THE UNIVERSITY OF DANANG
UNIVERSITY OF FOREIGN LANGUAGE STUDIES

VÕ THỊ THU HIỀN

DEVELOPING A VALIDITY ARGUMENT FOR THE ENGLISH PLACEMENT TEST AT BTEC INTERNATIONAL COLLEGE DANANG CAMPUS

Major: ENGLISH LANGUAGE
Code: 822.02.01

MASTER THESIS IN
LINGUISTICS AND CULTURAL STUDIES
OF FOREIGN COUNTRIES
(A SUMMARY)

Da Nang, 2020

This thesis has been completed at University of Foreign Language Studies,
The University of Da Nang

Supervisor: Võ Thanh Sơn Ca Ph.D.

Examiner 1: Assoc. Prof. Dr. Phạm Thị Hồng Nhung

Examiner 2: Nguyễn Thị Thu Hương Ph.D.

The thesis was orally defended at the Examining Committee
Time: July 3th, 2020
Venue: University of Foreign Language Studies -The University of Da Nang

This thesis is available for the purpose of reference at:

- *Library of University of Foreign Language Studies, The University of Da Nang.*
- *The Center for Learning Information Resources and Communication - The University of Da Nang.*

CHAPTER 1
INTRODUCTION

This chapter presents the introduction to test validity and the purpose of this thesis. The chapter concludes with the significance of this thesis.

1.1 Introduction to Test Validity

Language tests are needed to measure students' ability in English in college settings. One of the most common tests developed is entrance tests or placement tests which are used to place students into appropriate language courses. Thus, the use of test scores cannot be denied as a very important role. The placement test at BTEC International College is used as an example for building this research study and helping to build up validity argument with further research purposes.

Test validity is the extent to which a test accurately measures what it is supposed to be measure and validity refers to the interpretations of test score entailed by proposed uses of tests which is supported by evidence and theory (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999).

1.2 The study

BTEC International College – FPT University administers its placement test (PT) every semester to incoming students to measure their English proficiency for university studies. The test is composed of four skills: reading, listening, speaking, and writing. Only writing skill is the focus of this study.

This study developed a validity argument for the English Placement Writing test (EPT W) at BTEC International College – FPT

University. Developed and first administered in Summer 2019, the EPT W is intended to measure test takers' writing skills necessary for success in academic contexts. (See Table 1.1 for the structure of the EPT W.) Therefore, building a validity argument for this test is very important. It is helpful for many educators and researchers to understand the consequences of assessment. Particularly, this study investigated: 1) the extent to which tasks and raters attributed to score variability; 2) how many tasks and raters are needed to get involved in assessment to obtain the test score dependability of at least .85; and 3) the extent to which vocabulary distributions are different across proficiency levels of academic writing.

Table 1.1 The structure of the EPT W

Total test time	30 minutes
Number of parts	2
Part 1	
Total time	15 minutes
Task content	Write a paragraph using one tense on any familiar topics. For example: Write a paragraph (100-120 words) to describe an event you attended recently.
Part 2	
Total time	15 times
Task content	Write a paragraph using more than one tense on a topic that relates to publicity. For example: Write a paragraph (100-120 words) to describe a vacation trip from your childhood. Using these clues:

	Where did you go? When did you go? Who did you go with? What did you do? What is the most memorable thing? Etc.

The EPT W uses a rating rubric to assess test takers' performance. The appropriateness of a response is based on a list of criteria, such as task achievement, grammatical range and accuracy, lexical resource, coherence and cohesion.

1.3 Significance of the Study

The results of the study should contribute theoretically to the field of language assessment. By providing evidence to support inferences based on the scores of the EPT W test, this current study attempts to provide the discussion of test validity in the context of academic writing.

Practically, the results should contribute to the possible use of quantity of tasks and raters to assess writing ability. The findings of this study should provide an understanding of how different components affect variability of test scores and the kind of language elicited. This would offer guidance on choosing an appropriate task for measuring academic writing.

CHAPTER 2
LITERATURE REVIEW

This chapter discusses previous studies on validity and introduces generalizability theory (G-theory) that was used as background for data analyses.

2.1 Studies on Validity discussion

2.1.1 The conception of validity in language testing and assessment

What is validity?

The definition of validity in language testing and assessment could be given in three main time periods.

Different aspects of validity

Both Bachman (1990) and Brown (1996) agreed on the three main aspects of validity: content relevance and content coverage (or content validity), criterion relatedness (or criterion validity), and meaningfulness of construct (or construct validity).

2.1.2 Using interpretative argument in examining validity in language testing and assessment

The argument-based validation approach in language testing and assessment views validity as an argument construed by an analysis of theoretical and empirical evidences instead of a collection of separately quantitative or qualitative evidences (Bachman, 1990; Chapelle, 1999; Chapelle, Enright, & Jamieson, 2008, 2010; Kane, 1992, 2001, 2002; Mislevy, 2003). One of the widely-supported argument-based validation frameworks is to use the concept of interpretative argument (Kane, 1992; 2001; 2002). Figure 2.1 shows the inferences in the interpretative argument.

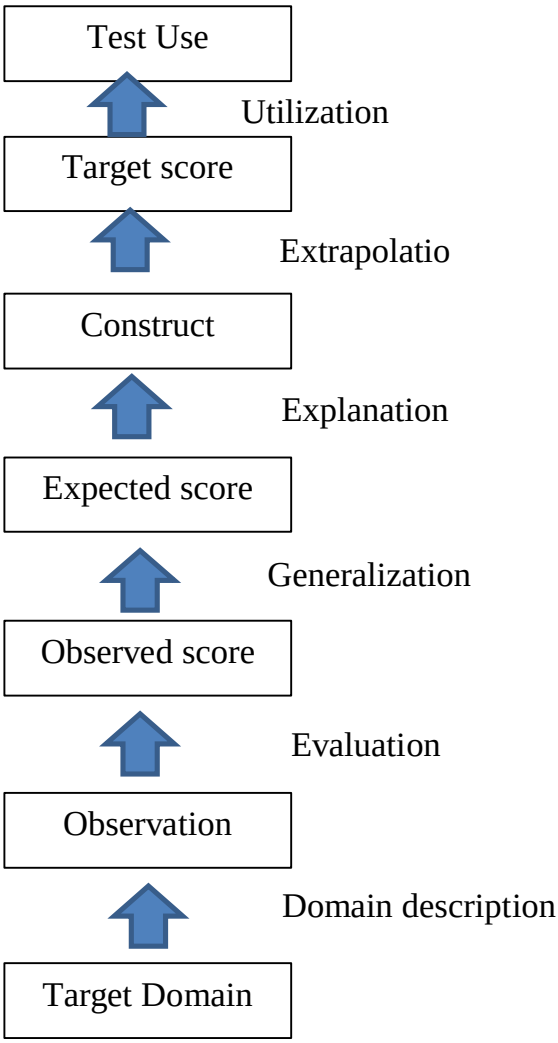

Figure 2.1 An illustration of inferences in the interpretative argument (adapted from Chapelle et al. 2008)

Structure of an interpretative argument

Kane (1992) argued that multiple types of inferences connect observations and conclusions. The idea of multiple inferences in a chain of inferences and implications is consistent with Toulmin, Rieke, and Janik's (1984) observation:

Kane et al. (1999) illustrated an interpretive argument that might underlie a performance assessment. It consists of six types of inferential bridges. These bridges are crossed when an observation of performance on a test is interpreted as a sample of performance in a context beyond the text. The Figure 2.2 shows the illustration of inferences in the interpretive argument.

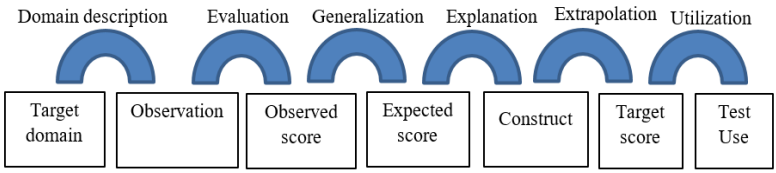

Figure 2.2 Bridges that represent inferences linking components in performance assessment (adapted from Kane et al., 1999)

2.1.3 The argument-based validation approach in practice so far

Chapelle et al. (2008) employed and systematically developed Kane's conceptualization about an interpretative argument in order to build a validity argument for the TOEFL iBT test.

The main components of the interpretative argument and the validity argument are illustrated in Table 2.1 and Figure 2.3 respectively.

Table 2.1 Summary of the inferences, warrants in the TOEFL validity argument with their underlying assumptions (Chapelle et al., 2010, p.7)

Inference	Warrant Licensing the Inference	Assumptions Underlying Inferences

Domain description	Observations of performance on the TOEFL reveal relevant knowledge, skills, and abilities in situations representative of those in the target domain of language use in the English-medium institutions of higher education.	1. Critical English language skills, knowledge, and processes needed for study in English-medium colleges and universities can be identified. 2. Assessment tasks that require important skills and are representative of the academic domain can be simulated.
Evaluation	Observations of performance on TOEFL tasks are evaluated to provide observed scores reflective of targeted language anilities.	1. Rubrics for scoring responses are appropriate for providing evidence of targeted language abilities. 2. Task administration conditions are appropriate for providing evidence of targeted language abilities. 3. The statistical characteristics of items, measures, and test forms are appropriate for norm-referenced decisions.
Generalization	Observed scores are	1. A sufficient number of

	estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters.	tasks are included in the test to provide stable estimates of test takers' performances. 2. Configuration of tasks on measures is appropriate for intended interpretation. 3. Appropriate scaling and equating procedures for test scores are used. 4. Task and test specifications are well defined so that parallel tasks and test forms are created.
Explanation	Expected scores are attributed to a construct of academic language proficiency.	1. The linguistic knowledge, processes, and strategies required to successfully complete tasks vary across tasks in keeping with theoretical expectations. 2. Task difficulty is systematically influenced by task characteristics. 3. Performance on new test measures relates to

		performance on other test-based measures of language proficiency as expected theoretically. 4. The internal structure of the test scores is consistent with a theoretical view of language proficiency as a number of highly interrelated components. 5. Test performance varies according to the amount and quality of experience in learning English.
Extrapolation	The construct of academic language proficiency as assessed by TOEFL accounts for the quality of linguistic performance in English-medium institutions of higher education.	Performance on the test is related to other criteria of language proficiency in the academic context.
Utilization	Estimates of the quality of performance in the English-medium	1. The meaning of test scores is clearly interpretable by

	institutions of higher education obtained from the TOEFL are useful and appropriate curricula for test takers.	admissions officers, test takers, and teachers. 2. The test will have a positive influence on how English is taught.

2.1.4 English placement test (EPT) in language testing and assessment

What is EPT?

Placement test is a widespread use of tests within institutions and its scope of use varies in situations (Brown, 1989; Douglas, 2003; Fulcher, 1997; Schmitz & C. Delmas, 1991; Wall, Clapham & Alderson, 1994; Wesche et al., 1993). Regarding its purpose, Fulcher (1997) generalized that "the goal of placement testing is to reduce to an absolute minimum the number of students who may face problems or even fail their academic degrees because of poor language ability or study skills" (p. 1).

2.1.5 Validation of an EPT

2.1.6 Testing and assessment of writing in a second language

Writing in a second language

Raimes (1994) indicates it as "a difficult, anxiety-filled activity" (p. 164). Lines (2014) took it into details: for any writing task, students need to not only draw on their knowledge of the topic, its purpose and audience but also make appropriate structural, presentational and linguistic choices that shape meaning across the whole text.

Testing and assessment of writing in second language

Table 2.2 A framework of sub-skills in academic writing (McNamara, 1991)

Criterion (sub-skill)	Description and elements
Arrangement of Ideas and Examples (AIE)	1. presentation of ideas, opinions, and information 2. aspects of accurate and effective paragraphing 3. elaborateness of details 4. use of different and complex ideas and efficient arrangement 5. keeping the focus on the main theme of the prompt 6. understanding the tone and genre of the prompt 7. demonstration of cultural competence
Communicative Quality (CQ) of Coherence and Cohesion (CC)	1. range, accuracy, and appropriacy of coherence-makers (transitional words and/ or phrases) 2. using logical pronouns and conjunctions to connect ideas and/ or sentences 3. logical sequencing of ideas by use of transitional words 4. the strength of conceptual and referential linkage of sentences/ ideas
Sentence Structure Vocabulary (SSV)	1. using appropriate, topic-related and correct vocabulary (adjectives, nouns, verbs, prepositions, articles, etc.), idioms,

	expressions, and collocations. 2. correct spelling, punctuation, and capitalization (the density and communicative effect of errors in spelling and communicative effect of errors in word formation (Shaw & Taylor, 2008, p.44)) 3. appropriate and correct syntax (accurate use of verb tenses and independent and subordinate clauses) 4. avoiding use of sentence fragments and fused sentences 5. appropriate and accurate use of synonyms and antonyms

2.2 Generalizability theory (G-theory)

What is Generalization theory (G-theory)?

Generalizability (G) theory is a statistical theory about the dependability of behavioral measurements.

2.2.1 Generalizability and Multifaceted Measurement Error

2.2.2 Sources of variability in a one-facet design

2.3 Summary

Based on the above review of current validation studies in language testing and assessment, especially EPT in colleges and universities, I would like to investigate the validity of the English placement writing test (EPT W) used at BTEC International College – Da Nang Campus, which is administered to new comers whose first language is not English. Using the framework of interpretative argument for

the TOEFL iBT test developed by Chapelle et al. (2008), I propose the interpretative argument for the Writing EPT W test by focusing on the following inferences: generalization, and explanation. To achieve those aims, this study sought to answer these three research questions. The first two questions aimed to provide evidence underlying the inferences of evaluation and generalization. The third question, which involved in an analysis of linguistic features from the hand-typed writing record of 21 passed tests, backed up the evidence for the explanation inference.

CHAPTER 3
METHODOLOGY

This chapter first provides the information about the research design of the study. The chapter presents knowledge about the participants including test takers, raters, materials, data collection procedures, and data analyses to answer each of research questions.

3.1. Research design

This study employed a descriptive design that involved collecting a set of data and using it in a parallel manner to provide a more through approach to answering the research questions. The qualitative data were the 21 typescripts of written exams by students who passed the entrance placement tests (79 out of 100 test takers did not pass that were placed into English class Level 0). The quantitative data included: 400 writing scores for two writing tasks from a total of 100 test takers (each task was scored by two raters).

3.2. Participants

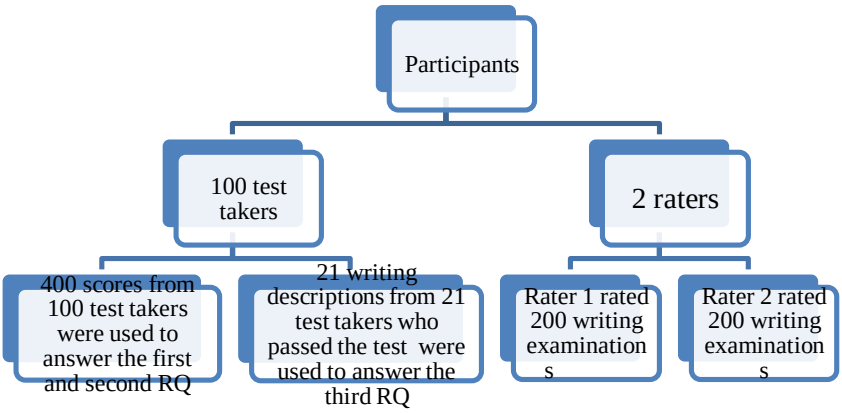

Figure 3.1 Participants

3.2.1 Test takers

3.3 Materials

The material used in this study included the English Placement Test, the writing task types, and the rating scale (rating rubric).

3.3.1 The English Placement Writing Test (EPT W) and the Task Types

There are two tasks in the EPT W. An example of each task is presented below:

Task 1: Write a paragraph (100-120 words) to describe an event you attended recently.

Task 2: Write a paragraph (100-120 words) to describe a vacation trip from your childhood. Using these clues:

Where did you go? When did you go? Who did you go with? What did you do? What is the most memorable thing? Etc.

For two tasks, 30 minutes was given and taking notes was allowed.

3.3.2 Rating scales

The test takers' written examinations were judged based on the rating from 0 to >7. More information about criteria and descriptors of each band can be found in Appendix A. Each of score level is equivalent to appropriate class that provides learners with learning materials, lectures and progress tests. This band is presented below:

- 0: not pass
- 0 – 4.5: level 1 class, material: Top Notch Fundamental
- 4.5 – 5.0: level 2 class, material: Top Notch 1
- 5.0 – 5.5: level 3 class, material: Top Notch 2
- 5.5 – 6.5: level 4 class, material: Top Notch 3
- 6.5 – 7.0: level 5 class, material: Summit 1
- >7: pass

3.4 Procedures

3.4.1 Rater training

3.4.2 Rating

3.4.3 Data analysis

- SPSS version 22 was used for data analysis for answering the first and the second research question. The third question relates to linguistic feature research was addressed by Vocabulary Profiler within Compleat Lexical Tutor software (www.lextutor.ca/vp/eng/)

3.5 Data analysis

3.5.1 To what extent is test score variance attributed to variability in the following: a. task?; b. rater?

3.5.2 How many raters and tasks are needed to obtain the test score dependability of at least .85?

3.5.3 What are vocabulary distributions across proficiency levels of academic writing?

CHAPTER 4
RESULTS

4.1 Results for Research Question 1

- Table 4.1 Variance components attributed to test scores

Source of Variance	Estimate	Percentage
Person	1.063	29%
Task	0.299	11%
Rater	0.009	0.2%
Person*Task	1.264	35%
Person*Rater	0.71	20%
Rater*Task	0.004	0.11%
Person*Rater*Task, error	0.277	8%

4.2 Results for Research Question 2

Table 4.2. Dependability Estimate

Source of variation	σ^2			G study				
				Alternative D studies				
		Nr =	1	2	3	3	10	14	10
		Nt =	1	2	5	6	5	10	12
Person (p)	σ^2_p		1.063	1.0633	1.063	1.063	1.0633	1.063	1.063
Raters (r)	σ^2_r		0.009	0.0045	0.003	0.003	0.00009	0.0299	0.0009

Tasks (t)	σ^2_t		0.299	0.1495	0.0598	0.0498	0.0598	0.0299	0.0249167
Pr	σ^2_{pr}		0.71	0.355	0.236667	0.23667	0.071	0.05071429	0.071
Pt	σ^2_{pt}		1.264	0.632	0.2528	0.210667	0.2528	0.1264	0.105333
Rt	σ^2_{rt}		0.004	0.001	0.0002667	0.0002222	0.00008	0.00002857	0.0000333
Prt, e	$\sigma^2_{prt,e}$		0.277	0.06925	0.01847	0.015389	0.00554	0.00197857	0.0023083
σ^2 Rel			2.251	1.05625	0.507937	0.462726	0.32934	0.17909286	0.1786413
Ep^2 Rel			0.32	0.5	0.68	0.7	0.76	0.8558	0.8561

4.3 Results for Research Question 3

Table 4.3 Distribution of vocabulary across proficiency levels

Vocabulary distributions	EPT L1	EPT L2
(1) Total number of types	450	524
(2) Total number of tokens	1260	1686
(3) Lexical diversity (type-token ratio)	0.36	0.31
	0.56 (701/1260)	0.52(859/1670)
(4) Lexical density		
(5) Lexical sophistication	0.887	0.896
(5a) K1 tokens	0.52	0.53
(5b) K2 tokens	0.05	0.15
(5c) AWL tokens	372	379

(6) Total number of word families		

CHAPTER 5
DISCUSSION AND CONCLUSIONS

The purpose of this study was to build a validity argument for the EPT W test. The study focused on two inferences of generalization and explanation (Chapelle et al., 2008). For the generalization inference, this study investigated the extent to which tasks and raters attributed to score variability and how many tasks and raters are needed to get involved in assessment to obtain the test score dependability of at least .85. For the explanation inference, this study analyzed discourse of written responses of two groups of students across different proficiency levels in terms of the extent to which vocabulary distributions are different across proficiency levels of academic writing. This chapter presents a summary and discussion of the findings of each question.

5.1 Generalization inference

5.2 Limitations of the study and suggestion for future research

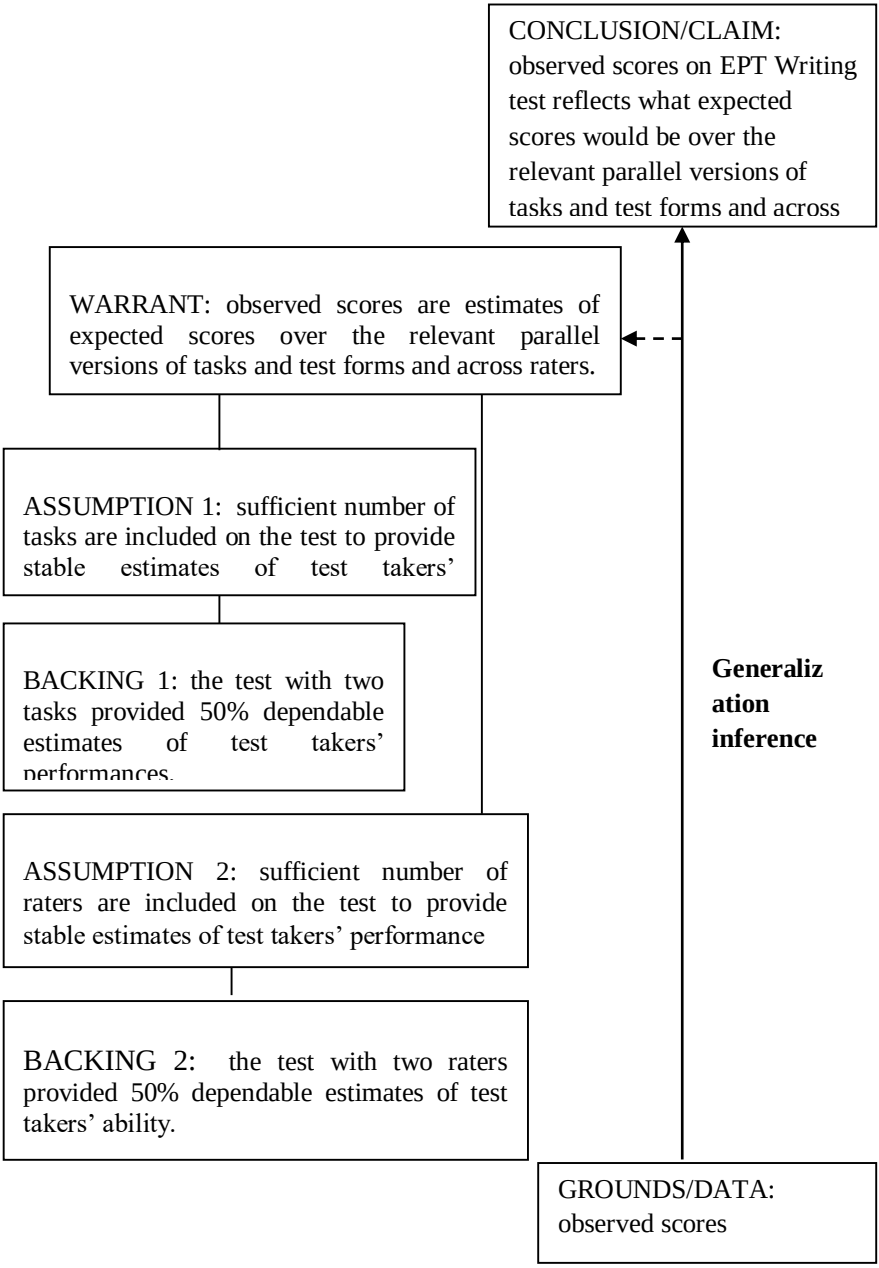
CONCLUSION/CLAIM: observed scores on EPT Writing test reflects what expected scores would be over the relevant parallel versions of tasks and test forms and across
WARRANT: observed scores are estimates of expected scores over the relevant parallel versions of tasks and test forms and across raters.
ASSUMPTION 1: sufficient number of tasks are included on the test to provide stable estimates of test takers'
BACKING 1: the test with two tasks provided 50% dependable estimates of test takers' performances.
ASSUMPTION 2: sufficient number of raters are included on the test to provide stable estimates of test takers' performance
BACKING 2: the test with two raters provided 50% dependable estimates of test takers' ability.
Generalization inference
GROUNDS/DATA: observed scores

Figure 5.1 Generalization inference in the validity argument for the PT W with 2 assumptions and backing

5.2 Explanation inference

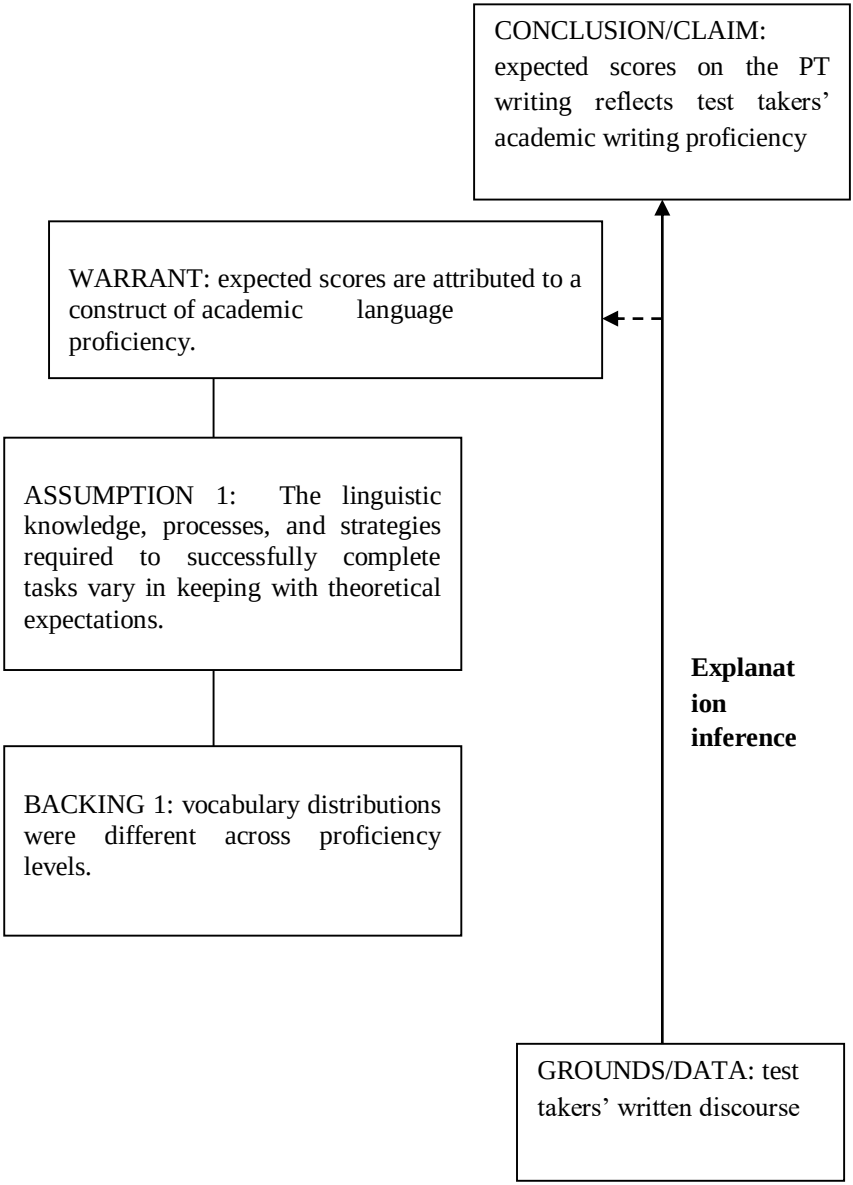

Figure 5.2 The explanation inference in the validity argument for the PT test with 1 assumption and backing

5.3 Summary and implications of the study

5.4 Limitations of the study and suggestion for future research