalized formulae, some of them rote learned; on the other (as previously discussed) it can show that the test taker has progressed as a speaker to a point where they have established a repertoire of word strings which can be produced with minimal effort.

Another factor that may affect fluency is length of uninterrupted speech. The level of cognitive demand is influenced greatly from the required length of speech. The VSTEP.3-5 specifications show that the required length of speech varies among the three parts of the speaking paper. Also, the descriptors of the VSTEP.3-5 rating scale show the construct of the VSTEP.3-5 with regards to length of uninterrupted speech. As can be clearly seen in the rating scale, the descriptors of band 2 to band 8 mention extended responses and extended stretches of language; however, with band 9 and band 10, the matter of length of speech is mentioned in form of pauses to show that the test taker of this level of proficiency in general speak at length.

Duration of planning pauses at syntactic boundaries and frequency of hesitation pauses are described as part of the phonological encoding as well. In the VSTEP.3-5 speaking rating scale, such characteristics are also described. As can be observed in the VSTEP.3-5 speaking rating scale, pauses and hesitation are mentioned from lower levels of

proficiency to higher levels of the scale. The examiners should identify the difference between the pauses and hesitations because of limited linguistic repertoire of because of test takers' searching for appropriate and accurate words and phrases to be used.

8.1.1.4. Phonetic encoding, articulation

Levelt (1989) pointed out that a person speaking a language that is not his/her mother tongue may be influenced by the first language in terms of a set of phonological representation in the mind which serve to define the phoneme values of the first language and a set of highly automatic process, attuned to the articulatory settings of the first language and the movements which link one to another. The accommodation of these two elements to the unfamiliar values of the target language forms the basis of any cognitive account of how L2 articulation is acquired. One should not lose sight of the fact that poor L2 pronunciation does not result solely from an inability to form the target sounds but also from an inadequate representation of those sounds in the mind and/or the inability to communicate the appropriate signals to the articulators. This explains well the presentation of requirement toward the intelligibility of the VSTEP.3-5 test takers' pronunciation from unintelligibility to intelligibility with the level of being intelligible to a native speaker increases descriptors of lower level of proficiency to higher level of proficiency. The descriptors of the VSTEP.3-5 show a gradient of elements which reflect the influence of the stored phoneme values of the first language of the test takers and the automatized articulatory processes that are associated with them.

8.1.1.5. Self-monitoring

Self-monitoring and self-repair are aspects of speaker's performance which are difficult to capture in the form test specifications. As mentioned by Field (2011), the nature of monitoring and repair are limited to the lower levels of proficiency of the L2 speakers and higher level of competence could be characterized by evidence of monitoring for pragmatic effectiveness as well as for linguistic accuracy.

Relating test takers' self-monitoring and repair, the certain descriptors are found in the VSTEP.3-5 specifications in the rating scale. It can be seen from those descriptors of the VSTEP.3-5 rating scale that the level of error correction decreases from lower level of proficiency to higher level of proficiency. The evidence of monitoring for pragmatic

effectiveness for linguistic accuracy is also apparent in the VSTEP.3-5 rating scale in the descriptors of vocabulary.

In a nutshell, though the Weir's socio cognitive model was not the theoretical framework applied when the VSTEP.3-5 speaking section was developed, the cognitive processes of the model provides a good framework for cognitive validity study of the test. All the patterns described in the model could be found in the VSTEP.3-5 speaking section, showing the effort of the VSTEP.3-5 development team in dealing with the cognitive demands that the test takers may encounter when taking the test. Besides, when applying the model to establish cognitive validity evidence for the speaking section, several issues have been identified including:

(1) The use of colloquialisms is only mentioned at C2 level of proficiency of the CEFR and the CEFR-VN; however, they are found in the descriptors of proficiency bands 9 and 10 of the VSTEP.3-5 speaking rating scale.
(2) Pauses and hesitation are found in the descriptors of almost all the bands of the VSTEP.3-5 rating scale; however, no further explanation is remarked of the difference of such pauses and hesitation across the bands.
(3) The interlocutor and assessor's cognitive load when interviewing and assessing the test takers was described in the CEFR and so the CEFR-VN, the theoretical framework applied when designing the VSTEP.3-5 speaking test; however, such load is not mentioned in the specifications or the examiner training manuals.

8.1.2. Calibration of cognitive demands across different levels of the VSTEP.3-5 speaking test

In general, the descriptors of the VSTEP.3-5 speaking rating scale are arranged properly with increasing level of difficulty from band 1 to band 10, except for some of the descriptors as below:

(1) When running the Rating Scale Model of Rasch analysis, the bands of different criteria of the rating scale should be placed more concern;, including Vocabulary bands 4,5, 9,10 and Fluency bands 6 and 7. No significant difference found among the logits arranged for the other descriptors of the same bands of different criteria.

(2) When running t-test on the difference between the speaking scores and the overall scores under different groupings of students of different levels of proficiency, significant difference found between the speaking scores and the overall scores of students of C1 level. The speaking score bands of 9 and 10 of all speaking criteria should be studied and the oral examiners should be informed of this pattern of the speaking scores.

(3) Arranging the descriptor difficulty on a same chart, certain descriptors stand out, among those are grammar bands 3, 4, 6, 7; vocabulary bands 4, 5, 9, 10; discourse management band 3, 9, 10; pronunciation bands 7, 8; fluency bands 6, 7. These are the descriptors with difficulty levels arranged into special positions as compared to the other descriptors of the same bands.

On the whole, the descriptors which are adjacent to each other and correspond to the same level of proficiency of the test takers are of relatively same level of difficulty. Besides, when studying the descriptors carefully, it seems to be difficult to see the difference between the performance of the test takers that correspond to those adjacent bands of a same criterium of the scale. Another matter identified is the descriptors of Vocabulary bands 9 and 10 include requirement about the ability to use idiomatic expressions and colloquialisms, which is not mentioned in the CEFR or CEFR-VN. Then, it is noticeable that bands 9 and 10 of all the criteria seem to be placed higher level of cognitive demand as compared to all the other bands. Some of the bands such as the ones for Vocabulary describe the performance that is not mentioned in the construct of the test; however, the same feature has not been observed in the other bands.

8.1.3. Similarities between cognitive processes in VSTEP.3-5 speaking test and non-test conditions

The survey questionnaires and stimulated recall interviews' results showed that

(1) The test takers experienced all the stages of cognitive processes when performing in the test condition. All five stages of the processes are observed including conceptualization, phonological encoding, grammatical encoding, phonetic encoding and articulation and self-monitoring.

(2) More than half of the VSTEP.3-5 test takers experienced similar speaking cognitive processes in the test and non-test conditions. All five stages of the

processes are observed including conceptualization, phonological encoding, grammatical encoding, phonetic encoding and articulation and self-monitoring.

(3) As for those who said that the processes were not the same thought that they would have performed better than what they performed on the day. They claimed that the information about VSEP.3-5 speaking section including the speaking rating scale and the same tests were of limited access to them. The test takers who prepared better for the test tend to get higher scores than those who did not.

(4) One situation of Part 2 of the test seems to be difficult to one group of test takers as compared to the other groups. This suggests that the test forms should be piloted more carefully.

8.2. Suggestions to VSTEP.3-5 speaking section administration, new language test development and validation, and the Weir's socio cognitive framework

8.2.1. VSTEP.3-5 development and administration

Though the cognitive processes that the test takers may encounter when taking the VSTEP.3-5 have been addressed at both the development and administration stages of the test, certain issues have been identified including:

(1) The use of colloquialisms is only mentioned at C2 level of proficiency of the CEFR and the CEFR-VN; however, they are found in the descriptors of proficiency bands 9 and 10 of the VSTEP.3-5 speaking rating scale. In order that the examinees are well aware of such inclusion, such information should be included in the oral examiners' training manual. Another way to deal with such descriptors is to remove them from the rating scale. In order to do so the test specifications of the VSTEP.3-5 test should be amended, which leads to the amendment of the Decision No. 729/QD-BGDDT dated March 11th 2015 issued by the Ministry of Education and Training of Vietnam.

(2) Pauses and hesitation are found in the descriptors of almost all the bands of the VSTEP.3-5 rating scale; however, no further explanation is remarked of the difference of such pauses and hesitation across the bands. Thus, further explanation should be included in the oral examiners' training manual so as that

the examiners are not confused of the signals when grading the performance of the test takers.

(3) The interlocutor and assessor's cognitive load when interviewing and assessing the test takers was described in the CEFR and so the CEFR-VN, the theoretical framework applied when designing the VSTEP.3-5 speaking test; however, such load is not mentioned in the specifications or the oral examiner's training manuals. The assessor's cognitive load when interviewing and assessing the test takers may lead to the fairness of their interviewing and grading work, and so affect the performance of the test takers and scores of them. It is highly recommended that the factors that affect the interlocutor and assessor's cognitive load interviewing and assessing the VSTEP.3-5 test takers should be included in the oral examiners training manual so that the interlocutors/assessors are aware of such issues when interviewing and rating. One particular suggestion is that the interlocutors/assessors are not provided with time to do grading work between adjacent test takers. They would find it to be practical when around one minute should be provided between two adjacent test takers so that they can complete the grading of the test taker who performed.

(4) The descriptors which are adjacent to each other and correspond to the same level of proficiency of the test takers are of relatively same level of difficulty. Besides, when studying the descriptors carefully, it seems to be difficult to see the difference between the performance of the test takers that correspond to those adjacent bands of a same criterium of the scale. Such issue may hinder the oral examiners from putting the test takers into the correct bands of the rating scale. In order that the examiners could precisely put the test takers' performance on the correct bands of the scale, quantified features of the bands should be included in the oral examiners' training manual or sample performance of the test takers in the forms of audio and/or videos should be developed for all the bands of the rating scale. Another way to deal with such situation is to merge the descriptors of the bands which correspond to a same level of English proficiency, in so doing, the number of bands of the VSTEP.3-5 rating scale will reduce to 5 bands. This will simplify the grading work of the examiners and will probably lead to better reliability of the work.

(5) All bands 9 and 10 are of significantly higher level of difficulty as compared to all bands 8 of the VSTEP.3-5 rating scale, which probably led to the low percentage of test takers who were graded C1 level as compared to other proficiency levels gauged by the test. In order that the speaking scores are more reliable for those test takers, the examiners should be provided with such data when they are trained. One possible reason for such significantly high level of difficulty of those bands is the examiners reluctantly gave highest scores to the test takers. In order to deal with this issue, it is recommended that the highest band of the rating scale should reflect the performance of the CEFR C2 level of proficiency, by so doing, the examiners would clearly see the expectation of CEFR C2 level of proficiency and better grade the CEFR C1 level performance.

(6) The test takers claimed that the information about VSEP.3-5 speaking section including the speaking rating scale and the sample tests were of limited access to them. And the test takers who prepared better for the test tend to get higher scores than those who did not. It is recommended that VSTEP.3-5 administration organizations should provide test takers with access to not only the test format, but also the rating scale and sample test forms so that the prospective test takers understand well what are expected of them to perform at different levels of proficiency, and so they could be better prepared for the test.

(7) One situation of Part 2 of the test seems to be difficult to one group of test takers as compared to the other groups. This suggests that the test forms should be piloted more carefully. It's highly recommended that all test forms should be developed strictly following the 12 stages of test development provided by the Vietnam Ministry of Education and Training, with which, each test task should be pretested twice. Besides, in order that the pretest scores are analyzed properly, specifically for speaking and writing sections, applications like FACETS and R, on which the Rating Scale Model and/or the Partial Credit Model can be run, should be used for test scores analysis.

8.2.2. Language test development and validation

From the case of the VSTEP.3-5 cognitive validity study, it is proved that the Weir socio cognitive model provides a comprehensive framework for language test

development and validation. Specifically for the validity of a certain aspect of a language test like cognitive validity, the Weir's socio cognitive model is approvingly applicable for it not only helps develop systematic validity argument for a particular language test but helps identify problem that face such studied language test at both the development and administration stages of the test. Thus, it is highly recommended that Weir's socio cognitive model should be applied when a new language test is developed and validated.

8.2.3. Weir's socio cognitive framework

The Weir's socio cognitive framework proves to be of highly applicable to language test development and validation, yet when applying the framework in validating a foreign language test like VSTEP.3-5, the specific features of learning English as a foreign language should be studied carefully to better apply the framework. In case of cognitive validity, the cognitive processes that the test taker may go through when taking the test follow Levelt's cognitive processes of spoken production for L1 speakers (1988, 1989) and Weir's adopted cognitive processes of spoken production for L2 speakers (2005) should be are almost the same. Nevertheless, when investigating the cognitive validity of the VSTEP.3-5 speaking section, the need to develop an updated cognitive model for oral production of foreign language speakers is arisen following the fact that a number of the VSTEP.3-5 studied test takers think in their L1 Vietnamese, which to certain extent affected their speaking performance in the test condition and probably in the non-test conditions as well, specifically in the cases of the test takers of CEFR B1 and lower proficiency.

8.3. Limitations and further studies

First, tt is recommended by the researcher hereof that Decision No. 729/QD-BGDDT dated March 11th 2015 issued by the Ministry of Education and Training of Vietnam should be amended with regards to the specifications of the VSTEP.3-5 test. In order to convince researchers and policy makers to make such a change to the Decision, further studies should be conducted to place stronger evidence on the urge to make amendment to the test specifications of the test and which part of the specifications should be amended as well.

Second, the study herein was accomplished on a small data collection size (288 test takers). In order that the data collected are of higher validity and reliability levels, similar studies with bigger data collection size are highly recommended to be conducted.

Third, for the moment, the Weir's socio cognitive framework proves to be of highly applicable to conduct cognitive validity studies; however, the model of cognitive processing in oral production was only developed for L1 and L2 speakers. A model that takes into consideration the features of foreign language acquisition should be developed to provide better impeccable model for establishing the cognitive validity of a foreign language test, specifically when the test is designed for assessment of test takers of CEFR B1 and lower levels of proficiency. Thus, prospective studies could be conducted to develop speaking cognitive processing model for foreign language speakers, which could be placed feasible by adopting the Levelt's model for L1 speakers or such a model could be newly developed for the fact that by far Levelt's model of L1 speaking cognitive processing is the only model applicable in language test development and validation.

To end with, further critical comments and contribution to the study herein are highly appreciated.

THESIS RELATED PUBLICACTIONS

1. Nguyen Thi Mai Huu (2019). *Stimulated recall – A practical data collection technique for cognitive studies in language testing.* Journal of Foreign Languages Studies, 58/2019, 38-50.

2. Nguyen Thi Mai Huu (2020). *Application of the socio-cognitive framework in language test development and validation*. 2020 International graduate research symposium ND 10TH East Asia Chinese teaching forum, Volume 1, 727-735.

3. Phuong Tran, Hoa Nguyen, Trang Dang, Minh Nguyen, Lan Nguyen, Tuan Huynh, Ha Do, Huu Nguyen, Fred Davidson (2015). *A validation study on the newly-developed Vietnam standardized English proficiency test*. 37th Language Testing Research Colloquium, 142.

4. Victoria Clark, Nguyen Thi Mai Huu, Jessica Wu, Jamie Dunlea, Richard West (2016). *Test centers and standardized Testing: Challenges, issues and benefits*. 4th British Council New Directions in English language assessment conference, 19-20.

5. Nathan Carr, Quynh Nguyen, Huu Nguyen, Yen Nguyen, Thao Nguyen (2016). *Systematic support for a communicative standardized proficiency test in Vietnam.* 4th British Council New Directions in English language assessment conference, 26-27.

6. Huu Nguyen (2018). *An Investigation into the Cognitive Validity of the Speaking Section of the Vietnam's Standardized Test of English Proficiency*. 40th Language Testing Research Colloquium, 114.

7. Huu Nguyen (2019). *An investigation into the cognitive processes reflected in the VSTEP speaking scoring rubrics*. 7th British Council New Directions in English language assessment conference, 22.

8. Barry O'Sullivan, Josep Lo Bianco, Huu Nguyen, Mitsuharu Ota, Yoshinori Watanabe (2019). *Language Assesment Policy*. 7th British Council New Directions in English language assessment conference, 18.