DEVELOPMENT AND VALIDATION OF THE ACADEMIC COLLOCATIONAL COMPETENCE TEST FOR EFL UNIVERSITY STUDENTS: AN APPLICATION OF THE ARGUMENT-BASED APPROACH

Mr. Apichat Khamboonruang

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Arts Program in English as an International Language
(Interdisciplinary Program)
Graduate School
Chulalongkorn University
Academic Year 2013

การพัฒนาและตรวจสอบคุณภาพแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการสำหรับนิสิตที่เรียนภาษาอังกฤษเป็นภาษาต่างประเทศ: การประยุกต์ใช้วิธีการอ้างเหตุผลโต้แย้ง

นายอภิชาติ คำบุญเรือง

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาศิลปศาสตรมหาบัณฑิต
สาขาวิชาภาษาอังกฤษเป็นภาษานานาชาติ (สหสาขาวิชา)
บัณฑิตวิทยาลัย จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2556

Thesis Title	DEVELOPMENT AND VALIDATION OF THE ACADEMIC COLLOCATIONAL COMPETENCE TEST FOR EFL UNIVERSITY STUDENTS: AN APPLICATION OF THE ARGUMENT-BASED APPROACH
By	Mr. Apichat Khamboonruang
Field of Study	English as an International Language
Thesis Advisor	Assistant Professor Jirada Wudthayagorn, Ph.D.

Accepted by the Graduate School, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

.....Dean of the Graduate School
(Associate Professor Amorn Petsom, Ph.D.)

THESIS COMMITTEE

.....Chairman
(Associate Professor Prakaikaew Opanonamata)

.....Thesis Advisor
(Assistant Professor Jirada Wudthayagorn, Ph.D.)

.....External Examiner
(Assistant Professor Sungworn Ngudgratoke, Ph.D.)

อภิชาติ คำบุญเรือง : การพัฒนาและตรวจสอบคุณภาพแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการสำหรับนิสิตที่เรียนภาษาอังกฤษเป็นภาษาต่างประเทศ: การประยุกต์ใช้วิธีการอ้างเหตุผลโต้แย้ง. (DEVELOPMENT AND VALIDATION OF THE ACADEMIC COLLOCATIONAL COMPETENCE TEST FOR EFL UNIVERSITY STUDENTS: AN APPLICATION OF THE ARGUMENT-BASED APPROACH) อ.ที่ปรึกษาวิทยานิพนธ์หลัก: ผศ. ดร. จิรดา วุฑฒยากร, 195 หน้า.

การใช้แบบสอบสามัตถิยะคำปรากฏร่วมที่ผ่านการพัฒนาและตรวจสอบคุณภาพที่ดีอาจให้คะแนนที่เป็นสารสนเทศให้ผู้ใช้แบบสอบได้รู้ถึงความสามารถด้านภาษาอังกฤษเพื่อวัตถุประสงค์ในการใช้จัดวางตำแหน่งหรือคัดเลือก การตัดสินใจที่เหมาะสมนั้น ผู้ใช้แบบสอบจำเป็นต้องอาศัยข้อมูลที่น่าเชื่อถือจากแบบสอบคำปรากฏร่วมที่ผ่านการพัฒนาและตรวจสอบคุณภาพอย่างดี ดังนั้นวัตถุประสงค์หลักของการวิจัยนี้ คือ การประยุกต์ใช้วิธีการอ้างเหตุผลโต้แย้ง (Kane, 1992, 2006, 2011, 2013) เพื่อพัฒนาและตรวจสอบคุณภาพแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการสำหรับนิสิตที่เรียนภาษาอังกฤษเป็นภาษาต่างประเทศ วิธีการอ้างเหตุผลโต้แย้งประกอบด้วยการสร้างเหตุผลโต้แย้ง 2 ขั้นตอน ขั้นตอนแรก คือ การสร้างเหตุผลโต้แย้งเชิงแปลความ (interpretive argument) โดยการระบุการแปลผลและใช้คะแนนสอบที่ต้องการ และขั้นที่สอง คือ การสร้างเหตุผลโต้แย้งเชิงความตรง (validity argument) โดยการประเมินหลักฐานเชิงทฤษฎีและเชิงประจักษ์ที่รวบรวมเพื่อสนับสนุนการแปลผลและการใช้คะแนนสอบที่ระบุไว้ในเหตุผลโต้แย้งเชิงแปลความ การวิจัยนี้ยังประยุกต์ใช้วิธีการวัดโมเดลราส์ชเพื่อให้หลักฐานเชิงประจักษ์สนับสนุนเหตุผลโต้แย้งเชิงความตรงของแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ

กลุ่มตัวอย่าง คือ นิสิตระดับบัณฑิตศึกษาที่เรียนภาษาอังกฤษในฐานะภาษาต่างประเทศ จำนวน 193 คน จากหลากหลายสาขาวิชาในจุฬาลงกรณ์มหาวิทยาลัย การรวบรวมหลักฐานเชิงทฤษฎีทำในช่วงการพัฒนาแบบสอบและการสร้างเหตุผลโต้แย้งเชิงแปลความของแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ การรวบรวมหลักฐานเชิงประจักษ์ใช้ แบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ แบบสอบระดับคำศัพท์เชิงวิชาการพัฒนาโดย Schmitt, Schmitt, และ Clapham (2001) และแบบสอบถามความคิดเห็นต่อแบบสอบพัฒนาโดย Voss (2012) แบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการสร้างขึ้นโดยใช้คำปรากฏร่วมกริยาและนามที่มีความถี่สูงจากภาษาเขียนเชิงวิชาการในหลายสาขาวิชาที่อยู่ในคลังข้อความ British National Corpus (BNC) แบบสอบนี้สร้างขึ้นเพื่อเป็นแบบสอบจัดระดับแบบอิงกลุ่มเพื่อวัดสามัตถิยะคำปรากฏร่วมเชิงรับ (receptive collocational competence) ของนิสิตระดับบัณฑิตศึกษาที่เรียนภาษาอังกฤษในฐานะภาษาต่างประเทศ การวิเคราะห์ข้อมูลเชิงประจักษ์ใช้ สถิติพรรณนา การวิเคราะห์โมเดลราส์ช การวิเคราะห์สหสัมพันธ์ การวิเคราะห์ความแปรปรวน การวิเคราะห์ไคสแควร์ การวิเคราะห์เนื้อหา การวิเคราะห์คะแนนจุดตัด และการวิเคราะห์ความคลาดเคลื่อนของการจัดกลุ่ม

ผลการวิจัยพบว่า วิธีการอ้างเหตุผลโต้แย้งช่วยในการพัฒนาและตรวจสอบคุณภาพแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ เหตุผลโต้แย้งเชิงแปลความใช้เป็นแนวทางในการออกแบบและพัฒนาแบบสอบและรวบรวมหลักฐานซึ่งผ่านการประเมินเพื่อสร้างเหตุผลโต้แย้งเชิงความตรงของแบบสอบ กระบวนการพัฒนาแบบสอบและเหตุผลโต้แย้งเชิงแปลความเป็นกระบวนการที่สัมพันธ์กัน และผ่านการแก้ไขปรับปรุงให้สอดคล้องกับการแปลผลและการใช้คะแนนสอบที่ได้ระบุไว้และเหมาะสมกับบริบทของการวิจัยนี้ เหตุผลโต้แย้งเชิงความตรงเป็นตัวบ่งชี้ถึงระดับความตรงหรือความเหมาะสมของการแปลผลและการใช้คะแนนสอบ ระดับของความตรงหรือความเหมาะสมขึ้นอยู่กับหลักฐานที่รวบรวมเพื่อสนับสนุนการแปลผลและการใช้คะแนนสอบที่ระบุไว้ในเหตุผลโต้แย้งเชิงแปลความ

เหตุผลโต้แย้งเชิงความตรงของแบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการบ่งชี้ถึงระดับความตรงที่เหมาะสมของการแปลผลและการใช้คะแนนสอบ กล่าวคือ การแปลผลและการใช้คะแนนสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการมีความเหมาะสม เหตุผลโต้แย้งเชิงความตรงของแบบสอบอิงหลักฐานเชิงทฤษฎีและเชิงประจักษ์ที่น่าเชื่อถือและเพียงพอในการสนับสนุนข้อสมมุติฐานในการอนุมานด้านการบรรยายเนื้อหา การประเมิน การสรุปอ้างอิง การอธิบาย การประมาณค่า และ การใช้ หลักฐานสนับสนุนการอนุมานด้านผลที่ตามมาไม่ได้ศึกษาในการวิจัยครั้งนี้ นอกจากนี้ยังพบว่า วิธีการวัดโมเดลราส์ชให้หลักฐานเชิงประจักษ์ที่น่าเชื่อถือในการสนับสนุนเหตุผลโต้แย้งเชิงความตรงของแบบสอบ หลักฐานจากโมเดลราส์ชได้แก่ ด้านความเป็นเอกมิติ ความสอดคล้องภายใน การกระจายและลำดับชั้นความสามารถของผู้สอบ การกระจายและลำดับชั้นความยากของข้อสอบ การทำหน้าที่ของตัวลวง การทำหน้าที่ต่างกันของแบบสอบ และการทำหน้าที่ต่างกันของข้อสอบแบบเอกรูป

สาขาวิชา ภาษาอังกฤษเป็นภาษานานาชาติ ลายมือชื่อนิสิต .....

ปีการศึกษา 2556 ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก .....

5587642120 : MAJOR ENGLISH AS AN INTERNATIONAL LANGUAGE
KEYWORDS: ACADEMIC COLLOCATIONAL COMPETENCE TEST / EFL UNIVERSITY STUDENTS / ARGUMENT-BASED APPROACH / RASCH MEASUREMENT APPROACH

APICHAT KHAMBOONRUANG: DEVELOPMENT AND VALIDATION OF THE ACADEMIC COLLOCATIONAL COMPETENCE TEST FOR EFL UNIVERSITY STUDENTS: AN APPLICATION OF THE ARGUMENT-BASED APPROACH. ADVISOR: ASST. PROF. JIRADA WUDTHAYAGORN, Ph.D., 195 pp.

Using a well-developed and validated test of specific English collocational competence may provide meaningful scores that partly inform test users of to what extent test-takers are proficient in English for the purposes of placement or screening uses. To make proper decision, test users need to depend remarkably on trustworthy information provided by a well-developed and validated collocation test. The primary purpose of the present study was, therefore, to apply the argument-based approach (Kane, 1992, 2006, 2011, 2013) to develop and validate the Academic Collocational Competence Test (ACCT) for EFL graduate students. The argument-based approach involves two argument development stages. The first stage is to develop the interpretive argument by specifying the intended interpretation and use of test scores and the second stage is to build the validity argument by evaluating theoretical and empirical evidence collected to support such intended score interpretation and use specified in the interpretive argument. This study also aimed to apply the Rasch measurement approach to provide empirical evidence in support of the ACCT validity argument.

A total of 193 EFL graduate students from various academic disciplines at Chulalongkorn University participated in this study. Theoretical evidence was collected during the development of the ACCT and the ACCT interpretive argument. Empirical evidence was gathered using the ACCT, the Academic Vocabulary Level Test (AVLT) developed by Schmitt, Schmitt, and Clapham (2001), and the test reflection questionnaire adopted from Voss (2012). The ACCT was developed using high-frequency verb-noun collocations from varying domains of the academic written discourse in the British National Corpus (BNC) and developed primarily as a norm-referenced placement test of receptive collocational competence of EFL graduate students. Empirical data were analysed using descriptive statistics, Rasch model analysis, correlation analysis, analysis of variance, chi-square analysis, content analysis, cut score analysis, and classification error analysis.

Research results revealed that the argument-based approach helped the development and validation of the ACCT. The interpretive argument served as the guideline for designing and developing the ACCT and also for assembling evidence that was later appraised to construct the validity argument of the ACCT. The development process of the ACCT and the ACCT interpretive argument was an interactive process and was modified until they were consistent with the intended score interpretation and use as well as the context of the current study. The validity argument indicated to what degree the ACCT score interpretation and use were valid or appropriate based on collected evidence collected to support the score interpretation and use specified in the ACCT interpretive argument.

The ACCT validity argument revealed a reasonable degree of validity of the ACCT score interpretation and use. That is, the ACCT scores were appropriately interpreted and used as intended. The ACCT validity argument was based on sound and sufficient theoretical and empirical evidence supporting assumptions in domain description, evaluation, generalization, explanation, extrapolation, and utilisation inferences in the ACCT interpretive argument. Backing for the consequence inference is beyond the scope of this study. The Rasch measurement approach provided sound empirical evidence in support of the ACCT validity argument. Rasch-based evidence included unidimensionality, internal consistency, examinee competency dispersion and hierarchy, item difficulty dispersion and hierarchy, multiple-choice distractor functioning, differential test functioning, and uniform differential item function.

Field of Study: English as an International Language Student's Signature .....

Academic Year: 2013 Advisor's Signature .....

ACKNOWLEDGEMENTS

This master's thesis could not have been completed without the guidance, assistance, and support from a number of people to whom I am very much indebted. First and foremost, I would like to thank my advisor, Asst. Prof. Dr. Jirada Wudthayagorn, for providing me with professional guidance, spending valuable time revising and editing my thesis writing, strengthening my knowledge of language assessment, and taking very good care of my thesis project from beginning to end. I am indeed very grateful to her careful supervision. I would also like to appreciate insightful comments and excellent suggestions from other members of my thesis committee, Assoc. Prof. Prakaikaew Opanonamata and Asst. Prof. Dr. Sungworn Ngudgratoke. Their incisive remarks and invaluable recommendations actually shaped my way of doing research and tremendously enhanced the quality of my thesis study. In particular, I would like to thank Dr. Sungworn Ngudgratoke for giving me such well-versed advice on psychometric concepts and quantitative data analysis. Without the assistance and contribution of my advisor and thesis committee, I could not have finished writing up this thesis and completed my M.A. degree.

There are several people and professors at Chula and at home to whom I wish to thank and express my sincere gratitude. My heartfelt thanks go to Dr. Sutthirak Sapsirin, Dr. Tanyaporn Arya, and Professor Michale Alroe at Chulalongkorn University Language Institute for their valuable time and willing assistance in evaluating the test and giving several useful suggestions to improve the test. I am deeply grateful to all my professors at the department of educational research and psychology for introducing me into the realm of statistics, psychometrics, and research methodology and inculcating into me a sense of a decent academic and a professional researcher while I was pursuing my M.Ed. degree in educational research. My genuine gratefulness also goes to all my professors at EIL for sharpening my knowledge of English linguistics, English language instruction, and in particular language assessment and evaluation which is the area I am exceptionally keen on and committed to working in for the rest of my life. I would also like to acknowledge any support and encouragement from friends during my M.A. study voyage at EIL and we had a very good time studying at EIL. Last but not least, I would like to give my deep gratitude to my parents for their long-running care and support. They tried as best they could to help me accomplish my goals. Without their financial and spiritual support, I could not have achieved my academic goals.

It was very well worthwhile sharing and gaining intellectual experience from all my professors and fellows on my long academic journey at Chula since 2003. Of course, my thesis study is not by any means impeccable and any reckless faults that remain in this thesis are on no account my own. I will take my thesis study as a steppingstone towards my prospective career as a lecturer and a linguametrician.

CONTENTS

Page

Page

Page

Page

Page

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1
INTRODUCTION

Chapter 1 is intended to provide an introduction to the current research on the development and validation of the Academic Collocational Competence Test (henceforth referred to as ACCT). In this chapter, I begin by describing the background of the current study. Following this, I address research questions, specify research objectives, determine the scope of the study, and present the definitions of key terms. After that, the significance of the present study is discussed. This chapter ends with a brief summary of this chapter

1.1 Background of the study

The pivotal role of phraseological units, otherwise called formulaic sequences, prefabricated language and so forth, has long been acknowledged in second language development (e.g., Conklin & Schmitt, 2008; Firth, 1957; Hoey, 2005; Lewis & Conzett, 2000; Nation, 2001; Nattinger & DeCarrico, 1992; Schmitt, 2004a, 2004b; Sinclair, 1991; Wray, 2005, 2008). Collocation, one of the phraseological units, is widely recognised by several scholars as a necessary part of second language learning and teaching and by far one of the most extensively-studied features (e.g., Bahns, 1993; Bahns & Eldaw, 1993; Benson, Benson, & Ilson, 2010; Howarth, 1998; Laufer, 2011; Laufer & Waldman, 2011; Nattinger & DeCarrico; Nesselhauf, 2003, 2005).

Up to the present day, collocation has been extensively researched in different trajectories. In the realm of language instruction, a number of studies, for instance, aimed primarily at investigating the effects of teaching collocation on several dimensions of second language development (e.g., Boers, Demecheleer, Coxhead, & Webb, 2013; Hsu, 2007, 2010; Hsu & Chiu, 2008; Rahimi & Momeni, 2012; Webb & Kagimoto, 2011; Webb, Newton, & Chang, 2013) or examining the effects of interventions on collocational knowledge enhancement (e.g., Chan & Liou, 2005; Daskalovska, 2013; Goudarzi & Momi, 2012; Molina-Plaza & de Gregorio-Godeo, 2010). With regard to linguistics, several studies, for example, aimed primarily to examine collocational behaviour (e.g., Walker, 2011a; Walker, 2011b) or investigate the use of collocations by L2 learners based on corpora of different genres (e.g., Bazzaz & Samad, 2011; Durrant & Schmitt, 2009; Gao & Zhang, 2009; Hashemi, Azizinezhad, & Dravishi, 2012a, 2012b; Laufer & Waldman, 2011) or analyse collocations in different

English language teaching materials (e.g., Durrant, 2009; Menon & Mukundan, 2012). In the field of language assessment, a body of research was set out to explore collocation use in L2 learners corpora or assess L2 collocational knowledge through developing and validating collocational measures based on different testing purposes, perspectives, and psychometric methods (e.g., Alsakran, 2011; Gitsaki, 1999; Jaén, 2007; Keshavarz & Salimi, 2007; Kim, 2008; Sadeghi, 2009; Voss, 2012; Webb & Kagimoto, 2011; Wolter & Gyllstad, 2011).

The present study attempts to contribute to the later line of research by drawing upon one of the recent comprehensive approaches "the argument-based approach" and one of the advanced psychometric methods "the Rasch psychometric model" to the development and validation of the ACCT which would provide scores that could be accurately interpreted as reflecting collocational competence and appropriately used as a norm-referenced test for placement or screening decision in English language courses in universities or other academic institutions of higher education. The motivation for developing the ACCT is resulted from the fact that English is now widely recognised as the lingua franca in the academic world (Jenkins, 2007; McKay & McKay, 2002; Sowden, 2012). That is to say, English is mostly and globally used by non-native speakers for academic purposes. In Thailand where people use English as a Foreign Language (EFL), a large number of students enter universities each year to pursue their advanced studies and they are required to pass one of standardised English tests such as TOEFL, IELTS, or Chulalongkorn University Test of English Proficiency (CU-TEP) in order to be accepted. Although these proficiency tests are assumed to assess students' English proficiency to survive in an academic context, it may nevertheless be needed to use supplementary testing to screen who should or should not take more English courses by finding each student's appropriate level of English proficiency and place them accordingly into proper class or group levels in English courses with particular emphasis on academic language and skills needed for academic success in university.

As such, if teachers of English know to what extent learners possess academic collocational ability, this may help them determine how proficient learner are in English and who should or should not take more English courses in order to survive their advanced studies in university or other higher-education settings where English is a tool for learning. To make proper decision as such, teachers need to rely hugely on sound and sufficient information provided by a well-developed and validated collocation test. In this regard, it is essentially of great use that an additional English placement test be developed and validated carefully to provide meaningful scores

that can be interpreted and used to inform a decision-making process regarding placement or screening. Since there is no single test that can perfectly measure psychological traits, using multiple tests may help ensure that intended decision is made as appropriately as possible. Accurate score interpretation and use can indeed be highly beneficial for both test users and test-takers, while misinterpretation or misuse of test scores might go the other way round.

Successful assessment is, of course, resulted from successful score interpretation and use. Over the past decade or so, the concept of validity has been extended to encompass empirical evidence and relevant theory which can be used to argue in favour of the proposed interpretation and utilisation of test scores. This validity concept is regarded as the contemporary perspective on validity which is in line with, for example, Kane (1992, 2006, 2011, 2013), Messick (1994), and American Educational Research, Association, American Psychological Association, and National Council on Measurement in Education (1999). Based on this contemporary perspective, validity is conceptualised as the degree to which test scores can be validly interpreted as reflecting a construct and validly used as intended purposes. This means that there must be relevant theory and empirical evidence supporting the proposed interpretation and use of test score. In essence, contemporary validity is the degree to which the interpretation and use of test scores are valid based on theory and evidence. One effective approach to accomplishing this contemporary validity is the argument-based approach, proposed by Kane (1992, 2006, 2011, 2013), where what is validated is the interpretation and utilisation of test scores, not the test proper. The argument-based validation approach has recently come into sharp focus in modern validity theory and has been acknowledged by several scholars (e.g., Brennan, 2013; Carol A Chapelle, 2012; Carol A. Chapelle, Enright, & Jamieson, 2010; Carol A Chapelle, Enright, & Jamieson, 2008; LeBaron Wallace, 2011; Oller, 2012; Stephen G. Sireci, 2007; Stephen G Sireci, 2013).

Kane's argument-based approach to validation is crucially composed of two interconnected argument development procedures. The first procedure is the development of the interpretive argument by specifying the proposed claims concerning the intended interpretations and uses of test scores. The second procedure is the development of the validity argument which is a comprehensive appraisal of the evidence collected to evaluate the interpretive argument. In essence, the argument-based approach provides the framework for evaluating the proposed interpretations and utilisation of test scores based on theory and evidence. It is not surprising then that an increasing number of studies have recently adopted

the argument-based approach to validating language assessment tools (e.g., Le, 2011; Pardo-Ballester, 2010; Voss, 2012). Precisely for this reason, the present study aimed to apply Kane's argument-based approach to the development and validation of the current collocation test and based the interpretive argument on the framework developed by Carol A Chapelle et al. (2008), and Voss (2012).

In assessing vocabulary knowledge, scholars typically divide vocabulary knowledge into breadth and depth aspects (Daller, Milton, & Treffers-Daller, 2007; Haastrup & Henriksen, 2000; Milton, 2009; Read, 2000, 2007; Read & Chapelle, 2001), each of which can be measured either receptively or productively. Vocabulary breadth refers to vocabulary size or how many words learners know, while vocabulary depth refers to how well the words are known in terms of the different meanings of a single word or knowledge of other words that frequently co-occur when produced (Daller et al., 2007; Milton, 2009; Read, 2000, 2007; Read & Chapelle, 2001). The literature illuminates that breadth and depth aspects of vocabulary knowledge are closely related (Akbarian, 2010; Qian, 1999, 2002; Read, 2000, 2007). Studies aiming to assess a receptive knowledge typically used selected-tem formats such as a multiple-choice test to elicit such knowledge (Gyllstad, 2005, 2007; Jaén, 2007; Keshavarz & Salimi, 2007; Webb & Kagimoto, 2011; Webb et al., 2013). As such, the present study uses a multiple-choice test with five options to elicit examinees' receptive collocational competence.

The ability to combine words into larger phrasal units properly is also called the lexical, phraseological or collocational competence. Learners need to know a large number of lexical items and know a great deal how words combine or collocate with each other if they wish to express themselves accurately, fluently, and naturally in their language performance (Benson et al., 2010; Lewis & Conzett, 2000; Nesselhauf, 2003, 2005; O'Dell & McCarthy, 2009; Read, 2000; Schmitt, 2004a, 2010; Sinclair, 1991). According to Benson et al. (2010), word combinations can be divided into two categories: grammatical collocations and lexical collocations. Grammatical collocations consist mainly of a dominant word (noun, adjective, and verb) and a preposition or grammatical structure such as an infinitive or clause. Lexical collocations, by contrast, typically do not contain prepositions infinitives, or clauses. Lexical collocations consist of nouns adjectives, and verbs. These two categories exemplify the kind of collocational knowledge native speakers of English have in common.

The lexical verb-noun collocation is chosen in particular as a construct to be measured since a body of research has established that second language learners

have difficulty producing verb-noun collocations which are commonly found in the academic written discourse (e.g., Ganji, 2012; Laufer & Waldman, 2011; Marco & José, 2011; Nesselhauf, 2003, 2005). A verb-noun lexical collocation is thereby the focal interest of a measure in the present study. The definition of collocations in this study is based primarily on a phraseological approach (Carter, 1998; Cowie, 1998; Howarth, 1998). Based on a phraseological perspective, a verb-noun collocation is defined as habitually occurring lexical combinations that are characterised by restricted co-occurrence of elements and relative transparency of meaning. in the current study, the phraseologist-based collocation definition was included as part of the overall interactionist-based collocational construct definition, proposed by Carol A Chapelle (1998).

With the availability of large corpora of various genres, this study also takes advantage of a corpus-based approach to systematically sample high-frequency collocations from the British National Corpus (BNC), which contains a large collection of academic English texts. Corpus-based collocation sampling is of great benefit not only to enhance the authenticity of the task representing the target language use (TLU) in the academic setting, but also to connect language knowledge and content knowledge (Carr, 2011; Douglas, 2000), A measure of collocational competence based on restricted collocations sampled from a TLU corpus may to a larger extent provide helpful information that reflects language ability which is inferred from language performance in universities or other higher education institutions.

Also of focal interest in this study is apply the Rasch measurement approach, which was initiated by Rasch (1960) and acknowledged as superior in several ways to true-score theory or classical test theory (CTT), to investigate the psychometric quality of the ACCT. The Rasch model applies mathematical logistic models to put item and person estimates on the same latent metric and by so doing the probability of getting an item correct depends significantly on the person ability and the item difficulty (Bond & Fox, 2007; Boone, Staver, & Yale, 2014; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Iramaneerat, Smith, & Smith, 2008; Linacre, 2012; Rasch, 1960, 1980; Schumacker, 2004; Wilson, 2005). In the Rasch probabilistic model paradigm, the Rasch approach considers a measurement model as a tool for making sense of a particular theoretical framework. Therefore, the model is not chosen to fit the data but rather the data are required to fit the Rasch model (Bond & Fox, 2007; Iramaneerat et al., 2008). If the data fit the Rasch model, it can then be confident that estimates of persons and items provide meaningful measurement properties, contributing to sound empirical evidence.

For example, person ability and item difficulty measures are put on the common logit scale which has equal measurement units. Thus, person ability and item difficulty can be compared. Raw ordinal scores are converted into interval logits or mesures. Rasch-based person ability measures are free from any sets of Rasch-based validated items and Rasch-based items difficulty measures are free from any groups of persons. Moreover, individual person ability or individual item difficulty measure has a unique standard error associated with its estimate. What is more, anomaly responses can be detected using person ability and item difficulty measures (Bond & Fox, 2007; Boone et al., 2014; Iramaneerat et al., 2008; Linacre, 2012; Schumacker & Smith, 2007). While most of the studies on collocational assessment used CTT to investigate the psychometric quality of collocational tests, only very recently have there been a few studies applying the Rasch IRT approach (Voss, 2012). In the light of this, this study hence intends to apply the Rasch model to evaluate psychometric properties of the ACCT, thereby maximising the overall validity argument of the ACCT.

What I have rationalised previously essentially underpins the objectives of my research study. The primary purpose of this study is to develop and validate the ACCT that can provides scores which is meaningfully interpreted as an indicator of English collocational competence and used primarily for placement decision in courses related to academic English language or skills in universities or other institutions of higher education. The restricted verb-noun collocation is chosen as a construct to be measured and the argument-based approach was adopted as the framework for developing the ACCT and validating the claims about the proposed interpretation and use of the ACCT scores. What is also of focal interest is the use of a corpus-based approach to systematically sample collocations from BNC which is claimed to represent the academic written discourse of interest. Also of particular interest is applying the Rasch model to investigate and improve the psychometric quality of the ACCT.

It is my fervent hope that this master thesis would significantly shed more light on the applications of both the argument-based approach and the Rasch model to the development and validation of collocational tests, and make a valuable contribution to the theoretical and empirical validation of language assessment in general and collocational assessment in particular. The hybrid of two scientific models was of greater help to validate the score interpretation and use of the ACCT, which was developed using a five-option multiple-choice format, based on a corpus-driven method, and designed primarily as a norm-referenced placement test of EFL

graduate students' receptive collocational competence. All this could be of significant contribution to language teachers and those who are particularly interested in conducting test-developing research.

1.2 Research questions

The research questions of the present study are addressed as follows.

1) To what degree are scores on the ACCT interpreted as an indicator of collocational competence of EFL university students and used for placement decision in English language courses in universities or other academic institutions at tertiary level?

2) How does the argument-based approach to validation help develop the ACCT and validate the proposed interpretation and use of scores on the ACCT?

3) How does the Rasch psychometric model help validate psychometric properties of the ACCT?

To elaborate, the research questions addressed above necessitate clear coherent and complete developments of both the interpretive argument and the validity argument based on theoretically and empirically well-established evidence. Therefore, both relevant theoretical and empirical evidence need to be appropriately and adequately assembled in this research in order to accomplish such goal. Relevant theory is documented in the literature review, whereas empirical evidence is gathered since the ACCT was developed up until empirical data were collected and statistically analysed. Every detail of the entire process of the ACCT development and validation all provides the information in response to the current research enquiries. Responses to research question 1 was derived primarily from the construction of the validity argument for the ACCT in chapter 6 and responses to research question 2 was obtained from chapter 2 to chapter 6. Responses to research question 3 stemmed from empirical results of the Rasch analysis in chapter 5. Guiding responses to research questions were also presented in more detail in chapter 6.

1.3 Research objectives

The primary objectives of the present study are to:

1) Develop the ACCT for EFL university students that can provide meaningful scores which are interpreted as an indicator of collocational competence and used for placement decision in English language courses in universities or other academic institutions at tertiary level.

2) Apply the argument-based approach to develop the ACCT for EFL university students and validate the proposed interpretation and use of scores on the ACCT.

3) Apply the Rasch psychometric model to validate psychometric properties of the ACCT for EFL university students under the framework of the argument-based approach to validation.

1.4 Scope of the study

The present study was set out with the primary aim of developing and validating the ACCT for EFL university students by applying the argument-based approach and the Rasch psychometric model. The generalisation of findings from this study is based on characteristics, approaches, and frameworks defined and used in the current study. The design, development and validation of the ACCT were based predominantly on the argument-based approach to validation (Kane, 1992, 2006, 2011, 2013), which views validity as the meaningful interpretation and use of test score and relies heavily on two types of arguments: the interpretive argument and the validity argument. This study built upon the TOEFL interpretive argument framework (Carol A Chapelle, 2008, 2012; Carol A. Chapelle et al., 2010; Carol A Chapelle et al., 2008) and the interpretive argument framework developed by Voss (2012). However, investigation of evidence for the consequence inference is beyond the scope of the current study.

The design of the ACCT was based partly on a corpus-based approach to sampling collocations from the TLU domain of academic written English. Criteria for sampling collocation were based on frequency, statistics, and judgement. The validation of the psychometric quality of the ACCT was based primarily on the Rasch Imeasurement model (Rasch, 1960). Several applications of the Rasch model were mapped onto Kane's argument-based validity framework. In term of collocational construct definition under measure, although the argument-based approach does not necessarily call for a theory-based construct definition, the collocational construct

under study was defined based on an interactionist approach (Carol A Chapelle, 1998), which defines the construct as including assessment context, linguistics competence, and cognitive strategies. Collocation was also linguistically defined according to the phraseologist perspective (e.g., Carter, 1998; Cowie, 1998; Howarth, 1998) and the focus of the collocation type was on a lexical verb-noun combination as classified by Benson et al. (2010).

The assessment of collocational competence was focused on a receptive aspect of vocabulary depth or collocation knowledge. Collocational competence was operationalised or measured using a five multiple-choice item test. The standard setting methods for cut-score establishment and classification error estimation were based primarily on a contrasting-group approach (Livingston & Zieky, 1982) and secondarily on a Bayesian approach. Although the target test-taker population of interest is EFL university students, the samples of EFL university test-takers in the study were EFL graduate students with different English proficiency level from different fields of study at Chulalongkorn University and almost all of the students were Thai graduate students. Therefore, findings from this study should be interpreted and generalised on the basis of the scope of the present study.

1.5 Definitions of terms

1.5.1 Academic Collocational Competence Test

The ACCT refers to the multiple-choice test designed to measure EFL learners' collocational competence demonstrated in the academic writing discourse and facilitate norm-referenced placement decision. The ACCT was developed by the author in this study with a view to providing scores that are expected to provide meaningful information facilitating placement decision in academic English courses at university or other institutions of higher education in the EFL context. It is aimed in particular to measure a receptive dimension of collocational competence, which is part of vocabulary depth knowledge.

1.5.2 EFL university students

EFL university students refer to graduate student who study English as a Foreign Language (EFL), where English is learned and used primarily for academic purposes in the classroom or university and is not used outside the classroom for everyday purposes. In this study, the samples of EFL university students were EFL

graduate students with varying English proficiency levels and from different academic disciplines at Chulalongkorn University.

1.5.3 Argument-based approach

The argument-based approach (Kane, 1992, 2006, 2011, 2013) refers to the model or framework for developing the ACCT and validating the proposed claims based on the scores of the ACCT. The argument-based approach views validity as arguments or rationales supporting the claims about proposed score interpretation and use, as the degree of appropriate score interpretation and use in lieu of simply valid or invalid, and as supported by theoretical and empirical evidence. In this regard, argument-based validity requires validation process involving collecting evidence for proposed score interpretation and use. Kane's argument-based validation approach builds essentially on two interrelated arguments or rationales, the interpretation argument and the validity argument. The interpretive argument, sequently renamed by Kane in 2013 as interpretive/use argument, specifies the statements of the intended interpretation and use of tests cores. The validity argument was then developed through evaluating evidence collected to support score interpretation and use as stated in the interpretive argument.

In short, the argument-based approach focuses on validating test score interpretation and use by evaluating the feasibility of the proposed interpretation and use of test scores. Therefore, the proposed interpretation and use of test scores need to be initiated as clearly as possible. Kane’s argument-based approach involves two argument development stages. The first stage is to develop the interpretive argument by specifying the intended interpretation and use of test scores. The second step is to build the validity argument by evaluating a priori and empirical evidence sought to support such intended interpretation and use of test scores outlined in the interpretive argument.

1.5.4 Rasch measurement approach

The Rasch measurement approach (Rasch, 1960) is a family of model-based statistical techniques in measurement used to evaluate the psychometric quality of collocation tests. Based on the Rasch psychometric model, a test taker’s response to a binary/dichotomous item (i.e., agree/disagree, right/wrong, true/false) is determined by the test taker’s competency level and the difficulty of the binary item. The Rasch

model estimates competency levels or the probability of a correct response using a mathematical function of person ability and item difficulty parameters. Application of the Rasch model to a set of data provides a range of diagnostic information as to how well items work in measuring the collocational competence construct under investigation.

The Rasch model analysis enables the test to be modified by revising or removing items so that the test can better assess the competency. The Rasch model can also help establish the internal consistency and the construct validity of a set of items. Estimates of person ability are independent of which items are used for comparisons. Similarly, estimates of item difficulty are independent of which persons are used for comparisons. If the data fit the Rasch mode, estimates of persons and items provide meaningful measurement properties, contributing to sound empirical evidence the legitimate validity argument of the ACCT.

1.6 Significance of the study

It is very much hoped that the findings from this master's thesis study would potentially make several significant contributions to the study of collocation and the validation of language assessment. The significance of the current study is discussed in terms of theoretical and practical significance.

1.6.1 Theoretical significance

In terms of theoretical significance, the present study provides the way of applying the argument-based approach to define construct definition of collocational competence and model the framework for validating the interpretation and use of language test scores. Findings from this study could shed novel light into how to apply the argument-based approach to model a more thorough framework for developing and validating language assessment instruments that provide scores which can be appropriately interpreted with inference to the linguistic competence and used with regard to the placement decision on placing test-takers into appropriate English language courses in universities or other academic institutions at tertiary level. Another theoretical significance is that this study offers the way of measuring a specific collocational competence, which is one linguistic feature that may be embedded as part of a measure of writing ability or other English skills for

placement decision in academic English courses at universities or other academic institutions at tertiary level.

1.6.2 Practical significance

With regard to practical significance, this study exemplifies the way of developing a language test using a corpus-based approach for the specific purpose of eliciting performance of collocational ability as a language feature commonly used in the academic context. The use of a corpus-based method to sample linguistic features under measure from representative corpora significantly helps ensure that test tasks and inputs are representative of the language and tasks in the TUL domain of interest. A further practical significance is that this study presents the way of applying the Rasch psychometric model to examine the psychometric quality of language tests, which help enhance the precision and accuracy of statistical estimation and provides several sources of empirical evidence in support of the validity of score interpretation and use. As is evident by this study, the Rasch measure approach is proven to be a cost-effective, time saving approach for test validation and is well mapped with the argument-based approach.

While most of prior studies primarily applied CTT to investigate item and test characteristics, far fewer studies used the Rasch model or other IRT models to examine the psychometric quality of language tests. The findings from this study could draw more attention to several time-saving, helpful applications of the Rasch measurement model to the assessment of collocational competence and other language abilities. Finally, the present study could raises the awareness of introducing collocations in English language instruction and material development in English classroom since awareness is considered as an important aspect of language learning. If the awareness of the importance of teaching and learning collocations increases, this implies that the use of collocation tests could potentially lead to the intended consequences in the form of positive washback.

1.7 Chapter summary

This chapter introduces several key components that rationalise, underline, and direct the process whereby this master's thesis was carried out from beginning to end. The remaining chapters that follow are concerned with the literature review, test development, research methodology, results and discussion, and conclusion of

this study. In chapter 2, I discuss in depth the literature review relevant to fundamental concepts and issues related to the development and validation of the ACCT. The interpretive argument, the first stage in the argument-based approach, is also developed in this chapter. Throughout chapter 3, I delineate in detail the process of test development based on fundamental concepts and the specified ACCT interpretive argument presented in chapter 2. Details of test development in chapter 3 provide some theoretical and empirical evidence in support of domain, evaluation, generalisation, and explanation inferences. In chapter 4, I describe in clarity the research methodology of the present study, including issues ranging from sampling design, measurement design, and analysis design. In chapter 5, I present and discuss the results from empirical data analysis. In chapter 6, conclusion of the study is presented and it deals primarily with the construction of the validity argument of the ACCT. Guidelines for responses to research questions are also presented in the final chapter.

CHAPTER 2
LITERATURE REVIEW

Chapter 2 presents the review of related literature that underlies and informs the development and validation of the ACCT. In this chapter, I describe several key issues and concepts related to purposes of the test, target language use domain, contemporary perspective on validity, argument-based approach to validation, Rasch measurement approach to validation, notion of collocation, item response design, conceptual framework of construct definition, and theoretical relationships of collocational construct. All these provide theoretical support to the validity argument. Before leaving this chapter with a chapter summary, I present a specification of the ACCT interpretive argument which outlines the interpretation and use of the ACCT scores through inferences, warrants, assumption, and potential evidence backing. The ACCT interpretive argument is the first step of the argument-based approach that need to be properly developed, for it helps direct not only how the ACCT is developed but what sources of evidence that need to be assembled to support the ACCT validity argument in the second stage.

2.1 Purposes of the test

The first and foremost step in language test development is to set a clearly-defined test purposes, for it directs the way in which the test, the interpretive argument, and the validity argument are to be developed (Bachman & Palmer, 1996, 2010; Kane, 2013; Stephen G Sireci, 2013; Wolfe & Smith, 2007a). It is thus of importance that the use of a test be clarified at the outset so that the validity of the test can be justified based on the conclusion drown from test scores. In the argument-based approach, the purposes of the test are also stated in the interpretive argument (Kane, 1992, 2006, 2011, 2013). A number of studies have so far developed collocational tests in order to assess collocational knowledge for a variety of purposes.

Some of previous studies were conducted with a view to explore to what extent L2 learners know collocations without administering any teaching methods (e.g., Jaén, 2007; Keshavarz & Salimi, 2007; Sadeghi, 2009; Webb & Kagimoto, 2011; Wolter & Gyllstad, 2011). Other studies developed collocation tests so as to assess to what extent L2 learners' collocational knowledge was enhanced after assigning

collocational interventions (e.g., Chan & Liou, 2005; Daskalovska, 2013; Goudarzi & Momi, 2012; Molina-Plaza & de Gregorio-Godeo, 2010).

Despite the growing number of studies on collocational knowledge assessment, most of collocational tests were developed primarily for experimental or exploratory purposes. It is only relatively recently that a few studies were set out with the main aim of developing and validating collocational ability tests particularly for placement decision. For example, Voss (2012) developed a computer-based ESL academic collocational ability test to serve as an admission or placement test. In his study, he used a gap-filling short answer format to elicit ESL learners' verb-noun collocational ability produced in an academic written English domain in English-medium universities.

In the light of this lack of collocation placement testing, the present study, therefore, seeks to develop the ACCT that can be used as a placement test or a supplement test of existing placement tests for informing decision about screening or placing students into appropriate English language courses in university or other institutions of higher education in the EFL context. The current ACCT is aimed specifically to measures a receptive dimension of academic collocational competence, which is part of vocabulary depth knowledge. The scores of the ACCT are interpreted based on a norm-referenced evaluation where students' performance is compared to one another in the group.

2.2 Target language use domain

In the realm of language assessment and evaluation, the concept of target language use (TLU) domain is of paramount importance to language test development. TLU domain specifies the context to which test scores are to be generalised. On this account, whether the interpretation of the test score will be meaningful or not depends to a very large extent on the identification of TLU. This is precisely due to the fact that language users or test-takers demonstrate their language ability or competence based on various kinds of interactions when they perform language use tasks in the TLU situation or domain. For this reason, test developers need to understand the nature of language use in the context of interest where test-takers' language ability are interpreted and generalised to (Bachman, 1990; Bachman & Palmer, 1996, 2010; Carol A Chapelle, 1998).

Bachman and Palmer (2010) classified TLU domain into two general types. One type of TLU domain involves a setting where language is used for the purpose

of language teaching and learning or a language teaching domain. The other type includes a setting where language is used for the purpose other than teaching and learning language and it is referred to as a real-life domain. When a language task is within a specific TLU domain, then it is called a TLU task. In developing an assessment tool, test developers are required to identify and describe a specific TLU domain of interest and develop one or more TLU tasks representative of and relevant to the corresponding TUL domain. The TLU domain of interest in this study falls into the language teaching domain since collocations are used for learning or academic purposes in university setting.

TLU is also considered as part of authenticity which is part of test quality. TLU significantly helps ensure that the language used in a test does represent the language used in the TLU context to which test scores are interpreted and generalised. In the past, it seems very hard indeed to obtain a sample of language that is sufficiently representative of the TLU domain and consequently the degree of test score validity can be questioned. At present, advances in technological tools and corpus linguistics make it possible for test developers to compile a large number of texts representing the TLU domain of interest or take advantage of corpora which contain large and representative collections of written or spoken language from different discourses. By using linguistic inputs from corpora, test developers can be confident that the degree of the validity of the test is enhanced as a result.

To date, a number of collocation tests have been developed using collocations from a variety of language use sources other than corpora (Chan & Liou, 2005; Kim, 2008; Laufer, 2011; Sadeghi, 2009; Sonbul & Schmitt, 2013). However, as several colossal corpora have come into existence nowadays, no small amount of research has thereby used collocation items sampled from the TLU corpora of interest. For instance, Jaén (2007) sampled adjective-noun collocation items from Bank of English and BNC to develop a general English collocational test. Webb and Kagimoto (2011) developed their general English verb-noun collocation tests using items form Bank of English and BNC as well. More recently, Voss (2012) sampled verb-noun collocation items from BNC to construct a test of collocational ability demonstrated in the TLU domain of academic written English. Very recently, Webb et al. (2013) developed a lexical verb-noun collocation test using collocation item from Bank of English.

It is evident from previous research that corpora have received more attention from language test developers since corpora provide a wealth of linguistic features and especially collocations that represent a TLU domain under study. By

virtue of corpus benefit, the present study, thereby, sampled high-frequency verb-noun collocations from the British National Corpus (BNC) which is expected to contain a collection of texts representing the academic written language in different academic areas. By using high-frequency collcoations from BNC, it is claimed that examinee performance on the ACCT would to a maximum extent reflect collocational competence demonstrated in the academic written English domain

2.3 Contemporary perspective on validity

Based on the validity literature, it can be concluded that the validity concept has now been shifted from the classical or traditional perspective to the modern or contemporary perspective which focuses validity on the degree to which existing theory and evidence support the proposed interpretation and use of test scores. From the classical perspective, validity is regarded as a property of the test and typically defined as the degree to which a test measure what it claims to measure (Akbari, 2012; Furr & Bacharach, 2014). This concept is based on different types of validity: face validity, content validity, criterion-related validity, and construct validity. The traditional validity concept is criticised as somewhat vague and does not stress the importance of social or consequential dimension of test score use (Akbari, 2012; Furr & Bacharach, 2014). It was also criticised as adding too much weight to psychometric and cognitive aspects by trying to make inferences to theories underpinning the traits or abilities under measure and to the way in which individuals possess or demonstrate such abilities (Akbari, 2012) .However, the traditional approach to validity is still considered as a necessary part of sound validity argument in the contemporary validity.

From the standpoint of the contemporary perspective, validity is refreshingly defined as the degree to which existing theory and evidence support the proposed interpretation and use of test scores and is regarded as a property of score interpretation and use. The historical development of the contemporary concept can be traced back to , Kane (1992, 2006, 2011, 2013), Messick (1994), and AERA, APA, and NCME (1999). Viewed from the contemporary perspective, it becomes clear that validity is concerned with the appropriate interpretation and use of test score, it is conceived as a matter of degree, and it is based on backing from empirical evidence and theory.

To achieve the contemporary validity, an effective approach to validation is thus called for. The argument-based approach to validity sequentially proposed by

Kane (1992, 2006, 2011, 2013), has become indeed in the foreground recently and increasingly acknowledged as a rigorous approach to improving validation and accomplishing validity based on the contemporary point of view (e.g., Brennan, 2013; Carol A Chapelle, 2012; Carol A. Chapelle et al., 2010; Carol A Chapelle et al., 2008; LeBaron Wallace, 2011; Oller, 2012; Stephen G. Sireci, 2007; Stephen G Sireci, 2013). In the sequent section, I discuss in detail the argument-based approach to validation proposed by Michael Kane, who is considered as one of the greatest validity theorist of our time.

2.4 Argument-based approach to validation

Since the focus of current validity has gone far beyond the traditional face, content criterion and construct validity aspects to encompass the appropriate interpretation and use of test score, a more appropriate validation approach need to be used in correspondence with contemporary validity. The traditional view that validity includes face, content, criterion, and construct evidence has been expanded by the current view of validity which focuses validity on the interpretation and use of test score and thus the validity of the interpretation and use of test score is based on various sources of existing evidence. In the light of this, the traditional types of validity are simply considered to be convenient categories for assembling evidentiary supports to the validity of score interpretation and use (Waugh & Gronlund, 2013). An argument-based validation approach, which provides the framework for evaluating the proposed claims based on test scores, has recently come into sharp focus in validity theory and has been acknowledged by several scholars (e.g., Brennan, 2013; Carol A Chapelle, 2012; Carol A. Chapelle et al., 2010; Carol A Chapelle et al., 2008; LeBaron Wallace, 2011; Oller, 2012; Stephen G. Sireci, 2007; Stephen G Sireci, 2013).

The argument-based approach has sequentially been introduced by Michael T. Kane, who is regarded as one of the greatest validity theorists and has published a series of papers on validity theory and the argument-based approach to validity (Kane, 1992, 2006, 2011, 2013). In addition, Kane's argument-based approach was put forward due to the fact that no agreement exists concerning a single best way to clearly define constructs of language proficiency to serve as a defensible basis for score interpretation. Various theoretical frameworks of language proficiency construct can be put as part of the argument-based validity (Carol A Chapelle, 2012; Carol A. Chapelle et al., 2010). As pointed out by Kane (2013), the argument-based approach does not require a strongly developed formal theory required by the construct

validity which is not often clear-cut, ambiguous, and debatable. The theory-based construct validity can nevertheless be included in the interpretive argument as part of backing for the claims. Therefore, an argument-based approach to validity does provide the general principles of construct validity without necessarily calling for formal theories and provide a facilitating framework for validation process.

In Kane's argument-based approach to validation, the test score is essentially of central interest by reason of its use in support of the claims made far beyond the observed performances. The claims that the test score needs to support involve test-takers' attributes, traits, or constructs as well as decisions or purposes of the test. It is sometimes misunderstood that validity is a property of the test. In fact, validity from Kane's sense is a property of the proposed interpretation and use of the test score. The interpretation and use that are sound and substantiated by proper and sufficient evidence are considered as having high validity. Conversely, in case that the interpretation and use do not make sense and lack appropriate and adequate evidence, the degree of their validity is open to question and debate as a consequence.

Based on the argument-based approach, validating the interpretation and use of the test score is actually to evaluate the possibility of the claims which relies hugely on the test score. The claims, thereby, need to be clearly defined in the forms of the proposed interpretation and use of the test score. To state the claims is to propose the interpretation and use of the test score and to evaluate those claims is to evaluate the extent to which those proposed interpretation and use of the test score are plausible. This indeed is necessarily the central concept of the argument-based validation. The interpretation and use of the test score are inextricably linked in practice and both direct the way in which the test is to be designed, developed and eventually validated.

The interpretation involves the claim concerning test-takers, while the use concerns the claim regarding decisions impacting on those test-takers. The claims concerning the proposed interpretation and use of the test score are developed in the process Kane called "the interpretive argument" in the argument-based approach (Kane, 1992, 2006, 2011). Later on, however, Kane (2013) coined the new term "the interpretive/argument, which modified the previous term "the interpretive argument" that pays too much attention to the interpretation of test scores. This study uses "the interpretive argument" to cover both the interpretation and use of test scores outlined as a network of inferences and assumptions necessitate backing.

Kane (2013) also pointed out that to make validation manageable, it is of great help to set a clearly-defined statement of the claims about the interpretation and use of test scores in order to know precisely what to be evaluated and how to evaluate those claims. One way to accomplish this is to develop the interpretive argument. The interpretive argument illustrates the interpretation and use of the test score proposed by test developers. The interpretive argument can be laid out in terms of the network of inferences and their assumptions leading from the test performances to the conclusions to be reached and to any decisions to be made based on those conclusions. Once the interpretive argument is well developed, meaning that the claims in the form of the proposed interpretation and use of the test score are relatively clearly stated, the interpretive argument provides the framework or direction for validation and criteria for the evaluation of the plausibility of the proposed interpretation and use of the test score. If the argument is coherent and complete and its inferences and assumptions are theoretically or empirically plausible, then the interpretive argument is proven possible and hence the validity argument is feasible as a consequence.

To conclude, from Kane's perspective on validity, validation is to validate test score interpretation and use by evaluating the feasibility of the proposed interpretation and use of test scores and thus clear statements regarding the proposed interpretation and use of test scores need to be made before they are evaluated in the validity argument stage. The degree of validity depends on the extent to which the assumptions of the proposed interpretation and use of the test score are sufficiently supported by sound theoretical and empirical support. The claims will determine the sorts of evidence needed for substantiating proposed assumptions, making passible the proposed interpretation and use of the test score in a particular context and at a particular time. The claims will determine the sorts of evidence needed for substantiating proposed assumptions, making passible the proposed interpretation and use of the test score in a particular context and at a particular time.

Moreover, Kane articulated in his latest article in 2013 that it is not possible to gather all evidentiary information to support validity in the process of developing and using the test, for validation is a lengthy or even endless process. This necessarily implies that the evidence needed for supporting the inferences and assumptions in the interpretive argument called for different amount of effort and time to gather, depending on how complex and demanding the proposed claims are.

Kane's argument-based approach involves two interdependent arguments steps to validity. The first step is to develop the interpretive argument through clearly stating the intended interpretation and use of test scores. The second step is to build the validity argument by analysing theory and evidence to evaluate the interpretive argument in terms of the feasibility of such intended interpretation and use of test scores.

2.4.1 Interpretive argument

As previously mentioned, the argument-based approach takes advantage of two sources of arguments: the interpretive argument and the validity argument. The interpretive argument specifies what is claimed in the proposed interpretation and use of the test score. In this way, it provides the framework for the validity argument where the proposed claims in the interpretive argument are evaluated. Therefore, to claim that the proposed interpretation and use of the test score is valid (validity argument) is to claim that the developed interpretive argument is clear, coherent, and complete enough through checking as to whether its inferences are logical and its assumptions are feasible. The interpretive argument specifies the intended interpretation and use of test scores by outlining a network of inferences and assumptions in the interpretive argument framework. In this way, the interpretive argument not only helps identify the sources of theory and evidence to support the intended interpretation and use of test scores, but also serves as the blueprint for designing and developing a test and the guideline for conducting research based on the argument-based approach. Test developers can develop the interpretive argument, while at the same time designing and developing a test. The interpretive argument can also be revised until it is well suited to interpretation and use of tests cores (Kane, 1992, 2006, 2011, 2013).

Once the interpretive argument is established, the validity argument can be built by evaluating how well the stated interpretation and use of test scores in the interpretive argument are properly supported by theory and evidence. In other words, the validity argument evaluates to what extent the proposed interpretation and use of test scores are valid or feasible based on theory and empirical evidence gathered. It can thus be said that the interpretive argument is of central to the argument-based approach and it needs to be well developed prior to others processes. In the argument-based approach, the interpretive argument can flexibly be developed in the sense that test developers can specify the network of inferences and their assumption in the interpretive argument. It can thus be

concluded that the interpretive argument can helps test developers to propose the interpretation and use of the test score through the inferences and assumptions related to such interpretation and use, provide the guidelines for designing and developing the test, identify the types of theory and evidence to gather in support of the inferences and assumptions, and even direct the way in which test-developing research is to be conducted.

2.4.2 Validity argument

Once the interpretive argument is developed, the validity argument can then be constructed. The validity argument provides the framework for an overall evaluation of the plausibility of the proposed claims stated in the interpretive argument. The degree of validity for the proposed interpretation and use relies heavily on how clear, coherent, and complete the developed interpretive argument is. Therefore, in the interpretive argument, test developers need to show that each inference is logical, each warrant is supported by assumptions, and each assumption is backed up by theoretical or empirical backing. The first and foremost step in building the validity argument is to conduct a conceptual analysis of the interpretive argument to see whether the interpretive argument is coherent in the sense that it gives the plausible rationale of the proposed interpretation and use and make sure that essential inferences and assumptions are included, acknowledged and investigated. The next step is to evaluate the warrants and their assumptions in the interpretive argument. Some assumptions may be based on theoretical review while some may be contingent on empirical studies. Certain backing may require more time and effort to gather if assumptions are more strong and complex (Kane, 1992, 2006, 2011, 2013).

It is clear that different warrants call for different sorts of backing. If a warrant rests on multiple assumptions, then it requires more types of backing as well. This means that the validity argument needs to provide sufficient backing for all of the inferences in the interpretive argument and again the process of validation is a lengthy or even endless process since the claims being made vary from case to case and from time to time. Consequently, the evidence and theory to support the claims also vary. It is still important to keep in mind that the validation process always involves two interconnected parts: the interpretive argument specifying the proposed interpretation and use of the test score and the validity argument evaluating such proposed interpretation and use of the test score.

2.5 Rasch measurement approach to validation

2.5.1 Concept of the Rasch measurement model

In the realm of measurement or psychometric theories, it can be said that there are three measurement or psychometric models for latent trait measurement: classical test theory (CTT), confirmatory factor analysis (CFA), and item response theory (IRT). IRT differs from CTT and CFA in that its unit of analysis is the item-level binary or polytomous data which are categorical in nature. IRT is widely acknowledged as a modern and superior alternative to CTT (Bachman, 2004; De Ayala, 2009; Embretson & Reise, 2000; Hambleton & Swaminathan, 1985; Reckase, 2009; van der Linden & Hambleton, 1997). Another family of IRT models is the Rasch model which focuses primarily on person ability and item difficulty parameters (Bond & Fox, 2007; Boone et al., 2014; Engelhard, 2013; Linacre, 2012; Rasch, 1960; Wilson, 2005). The emergence of the Rasch measurement model can be traced back to as far as 1960 when Georg Rasch developed a family of IRT models to develop measures of reading and develop tests for use in the Danish military. It was handed down thereafter to those well-known psychometricians such as Benjamin Write, David Andrich, Geoffrey Master, Graham Douglas, Mark Wilson, Richard Woodcock, Trevor Bond, and Christine Fox, who make a great contribution to Rasch-family models (Bond & Fox, 2007; Embretson & Reise, 2000)

One important advantage of the Rasch measurement model over CTT and CFA is that it applies nonlinear response mathematical models to simultaneously account for differences between persons and differences between items. Items and persons are put on the same latent metric and thus the probability of getting an item right depends at least on the subject's ability and the item's difficulty. In this way, the ability is interpreted relative to item performance, not just relative to other people in the sample. Unlike CTT, IRT-based item statistics are independent of respondents who complete the test and IRT-based estimates of respondents' ability are independent of the items that the participants answer (Bond & Fox, 2007; Boone et al., 2014; Engelhard, 2013; Linacre, 2012; Rasch, 1960; Wilson, 2005)

The Rasch measurement model was developed to analyse both dichotomous or polytomous item responses through separately estimating person ability and item difficulty. In other words, it involves measures of person ability and item difficulty, while holding other item parameters (discrimination and guessing) constant across all items. It was applied to item analysis for the purpose of modelling test characteristics specifically at the item level. In addition, the Rasch model makes use of a logistic

technique to estimate item parameters and person abilities into relative logit measurements, thereby enabling person ability and item difficulty to be compared on the common scale. Moreover, the Rasch model has three qualities that make it attractive and advantageous: the ease of use due to fewer parameters, fewer estimation problems due to fewer parameters, and the specific objectivity concerning the estimation of the item and ability parameters, which was the reason for its emergence (Bond & Fox, 2007; Boone et al., 2014; Engelhard, 2013; Rasch, 1960). Additionally, the Rasch measurement model computes individual measurement errors for persons and items, thereby providing clearer proscriptive diagnostics (Schumacker, 2004; Schumacker & Smith, 2007). In the Rasch model, the data must fit the model to possess the properties of specific objectivity and sufficiency (Bond & Fox, 2007; Embretson & Reise, 2000; Iramaneerat et al., 2008).

Figure 2.1 shows the model representing the relationship between two latent variables and one observed variable. Latent variables are person ability and item difficulty and an observed variable is a dichotomous response to a particular item. The model represents how the person ability and item difficulty influence the probability of the response to the item either correctly or incorrectly.

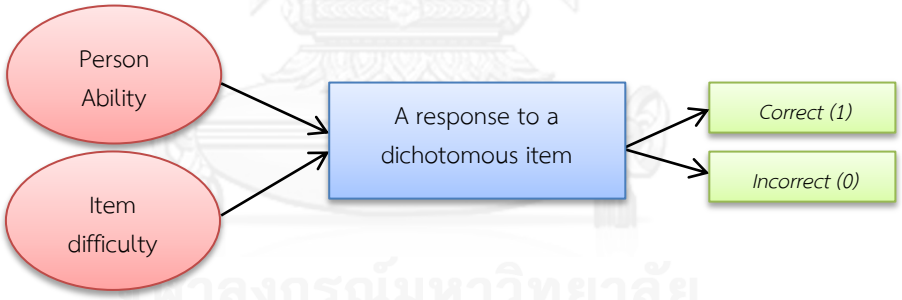

Figure 2.1. Model of the relationship between person ability, item difficulty, and a dichotomous response (modified from Embretson & Reise, 2000, p. 42)

2.5.2 Applications of the Rasch measurement model

A Rasch model offers several applications that can be used to provide empirical evidence supporting the inferences in the ACCT interpretive argument. In the domain inference, the point-measure correlation can be used to check the adequacy of item content and the congruency of a particular item with the remaining items on the instrument. The correlation should be positive to show the correlation between scores on the item and scores on the remaining items. The

value close to zero means that items are too easy or difficult to answer correctly or they do not measure the construct in the same manner as other items do (Wolfe & Smith, 2007b). The item fit indices can be used to investigate the unidimensionality of the items or other measurement problems. Item fit indices indicate whether the test content is relevant to the intended construct and assure that items elicit a relevant, unidimensional construct of interest, while misfit items may assess irrelevant, subdimensional constructs (Bond & Fox, 2007; Boone et al., 2014; Engelhard, 2013; Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b). The person-item variable map and the item strata index can be used to check the representativeness of the items. Noticeable gaps in the item difficulty hierarchy inform that certain area of the construct domain has not been covered by the test. Item difficulties should be widely spread, well matched with person abilities, and stratified into at least two levels to secure appropriate representativeness of the assessed content (Boone et al., 2014; Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b).

As for the evaluation inference, the principal component analysis of linearised Rasch residuals (PCAR) can be used to check the unidimensionality of the data by determining whether there is a sufficient amount of variance explained by the construct in question. If the data fit the model, it can then be confident that item scoring is appropriate for eliciting the construct under measure (Wolfe & Smith, 2007b). As for scoring, the dichotomous Rasch model scales observed scores into comparable measured scores, hence contributing to the standardisation of scoring process (Aryadoust, 2009; Boone et al., 2014; Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b). Transforming raw scores to measured scores in the Rasch analysis is of fundamental importance, for the distance between measured scores is equal and thereby item difficulties can be compared with person abilities (Bond & Fox, 2007; Boone et al., 2014; Engelhard, 2013; Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b). The person-item variable map can be used to check the appropriateness of norm-referenced interpretation. Linacre (2012) suggested that the distribution of person ability should relatively match the distribution of item difficulty in order to be appropriate for norm-referenced interpretations. Point-measure correlation coefficients exceeding 0.3 are appropriate for norm-referenced evaluation (Wolfe & Smith, 2007b).

In respect of the generalisation inference, the Rasch measurement model can also calculate reliability estimates of scores under different test circumstances. In contrast to CTT methods (i.e., KR-20 and a coefficient alpha) that use the variance for

an average sampled person, the Rasch measurement model should yield a better estimate of internal consistency because the numerical values are linear if the data fit the model, the actual average error variance of the sample is used lieu of the error variance of an average person, and Rasch-based methods typically compute reliability without regard to extreme scores (Schumacker, 2004; Schumacker & Smith, 2007).

Iramaneerat et al. (2008) and Wolfe and Smith (2007b) suggest that the item reliability informs how well examinee abilities spread out items difficulties or how well item difficulties are dispersed along the difficulty hierarchy. The item separation supplements the item reliability by checking how well items are classified into different levels on the item difficulty hierarchy. Another useful index is the item strata index which indicates whether person competencies statistically distinguish item difficulty levels. The person reliability (analogous to coefficient alpha and KR-20) can be employed to check how well item difficulties spread out examinee abilities or how well competencies are distributed along the competence hierarchy. The person separation supplements the person reliability by examining to what extent persons are separated into different competency levels on the competency hierarchy. The person strata index also indicates how well items statistically discriminate competence levels.

The strata index for person and item are calculated using the following formula: strata = $(4G_{Sep} + 1) / 3$, where G_{Sep} is the separation index. The item strata index informs the number of statistically distinct levels of item difficulty that a particular group of examinees could distinguish, while the person strata index indicates the number of statistically distinct levels of person competency that a particular set of items could distinguish (Wright & Masters, 1982, 2002). The higher the value of separation indices, the more spread out the persons and items are on the construct being measured (Linacre, 2012; Schumacker, 2004; Schumacker & Smith, 2007; Wolfe & Smith, 2007b). The person-item variable babble maps also provide visual information regarding the degree of instrument assessment precision for a particular group of examinees (Baghaei, 2008; Linacre, 2012).

Concerning measurement invariance, differential test functioning (DTF) can be performed to detect whether items function psychometrically invariantly for males and females on the test level and differential item functioning (DIF), on the item level, can be used to check the invariance of item quality across gender. DTD and DIF manifest when a particular item has different difficulty measures for males and females (Linacre, 2012; Wolfe & Smith, 2007b). DIF analysis is a method of

determining whether test items function differently across subgroups of test-takers upon controlling for person ability level. Results from DIF analysis can be used to evaluate validity arguments of the interpretation and use of test score. It is important to note, however, that empirical evidence of differential performance is necessary, but not sufficient to draw the conclusion that bias is actually present. The conclusion of bias goes beyond the empirical data, while DIF is typically used to describe the empirical evidence found in the investigation of bias (Boone et al., 2014; Hambleton, Swaminathan, & Rogers, 1991). The Rasch measurement model can thus be employed to examine invariance of item calibrations that are necessary to detect differential item functioning.

The hypothetical concept of DIF is that test items should not behave differently for particular subgroups (such as ability, gender, and ethnicity subgroups). If an item functions differently for certain groups, then the item decreases the validity of the measure for a construct, thereby giving rise to undesired test fairness. As Engelhard (2013) and Wright and Masters (1982) pointed out, meaningful comparisons of person measures can merely be drawn only when the item calibrations are invariant from one group to the next. It is necessarily of essence to investigate whether all items of assessment tools function in a similar fashion across subsamples. The present study employed a Rasch-based DIF analysis to ascertain whether all items of the ACCT function differently for gender subgroups (male and female).

With respect to the explanation inference, Wolfe and Smith (2007b) recommend that the multiple-choice distractor analysis inform whether responses to distractors are consistent with the intended cognitive process around which distractors are constructed. The examinee proportion (p-value) choosing each distractor indicates whether distractors equally attract a sizeable examinee proportion. Each distractor should attract at least 5% of the examinee proportion and should not attract a larger proportion than the correct choice. The average ability of respondents choosing each distractor determines the degree to which the option discriminates between respondents. On average, each distractor should be chosen by lower-ability persons, while the correct option should be selected by higher-ability persons. The distractor-measure correlation indicates whether a particular distractor is selected by lower-ability examinees. The distractor-measure correlation should be negative to indicate that lower-ability respondents choose that distractor more than higher-ability examinees. Linacre (2012) and Wolfe and Smith (2007b) suggest that the item difficulty distribution in the person-item variable map,

the item fit statistics, and the principle component analysis of Rasch residual (PCAR) all gives useful information on the relevancy and unidimensionality of the construct being measured.

Regarding the extrapolation inference, the person-item variable map provides visual information as to whether the instrument may detect change in the future (Wolfe & Smith, 2007b). The person competency distribution should be widely dispersed on the latent competency scale and well matched with the item difficulty distribution. Another indication is the person strata index which informs how well items statistically classify person abilities. The person strata index greater than 2 suffices to confirm that items distinguish the more competent from the less competent. Although the Rasch model has long taken its place in language testing (McNamara & Knoch, 2012), only a few collocation tests has been validated using the Rasch measurement approach, while much more vocabulary tests has been evaluated using the Rasch model and Messick' validity framework (e.g., Baghaei & Amrahi, 2011; Beglar, 2009). Voss (2012) conducted his dissertation to develop a collocation test and he used the Rasch model and an argument-based approach to build a sound validity argument for the test. However, the use of Rasch statistics was focused on item fit statistics.

While a lot of collocation assessment tools have been developed based primarily on CTT perspective on the one hand, little interest is taken in exploiting advanced IRT psychometric methods to validate the psychometric quality of the collocational test on the other. Investigating the psychometric quality of assessment tools is probably a challenging burden that many test developers have to come to shoulder. This is precisely due to the fact that CTT is more practical for most test developers while IRT or Rasch methods require more advanced knowledge and effort as well as a sufficient number of samples. There is no doubt then that much research on developing collocation tests applied CTT to validate psychometric quality of the tests, whereas little research validated psychometric properties of collocation tests using IRT models (e.g., Voss, 2012). With this in mind, the present study applied the Rasch psychometric model for dichotomous scoring method to investigate and enhance the psychometric properties of the ACCT, designed to measure a receptive knowledge of EFL learners' collocation competence, which is part of vocabulary depth and writing abilities.

2.6 Notion of collocation

2.6.1 Definition of collocation

The concept of collocation was initially introduced by Palmer and was sequentially brought into prominence by Firth in (1957). A large body of literature reveals that there are different approaches to the study of collocation and hence collocation can be defined in different ways. It is commonly recognised that collocation can be broadly defined based upon either a lexical approach or a frequency approach. The phraseological approach to collocation study is employed by those well-known scholars, for example, Carter (1998) Cowie (1998), and Howarth (1998), a while the frequency-based approach to collocation is deployed by such leading authorities as Nesselhauf (2003, 2005) and Sinclair (1991).

The frequency-based approach typically regards a collocation as a co-occurrence of words within a certain distance of each other. Nesselhauf (2005) mentioned that collocations are viewed as being co-occurrences that are more frequent than could be expected if words combined randomly in a language. The frequency-based approach was very much developed and made known by Sinclair, who in turn based his own notion of collocation on Firth (1957). This approach does not regard collocations as belonging to a distinct linguistic category but rather defines collocations in terms of probability. In the frequency-based approach, the strength of a particular word combination or collocation is assessed on the basis of how frequently it appears in a large representative sample of discourse. In this way, only certain combinations or collocations are much more likely to occur than others. That is to say, the frequency-based approach uses statistical criteria to define collocations.

Studies using the phraseological approach normally use lexical criteria to determine whether a particular combination can be classified as a collocation or not. The phraseological approach tends to formulate collocation categories according to phrasal characteristics exhibited by different word combinations and views collocation as exhibiting a degree of ‘fixedness’, “restriction” and/or “a lack of meaning transparency”. For instance, Carter (1998) drew remarkably upon the criterion of the degree of commutability concept in order to divide collocations into four categories: unrestricted, semi-restricted, familiar, and restricted collocations. According to Howarth (1998), word combinations can be classified into free combinations, restricted collocations, and idioms. His collocation continuum (as in Figure 2.2) provides fundamental concept in consistence with Sinclair (1991)’s open choice and idiom principles that distinguish different types of word combination on

the continuum as shown in Figure 2.2. These two principles are combinations of words or lexical composites chosen to form meaning. Based on the open choice principle, the interpretation of the meaning of words combined freely is far more transparent than the interpretation of the meaning of words combined according to the idiom principle. The meaning of each word in free combinations (e.g., *blow a trumpet*) is clear and understandable individually, whereas the first constituent "*blow*" in "*blow a fuse*" as an idiom has different meaning from the core meaning of "*to blow*". The intended meaning of this multiword lexical item is different from the original meaning of each individual lexical item in such multiword lexical item.

In addition, restricted co-occurrence differentiates collocations from free combinations in the sense that individual words are easily substituted or replaced in accordance with grammatical rules. Examples of restricted collocations are the following: rain collocates with heavy but not with strong; discussion collocates with have or hold but not with deliver; and speech collocates with deliver but not with hold. Therefore, heavy rain, hold/have discussion, and deliver speech are considered as restricted collocations. As in Figure 2.2, restricted collocations can further be divided into strictest, strict, and liberal applications based on the main criterion of commutability. Strictest application allows no substitution of either verb or noun element (e.g., *curry favour*), strict application allows some substitution of either verb or noun element (e.g., *pay/take heed and give the appearance/impression*), and liberal application permits limited substitution in both elements (e.g., *introduce/table/bring forward a bill/an amendment*). In this study, collocation is defined based primarily on the phraseological approach which describes collocations as habitually occurring lexical combinations that are characterised by restricted co-occurrence of elements and relative transparency of meaning.

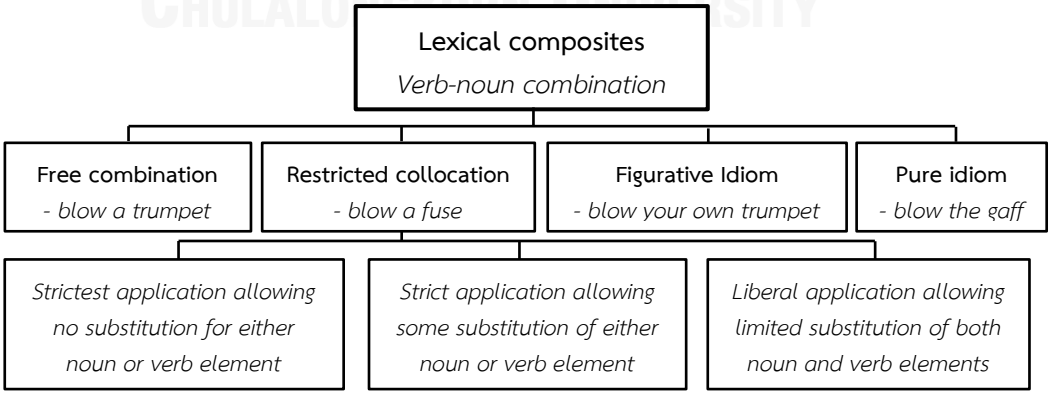

Figure 2.2. Collocation continuum (modified from Howarth, 1998)

2.6.2 Classification of collocation

It is widely acknowledged that EFL learners must learn how words combine or collocate with each other if they wish to express themselves accurately, fluently, and naturally in their language performance (Benson et al., 2010; Lewis & Conzett, 2000; Nation, 2001; O'Dell & McCarthy, 2009; Read, 2000, 2007; Read & Chapelle, 2001; Schmitt, 2004a, 2004b, 2010). This means that EFL learners need to pay special attention to how words are combined into phrases, sentences and discourses. Prior findings have established that EFL learners typically have difficulty using lexical collocations rather than grammatical ones in their language production. In particular, a verb-noun collocation is found to be a difficult collocation type for L2 learners' written and spoken language production (Ganji, 2012; Laufer, 2011; Laufer & Waldman, 2011; Marco & José, 2011; Miyakoshi, 2009; Molinaro, Canal, Vespignani, Pesciarelli, & Cacciari, 2013; Nesselhauf, 2003, 2005).

According to Benson et al. (2010), collocations are of two categories: grammatical collocations and lexical collocations. Grammatical collocations, as presented in Table 2.1, consist mainly of a dominant word (noun, adjective, and verb) and a preposition or grammatical structure (infinitive or clause). Lexical collocations, by contrast, typically do not contain prepositions infinitives, or clauses. As can be seen from Table 2.2, lexical collocations consist of nouns adjectives, and verbs. These two categories exemplify the kind of collocational knowledge native speakers of English have in common. Following phraseologists as well as Benson et al. (2010), the present study thus aims to investigate a verb-noun lexical collocation which is characterised by restricted co-occurrence of elements and relative transparency of meaning as already mentioned previously.

Table 2.1
Some examples of grammatical collocations (modified from Benson et al., 2010)

No.	Rules	Examples
1	noun + preposition	- The blockade of enemy ports by the US navy.
2	noun + to + infinitive	- Students made an effort to do the test.
3	noun + that-clause	- He took an oath that he would do his duty.
4	preposition + noun	- We discovered the species by accident.
5	adjective + preposition	- Teachers were very angry at students.
6	adjective + to + infinitive	- We are ready to go swimming.
7	adjective + that-clause	- It is crucial that students be placed properly.

Table 2.2

Some examples of lexical collocations (modified from Benson et al., 2010)

No.	Rules	Examples
1	verb + noun	- He does the laundry once a week.
2	adjective + noun	- There was a heavy rain last night.
3	noun + verb (action)	- Problems arose after the conflict.
4	noun (unit) + of + noun	- Peter gave Mary a bouquet of flowers.
5	adverb + adjective	- Two arguments are inextricably linked.

2.7 Item response design

2.7.1 Multiple-choice item response format

Different authors adopt different classification system or scheme when it comes to categorising test items. A more direct approach is to classify items as either selected-response or constructed-response formats. According to Reynolds, Livingston, and Willson (2008), if an item requires test-takers to select a response from available alternatives, it is classified as a selected-response item. Examples of this kind of item are a multiple choice item, a true-false item, and a matching-item. On the one hand, if an item requires examinees to create or construct a response, it is classified as a constructed response item. Essay and short-answer items are examples of a constructed-response item. There are strengths and weaknesses of either a selected-response format or a constructed-response format. The present study uses a selected-response format with specific focus on a multiple-choice format and therefore only a multiple-choice format is discussed in detail in the following section.

The multiple-choice item is generally recognised as the most practical and useful type of the objective test item. There are a number of advantages of a multiple-choice format. It can effectively assess many of simple learning outcomes measured by the short-answer item, the true-false item, and the matching exercise. It can also measure some of the more common complex learning outcomes in relation to knowledge, understanding, and application areas (Haladyna, 1994; Haladyna, Downing, & Rodriguez, 2002; Haladyna & Rodriguez, 2013; Miller, Linn, & Gronlund, 2008; Reynolds et al., 2008; Waugh & Gronlund, 2013). In addition, the multiple-choice item is also adaptable to most types of subject-matter content. As such, many standardised tests use multiple-choice items. Carr (2011) pointed out that in a discrete-point item, if test-takers reply an item wrong, it is assumed that they lack

ability in a specific area and thereby it is of great use when a measure of interest is a very specific knowledge of language. Strengths and weaknesses of a multiple-choice test item are shown in Table 2.3.

Table 2.3
Strengths and weaknesses of a multiple-choice format (modified from Waugh & Gronlund, 2013)

Strengths	Weaknesses
- Learning outcomes from simple to complex can be measured. - Highly structured and clear tasks are provided. - A broad sample of achievement can be measured. - Incorrect alternatives provide diagnostic information. - Scoring is easy, objective and reliable.	- The format tests only recognition knowledge. - Guessing may considerably affect test scores. - It is difficult to write successful items. - It is frequently difficult to find plausible distractors. - It is ineffective to measure some types of problem solving and the ability to organise and present ideas.

A multiple-choice item consists of a stem presenting a problem situation and alternatives (otherwise called choices or options) providing solutions to the problem. The problem may be stated as a direct question or an incomplete statement and the solutions may include words, numbers, symbols, or phrases. Alternatives include both a correct answer or the best answer and several plausible wrong answers called distractors. In using the best answer, care must be taken, however, to ascertain that the best answer is the one agreed on by experts so that the answer can be defended as clearly the best. The best-answer type of multiple-choice item is likely to be more difficult than the correct-answer type. A multiple-choice format is a receptive or selective response item in that test-takers choose from a set of responses in lieu of producing a response and therefore elicit test-takers' recognition knowledge which is a receptive aspect of lexical competency.

Previous studies used a multiple-choice format to elicit receptive dimension of vocabulary and collocational knowledge (Gyllstad, 2005, 2007; Jaén, 2007; Keshavarz & Salimi, 2007; Webb & Kagimoto, 2011; Webb et al., 2013). Despite the wide application of the multiple-choice item mentioned earlier, there are learning outcomes such as the ability to organise and present ideas that cannot be effectively

measured with any form of selection item. Another classic problem of using a multiple-choice item is that test-takers have the probability of guessing the correct answer. There is no exact number of alternatives to use in a multiple-choice item. Typically, three, four, or five choices are used. However, as presented in Table 2.4, there are chances of guessing the correct answers in three, four, or five choices.

Table 2.4.
Chances of guessing the correct answers (Reynolds et al., 2008)

Number of choices	Chances of a correct guess	Chance score of 100 items
Five-choice items	1 in 5	20
Four-choice items	1 in 4	25
Three-choice items	1 in 3	33

It is suggested by several scholars that a five-choice item test is used to reduce the chances of guessing the correct answer (e.g., Bachman & Palmer, 1996; Bachman & Palmer, 2010; Haladyna, 1994; Haladyna & Rodriguez, 2013; Reynolds et al., 2008; Waugh & Gronlund, 2013). Reducing the chances of guessing the correct answers by adding alternatives enhances reliability and validity, but only if all the distracters are plausible and the items are well-constructed (Reynolds et al., 2008). Precisely for this reason, the present study uses a five multiple-choice item test to measure test-takers' receptive knowledge of English verb-noun collocation, demonstrated in an academic written discourse at a university setting.

In the current research, the ACCT is designed particularly to assess a receptive dimension of collocational competence which requires test-takers to recognise lexical items and thereby a multiple-choice format is used as an item response format in the present study. There are particular reasons why the present study uses a multiple-choice format. Firstly, the construct to be measured in this study is a verb-noun collocational competence which is a very specific linguistic trait and thus a discrete multiple-choice item is suitable for assessment of a very specific construct under investigation. Secondly, most of EFL learners in Thailand are more familiar with standardised or high-stage multiple-choice tests. Therefore, using a task format with which test-takers are more familiar may not affect their performance on the test since task characteristics may impede the way in which test-takers perform on the test as well (Carr, 2011).

Thirdly, previous studies indicate that EFL learners acquire vocabulary receptively through reading, listening, and teaching in class (Webb, 2005, 2008). It is

thus more appropriate to measure receptive knowledge that learners have acquired and used rather than productive knowledge which is in reality less gained and used in the EFL contexts. This is evident partly from a small-scale trialling of the gap-filling productive test with low proficient EFL students. The test is probably too difficult for low proficient test-takers to elicit a productive knowledge of collocational competence. For a test to provide much information on test-takers' knowledge, the difficulty of the test should be matched with the ability levels of test-takers.

Finally, a multiple-choice format is widely acknowledged as a more practical format for a standardised large-scale test, in particular a placement test where placement decision need to be made as soon as possible before or during the beginning of the courses so that teachers can decide who should or should not take more courses and which proficiency level students should be on the basis of norm-referenced evaluation. A five-item multiple-choice test is used to elicit test-takers' receptive knowledge of verb-noun collocation competence expressed in EFL academic written context.

2.7.2 Dichotomous response scoring method

Methods of scoring item responses can be classified into two types: dichotomous and polytomous scoring. In binary or dichotomous scoring, item responses are scored into two categories to represent, for example, success (1) or failure (0) or represent true (1) or false (2). Although ability or achievement items are typically binary or dichotomous data, there are situations where other types of items are perhaps more appropriate, for example, rating scales, which are scored into more than two categories. If information regarding the ability or trait is lost by binary scoring, then a polytomous scoring should be better taken into consideration. Many polytomous scoring models are available for scoring item responses unless binary scoring may not be appropriate and may not perhaps provide accurate ability information (Embretson & Reise, 2000).

In this study, a dichotomous scoring (correct and incorrect) is employed to score a multiple-choice test. The dichotomous scoring method is based on the target responses that were identified in the collocation identification process. If test-takers choose a correct answer, they would get a full mark (1). If, on the other hand, they select an incorrect alternative, they would get no mark (0). A multiple-choice item consists of five options. Test-takers have to choose the best answer among five

options in order to gain 1 mark for that item. Answer keys for a multiple-choice test is also provided.

2.8 Conceptual framework of construct definition

As mentioned earlier, the argument-based approach does not focus on a theory-based construct definition. However, a theory-based construct definition can be included as part of the interpretive argument for enhancing sound validity argument. The present study sets out to measure English verb-noun collocation competency in the context of academic written genre. I base the conceptual framework of construct definition on an interactionist approach proposed by Carol A Chapelle (1998). An interactionist approach to construct definition posits that "*performance is viewed as a sign of underlying traits, and is influenced by the context in which it occurs, and is therefore a sample of performance in similar contexts."* An interactionst-based construct definition involves a trait-oriented perspective and a context perspective.

Drawing upon the interactionist perspective, the construct to be measured is thereby defined as a collocational competency demonstrated in the context of academic written English. In this sense, the performance of the current ACCT is a reflective indicator of collocational competence and a representative sample of the collocational performance produced in the academic written context and other related contexts alike. Also of interest in an interactionist approach is the metacognitive competence lying behind test-takers' behaviour or characteristics expressed in the context. It is through metacognitive strategies that test-takers use to appraise language use context and produce language that is proper to such context.

Viewed from an interactionism perspective, it can thus be concluded that the construct definition encompasses linguistic competence, contextual competence as well as strategic competence. This definition is consistent with the theoretical definition model of communicative language ability, proposed by Bachman and Palmer (1996, 2010). The intended score interpretation would be that the performance on the ACCT is supposed to reflect restricted collocational competency in academic written English, properly produced using metacognitive strategies. In this sense, a theory-based construct definition helps ensure that test scores are interpreted as an indicator of collocational competence. By reason of strict time constraint, investigating test-takers' meta-cognitive strategies is sufficiently investigated this study. The construct definition of this study, therefore, focuses

considerably on linguistic and contextual competence and only these two perspectives are discussed in depth in the following section.

2.8.1 Collocation definition

In this study, linguistic competence is collocation knowledge which is defined based on the phraseological approach (Carter, 1998; Cowie, 1998; Howarth, 1998). As described previously, the phraseology-based definition defines collocations as habitually occurring lexical combinations that are featured by two principal criteria, restricted co-occurrence of words and relative transparency of semantic meaning. Restricted combination of words differentiates collocations from free combinations on the ground that the individual words in free combinations are easily substituted or replaceable in accordance with grammatical rules. Examples of restricted collocations are the following: *rain* collocates with *heavy* but not with *strong*; *discussion* collocates with *have* or *hold* but not with *deliver;* and *speech* collocates with *deliver* but not with *hold*. Therefore, *heavy rain*, *hold/have discussion*, and *deliver speech* are considered as restricted collocations.

Relative transparency of semantic meaning, on the other front, distinguishes collocations from idioms on the ground that the meaning of idioms is far less transparent than that of collocations and is often very unclear because it cannot be deciphered simply from the words that compose idioms. Relative semantic transparency is illustrated by the following example: *face* in *"face a problem"* is not used with its original meaning, but the semantic meaning of *face* in *"face a problem"* is at least partially relevant to its original meaning, and the expression of *"face a problem"* is a great deal clearer than *"face the music"*, which is an idiom that means show courage. In this regard, when two words are combined to form a collocation, such words are much more likely to co-occur than others and the semantic meaning of two words remains relatively the same. Drawing on a phraseological perspective, the present study defines a collocation as "a habitually occurring combination, characterised by restricted co-occurrence of elements and relative transparency of meaning".

2.8.2 Academic discourse context

Contextual or pragmatic competence is the ability to know the context under which collocations are used. The target context or target language domain in which the language knowledge is used need to be taken into account when it comes to

defining construct to be measured from the point of view of an interactionist approach. In the present study, the context in which collocations are used and to which test scores are generalised is academic written English in universities or other institutions at tertiary level. A corpus, a large collection of textual data representing the target language use domain is of great helps ensure that verb-noun collocations sampled from the TLU corpus maximally represents the general academic written English discourse. By virtue of this, the British National Corpus (BNC) was chosen as it represents a large, representative source of general written academic language.

BNC contains a large number of texts from which the target language use domain can be investigated as a sub-corpus. This written academic sub-corpus of BNC consists of just about 16 million running words and is the largest collection of written academic texts at the time of the development of the ACCT. Lancaster BNCweb was used a tool for domain analysis and collocation sampling in the current research. Sub-corpora of the academic written sub-corpus of BNC were created to represent academic written text in different academic disciplines where representative collocations were extracted. To ensure that high frequent collocations do not represent only a few disciplines, a systematic sampling was employed to sample collocations that are representative of all academic disciplines in the academic written English. In this way, the test would contain not only high frequent collocations but also faire and representative collocations found in all academic fields. By doing this, topical or content knowledge does not affect test performance of students from different academic disciplines.

The current study sampled collocations based on the phraseological method of corpus identification using a frequent word-based approach (Gyllstad, 2005, 2007; Jaén, 2007). A frequent word-based approach begins by identifying a list of high frequent words selected prior to searching their collocate words which are confirmed thereafter as valid collocations by particular criteria set. The sampling method in this study is based on a word-list phraseological approach and a systematic sampling that sample collocations from seven academic sub-corpora in BNC through Lancaster BNCweb. The conceptual frameworks of construct definition in the present study is summarised in Table 2.5.

Table 2.5

Summary of an interactionist-based collocation construct definition (Carol A Chapelle, 1998)

Dimension	Variable	Description
Collocation aspect	Collocation definition	Based on a phraseological approach (e.g., Carter, 1998; Cowie, 1998; Howarth, 1998), a collocation is defined as occurring combination that are characterised by restricted co-occurrence of elements and relative transparency of meaning.
	Collocation type	Drawing on Benson et al. (2010) the present study focuses on a lexical collocation type with a mere emphasis on a verb-noun collocation type.
	Collocation knowledge	A receptive dimension of collocational competence in knowledge of vocabulary depth
Context aspect	Setting	The test is used for postgraduate or graduate studies in universities or other institutions of higher education
	Text type/discourse	Academic written English
	Subject matter/topic	Applied science, art, belief and thought, commerce and finance, natural and pure sciences, social science, world affairs
	English user	Postgraduate or graduate students who study English as a foreign language (EFL)
Cognition aspect	Metacognitive strategies	Metacognitive competence underlying test-takers' behaviour or characteristics expressed in the context assist test-takers in appraising language use context and produce language proper to such context.

2.9 Theoretical relationship of collocational construct

One way to evaluate construct validity for sound validity argument is to examine whether scores on the ACCT correlate positively to other tests of English language proficiency related to the construct and other measures of academic language performance. Theoretical relationship of collocational construct (aka nomological network) is one sort of evidence to substantiate the extrapolation inference in the interpretive argument This section discusses previous studies that revealed evidence of nomological construct network between collocation knowledge, vocabulary knowledge, and reading comprehension. It is clear from the literature that vocabulary knowledge and reading comprehension are positively correlated. Several studies have thus far explored the relationship between English vocabulary knowledge and reading comprehension of L2 learners. For example, Qian (1999) found a significant positive correlation between the scores of vocabulary size test and academic reading comprehension test and his later study (2002) also found the significant positive relationship between vocabulary knowledge and academic reading performance.

Recently, Baleghizadeh and Golbin (2010) investigated the effect of vocabulary size on reading comprehension of Iranian EFL learners and found that there is a very significant relation between vocabulary size and reading comprehension. Yamamoto (2011) examined the effect of reading combined with writing task on productive vocabulary growth of Japanese university students. The result of the study indicated that reading combined with writing task help retain receptive and productive vocabulary knowledge. Another study conducted by Chen (2011) explored the relation between EFL students' vocabulary breadth knowledge and literal reading comprehension and discovered that vocabulary breadth knowledge was significantly positively correlated to literal reading comprehension.

Very recently, Voss (2012) investigated the relationship between ESL learners' academic collocational knowledge and academic reading comprehension as well as academic collocational knowledge and academic vocabulary size. He found that collocational knowledge, which is part of vocabulary depth knowledge, had a significantly positive relationship with vocabulary size knowledge and reading comprehension. He also explained in a very clear manner the relationship amongst vocabulary knowledge, collocation knowledge, and reading comprehension. Voss clearly described the relationship among the constructs in a nomological network in the following formulas "R/VS > VS/VD > R/VD" It is predicted that the relationship between reading (R) and vocabulary size (VS) is stronger than the relationship

between vocabulary size (VS) and vocabulary depth (VD). Vocabulary depth is predicted to be represented by collocational ability in his study and the relationship between reading and vocabulary depth or collocational ability has the weakest relationship in the nomological network.

Figure 2.3 presents the theoretical relationships of collocational construct based on prior research. It is hypothesised that test-takers who have high vocabulary size knowledge and high reading competence ability are also very likely to possess high vocabulary depth or collocational knowledge. Therefore, if test-takers do well on the ACCT, they are supposed to do well on vocabulary size and reading comprehension tests. In the light of what discussed previously, the present study explores the relationship between receptive collocational competence, measured by the ACCT, and receptive vocabulary breadth knowledge, assessed by the Academic Vocabulary Level Test (AVLT) (Schmitt, Schmitt, & Clapham, 2001). If a significant correlation is found between the ACCT and the AVLTI ACCT, it can thus be more confident that the ACCT measures more accurately the latent construct of collocational competence. The investigation of collocational construct relationship provides partial empirical evidence in support of the explanation inference for enhancing the validity argument of the ACCT. Due to strict time constraint, exploring the correlation between scores on the ACCT and scores on a reading comprehension test is beyond the scope of the present study.

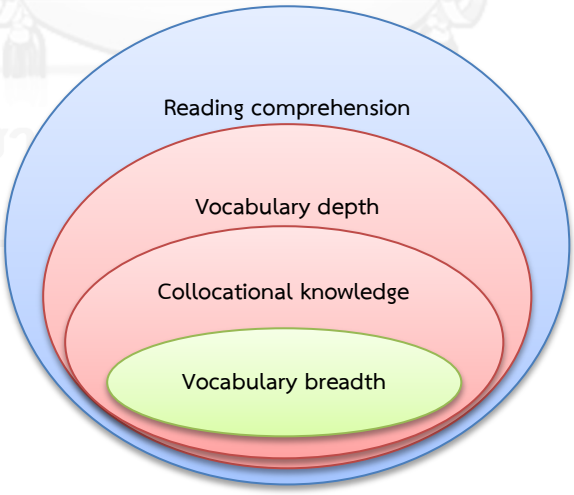

Figure 2.3. Theoretical relationships of collocational construct (modified from Voss, 2012, p. 46)

2.10 Specification of the ACCT interpretive argument

Following the argument-based approach (Kane, 1992, 2006, 2011, 2013), the validation process of the ACCT involves two stages. The first stage is to develop the interpretive argument specifying the proposed interpretation and use of test scores and the second step is to develop the validity argument evaluating the interpretive argument and plausibility of the proposed interpretation and use of test scores. The validity argument provides answers to the three research questions addressed earlier. The ACCT interpretive argument follows the TOEFL interpretive argument framework (Carol A Chapelle et al., 2008) which lays out six types of the inference: domain description, evaluation, generalisation, explanation, extrapolation, utilisation, and consequence.

The ACCT interpretive argument specifying the interpretation and use of the ACCT scores is based finally on seven types of inferences: domain description (henceforth referred to as domain inference), evaluation, generalisation, explanation, extrapolation, utilisation, and consequence. Each inference has its warrant which rests on the assumptions requiring different kinds of backing either theoretically or empirically. The interpretive argument provides the framework for proposing the interpretation and use of the ACCT score through laying out the inferences and assumption related to the proposed interpretation and use, providing the guidelines for designing and developing the ACCT, identifying the types of evidence to gather in support of the inferences and assumptions, and guiding the way in which the current research is conducted. Once the ACCT interpretive argument is relatively fully developed, the ACCT validity argument can then be established based on the evaluation of the ACCT interpretive argument.

It is important to realise that gathering all evidentiary information to support validity is not possible in the process of developing and using the test, for validation is a lengthy or even endless process (Kane, 2013). It is therefore impossible for this study to considerably investigate the utilisation and consequence inferences to support validity in the process of developing and using the ACCT. As such, investigating the consequence inference thus beyond the scope of this study.

2.10.1 Specifying the domain inference

Table 2.6 shows specification of warrants, assumptions and potential baking for the domain inference. The domain inference warrants that observations of performance on the ACCT reflect the collocational competence representing the

TLU domain of academic written English in universities or other institutions of higher education. This warrant assumes that: 1) performance on the ACCT reflects collocational competence which contributes partly to performance on the academic English writing task, 2) collocations on the ACCT are representative of the TLU domain of academic written discourse, and 3) the ACCT can elicit test-takers' performance reflecting collocational competence. These assumptions require potential backing from TLU domain and corpus analysis, systematic collocation sampling, Rasch person-item variable map, item response development, Expert review of the test, Rasch unidimensionality analysis, and Rasch item strata estimation.

Table 2.6
Summary of backing in support of the assumptions underlying the warrant of the domain inference

Warrant	Underlying assumptions	Backing evidence
Observations of performance on the ACCT reflect the collocational competence representing the TLU domain of academic written English in universities or other institutions of higher education	1) Performance on the ACCT reflects collocational competence which contributes partly to performance on the academic English writing task. 2) Collocations on the ACCT are representative of the TLU domain of academic written discourse. 3) The ACCT can elicit test-takers' performance reflecting collocational competence.	- TLU domain and corpus analysis - Systematic collocation sampling - Rasch person-item variable map - Item response development - Expert review of the test - Rasch unidimensionality analysis - Rasch item strata estimation

2.10.2 Specifying the evaluation inference

Table 2.7 outlines specification of warrants, assumptions and potential baking for the evaluation inference As in Table 2.6, the evaluation inference has the warrant

that observed performance on the ACCT is evaluated to provide observed scores reflective of the collocational competence. This warrant rests on the assumptions that: 1) scoring procedure is appropriate to elicit responses that serve as evidence of various collocation competence levels, 2) test administration condition is conducive for test-takers to maximally demonstrate collocational competence, and 3) psychometric properties of the ACCT are appropriate for norm-referenced evaluation. These assumptions require potential backing from data preparation and screening, scoring and rubric development, Rasch dichotomous scaling, Rasch unidimensionality analysis, test trialling and evaluation, sufficient testing time, descriptive statistics analysis, and point-measure correlation estimation.

Table 2.7

Specification of warrants, underlying assumptions and potential baking for the evaluation inference

Warrant	Underlying assumptions	Potential backing
Observed performance on the ACCT is evaluated to provide observed scores reflective of the collocational competence.	1) The scoring procedure is appropriate to elicit responses that serve as evidence of various collocation competence levels.	- Data preparation and screening - Scoring and rubric development - Rasch dichotomous scaling - Rasch unidimensionality analysis
	2) Test administration condition is conducive for test-takers to maximally demonstrate collocational competence.	- Test trialling and evaluation - Sufficient testing time
	3) Psychometric properties of the ACCT are appropriate for norm-referenced evaluation.	- Descriptive statistics analysis - Point-measure correlation estimation - Rasch person-item variable map

2.10.3 Specifying the generalisation inference

As unlined in Table 2.8, the generalisation inference warrants that scores on the ACCT are estimates of expected scores which are congruent across items and invariant across gender. This warrant assumes that: 1) estimates of test-takers' performance can consistently distinguish among test-takers, 2) psychometric properties of the ACCT item are invariant across males and females who have equal collocational competence levels, 3) the test specification of the ACCT is adequately detailed and consistent to develop equivalent task or test forms, and 4) the paper-based administration of the test is sufficiently uniform to produce consistent results. Expected backing for these assumptions can be derived from Rasch internal consistency estimation, visual investigation of person-item variable and babble maps, Rasch differential test functioning analysis, Rasch differential item functioning analysis, test specification development, and test trialling, monitoring and instruction.

Table 2.8

Specification of warrants, underlying assumptions and potential baking for the generalisation inference

Warrant	Underlying assumptions	Potential backing
Observed scores on the ACCT are estimates of expected scores which are congruent across items and invariant across gender.	1) Estimates of test-takers' performance can consistently distinguish among test-takers.	- Rasch internal consistency estimation - Rasch person-item babble map investigation - Rasch person-item babble map investigation
	2) Psychometric properties of the ACCT item are invariant across males and females who have equal collocational competence levels.	- Rasch differential test functioning analysis - Rasch differential item functioning analysis
	3) The test specification of the ACCT is adequately detailed and consistent to develop equivalent task or test forms.	- Test specification development

Table 2.8

Specification of warrants, underlying assumptions and potential baking for the generalisation inference

Warrant	Underlying assumptions	Potential backing
	4) The paper-based administration of the test is sufficiently uniform to produce consistent results	- Task trialling and monitoring, and instruction.

2.10.4 Specifying the explanation inference

As in Table 2.9, the explanation inference warrants that expected scores are attributed to the construct of collocational competence in academic English writing. This warrant assumes that: 1) performance on the ACCT reflects test-takers' collocational competence, 2) the construct to be assessed is collocational competence which is defined as a restricted lexical collocation in academic written texts, 3) scores on the ACCT correlate positively to other tests of English language proficiency related to the construct, and 4) while doing the test, test-takers use cognitive process related to collocation use in academic language. These assumptions require potential backing from construct definition review, coring and rubric development, Rasch unidimensionality analysis, Rasch person-item babble map, Rasch person-item babble map investigation, collocation definition review, Rasch unidimensionality analysis, Rasch person-item variable map investigation, Rasch person-item babble map investigation, correlation analysis between ACCT scores and AVLT scores, correlation analysis between ACCT theta and AVLT theta, Rasch multiple-choice distractor analysis, and test reflection survey.

Table 2.9

Specification of warrants, underlying assumptions and potential baking for the explanation inference

Warrant	Underlying assumptions	Potential backing
Expected scores are attributed to the collocational competence construct in the academic written discourse.	1) Performance on the ACCT reflects test-takers' collocational competence.	- Interactionist construct definition review - Scoring and rubric development - Rasch unidimensionality analysis

Table 2.9

Specification of warrants, underlying assumptions and potential baking for the explanation inference

Warrant	Underlying assumptions	Potential backing
		- Rasch person-item babble map - Rasch person-item babble map investigation
	2) The construct to be assessed is collocational competence which is defined as a restricted lexical collocation in academic written texts.	- Phraseologist collocation definition review - Rasch unidimensionality analysis - Rasch person-item variable map investigation - Rasch person-item babble map investigation
	3) Scores on the ACCT correlate positively to other tests of English language proficiency related to the construct	- Correlation analysis between ACCT scores and AVLT scores - Correlation analysis between ACCT theta and AVLT theta
	4) While doing the test, test-takers use cognitive process related to collocation use in academic language	- Rasch multiple-choice distractor analysis - Test reflection survey

2.10.5 Specifying the extrapolation inference

Table 2.10 lays out specification of warrants, assumptions and potential baking for the extrapolation inference The extrapolation inference warrants that the construct of collocational competence as measured by the ACCT accounts for collocation production in the academic written discourse in universities or other institutions at tertiary level. This warrant assumes that: 1) collocations on the ACCT reflect those that the test-takers will be exposed to in the context of the academic

written discourse and 2) scores on the ACCT distinguish among proficiency groups with and without experience and topical knowledge of academic language. These assumptions require potential backing from TLU domain and corpus analysis, Rasch person-item variable map investigation, Rasch person strata estimation, Rasch person-item variable map investigation, and analysis of variance.

Table 2.10
Specification of warrants, underlying assumptions and potential baking for the extrapolation inference

Warrant	Underlying assumptions	Potential backing
The collocational competence construct as measured by the ACCT accounts for relevant language performance in the academic discourse in university or other higher-education settings.	1) Collocations on the ACCT reflect those that the test-takers will be exposed to in the context of the academic written discourse. 2) Scores on the ACCT distinguish among proficiency groups with and without experience and topical knowledge of academic language.	- TLU domain and corpus analysis - Rasch person-item variable map investigation - Rasch person strata estimation - Rasch person-item variable map investigation - Analysis of variance

2.10.6 Specifying the utilisation inference

Table 2.11 shows specification of warrants, assumptions and potential baking for the utilisation inference. The utilisation inference warrants that performance on the ACCT contributes to making appropriate norm-referenced decisions about placement in English language courses in universities or other institutions of higher education. This warrant rests on the assumptions that: 1) the interpretation of the ACCT scores provides enough information which contributes to the decision making process and 2) the ACCT scores are intended to be used to contribute to and facilitate student placement decision in appropriate English language courses in universities or other institutions of higher education. Expected backing for these assumptions can be derived from cut-score study, classification error analysis, and correlation study with English class grades.

Table 2.11

Specification of warrants, underlying assumptions and potential baking for the utilisation inference

Warrant	Underlying assumptions	Potential backing
Performance on the ACCT contributes to making appropriate norm-referenced decisions about placement in English language courses in universities or other institutions of higher education	1) The interpretation of the ACCT scores provides enough information which contributes to the decision making process	- Cut-score study - Classification error analysis
	2) The ACCT scores are intended to be used to contribute to and facilitate student placement decision in appropriate English language courses in universities or other institutions of higher education	- Cut-score study - Classification error analysis - Correlation study with English class grades

2.10.7 Specifying the consequence inference

Table 2.12 reveals specification of warrants, assumptions and potential baking for the utilisation inference. The consequence inference, modified from Voss (2012), has the warrant that the interpretation and use of the ACCT scores are appropriate and advantageous for all test users and stakeholders. This warrant rests on the assumptions that: 1) the construct of the ACCT raises awareness about the importance of collocations in academic English and 2) the construct of the ACCT raises awareness of introducing the importance of collocations in English instruction and material developments. These assumptions can be backed up by washback study and stakeholder survey.

It should be reminded that the utilisation and consequence inferences can be extensively studied upon utilisation of the ACCT. For this reason, the present study did not provide sufficient sources of evidence in support of the utilisation and consequence inferences at the time this study is being conducted. Score interpretation evaluation and predictive validation study for backing the utilisation and backing for the consequence inference were not examined in this study and should be further examined in future research.

Table 2.12

Specification of warrants, underlying assumptions and potential baking for the consequence inference

Warrant	Underlying assumptions	Potential backing
The interpretation and use of the ACCT scores are appropriate and advantageous for all test users and stakeholders.	1) The construct of the ACCT raises awareness about the importance of collocations in academic English.	- Washback study - Stakeholder survey
	2) The construct of the ACCT raises awareness of introducing the importance of collocations in English instruction and material developments	- Washback study - Stakeholder survey

2.11 Framework of the ACCT interpretive argument

As pointed out earlier, the interpretive argument specifies the proposed interpretation and use of test scores through laying out a network of inferences and assumptions that need to be theoretically and empirically backed up. As shown in Figure 2.4, the ACCT interpretive argument framework, modified from on Carol A Chapelle et al. (2008), consists of 7 interrelated inferences. Each inference has its warrant that states the proposed score interpretation or the proposed score use. The warrant then rests on assumptions requiring theoretical and empirical backing. Backing is resulted from either theoretical justification or empirical investigation and backing supporting one inference can also substantiate other inferences. The empirical evidence or backing is gathered through analyses of empirical data after a priori or theory is well documented from the review of related literature.

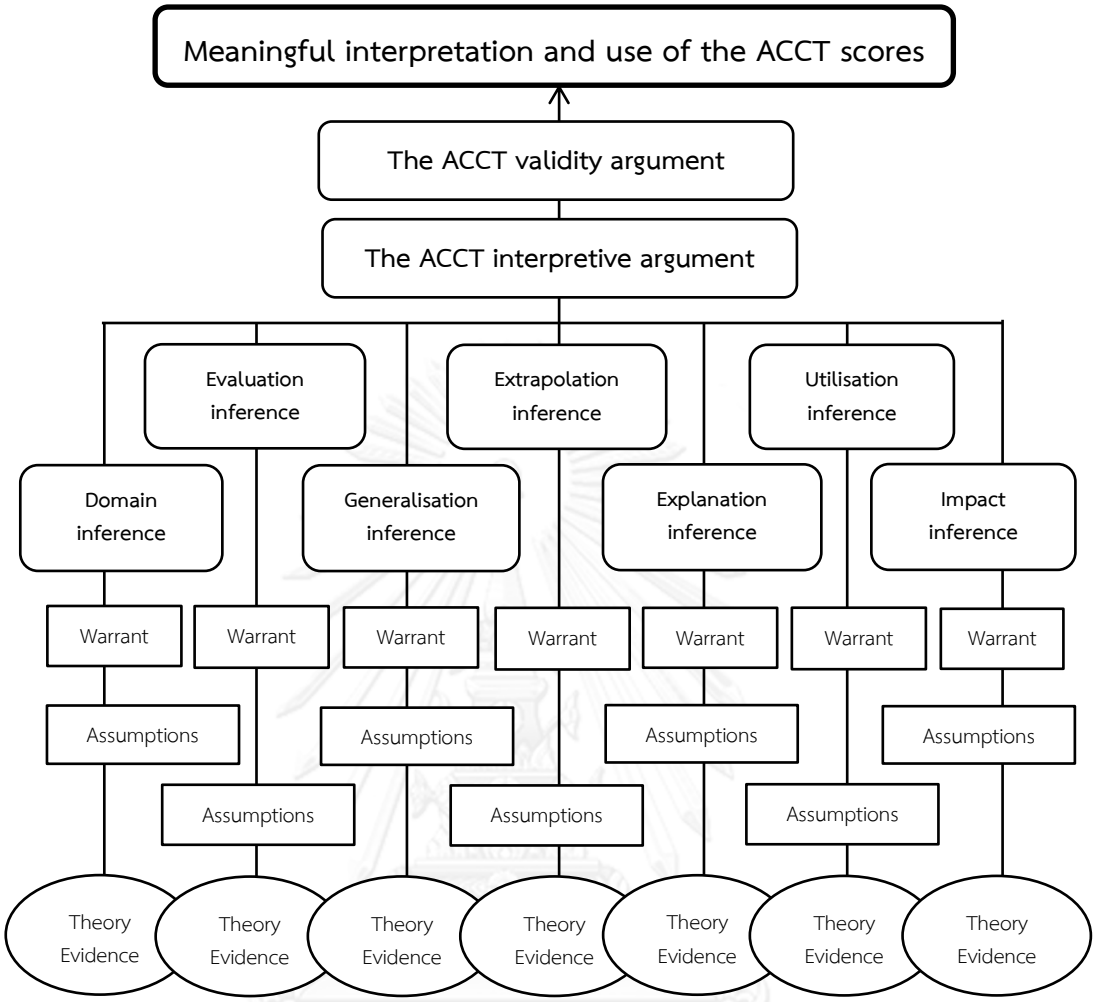

Figure 2.4. The ACCT interpretive argument framework (modified from Carol A Chapelle et al., 2008, p. 18)

2.12 Chapter summary

To recapitulate, several key issues and concepts have been thoroughly reviewed in this chapter and this theoretical review lays the foundation for the development and validation of the test. Up to this point, fundamental concepts were reviewed and the ACCT interpretive argument was specified, the ACCT was then developed and validated drawing upon these fundamental guidelines. In the next chapter, I present in detail the development of the ACCT and details of this not only draws on fundamental concepts and the ACCT interpretive argument presented in this chapter but also proves both theoretical and empirical backing in support of several assumptions.

CHAPTER 3
TEST DEVELOPMENT

Chapter 3 describes the process of developing the ACCT which is one of the focal objectives of the present study. The ACCT development is also based on previous findings, conceptual frameworks, and the ACCT interpretive argument structure in chapter 2. Details of the ACCT development documented in this chapter serve as both theoretical and empirical evidence in support of assumptions underpinning the warrants of domain description, evaluation, generalisation, and explanation inferences. This chapter begins by presenting the purposes, context, and TLU domain of the test. Then the process of selecting TLU corpus of academic written English is described, followed by a presentation of the construction of academic written sub-corpora. Following this, how TLU verb-noun collocations were sampled and how item responses were developed are delineated. The two penultimate sections that follow are concerned with test evaluation and revision as well as test trialling and quality evaluation. As always, this chapter ends with chapter summary

3.1 Defining test purposes, context, and TLU domain

The purposes of the ACCT were to provide meaningful scores which can be interpreted as reflecting collocationnal competence and used as a norm-referenced test for placement or screening decision. The ACCT was developed to assess collocationnal competence and facilitate placement or screening decision. In terms of testing context, test-takers are EFL graduate students with different proficiency levels and the setting is university or higher-education setting. The TLU domain of interest is the academic written English. To sum up, the ACCT was developed to provide scores which can be meaningfully interpreted as reflecting EFL graduate students' receptive collocational competence and appropriately used as a norm-referenced evaluation for facilitating placement or screening decision in university or higher-education setting.

3.2 Selecting TLU corpus of academic written English

With regard to TLU corpus selection, the British National Corpus (BNC) was chosen to represent the TLU domain of academic written English because it contains

a wealth of textual data and information about the frequency and distribution of words and phrases in many different registers of English. BNC is a carefully-gleaned collection of 4,124 contemporary written and spoken English texts, primarily from the United Kingdom. The corpus contains texts of over 100 million words and covers a representative range of domains, genres and registers. Therefore, BNC contains a large number of texts through which the TLU domain of academic written English can be representatively and relevantly investigated. Only the academic written discourse in BNC was used for collocation sampling and it consists of practically 16 million running words and is the largest collection of written academic texts at the time when this ACCT was developed.

The Lancaster BNCweb was used as an online tool for corpus-based TLU domain analysis and collocation sampling in this study. As shown in Figure 3.1, seven academic domains in the academic written discourse in BNC were located to represent academic written texts from seven main academic disciplines. Several academic written domains in BNC were located in this study in an attempt to ensure that English verb-noun collocations were representative and commonly found in various academic written English.

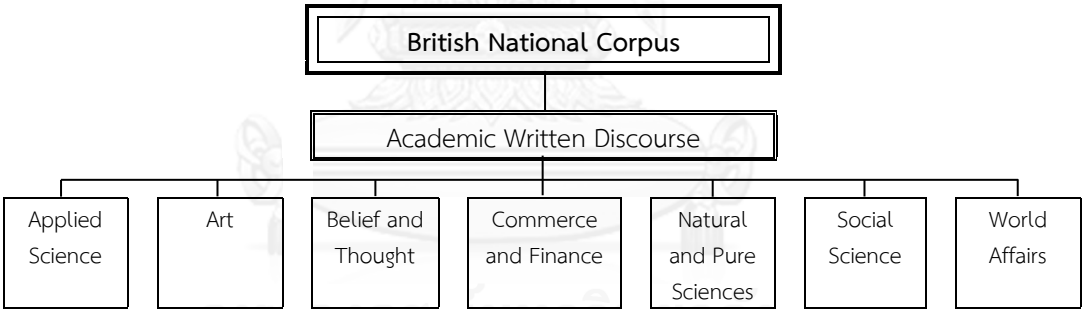

Figure 3.1. Diagram showing the construction of sub-corpora structure in BNC

3.3 Constructing academic written sub-corpora

After seven academic domains in the academic written discourse in BNC were identified, an academic written sub-corpus and seven academic written sub-corpora were constructed to represent academic written English used in seven main academic disciplines. Firstly, an academic written sub-corpus was built from the academic written domain in BNC and secondly seven academic written sub-corpora were further created from seven academic domains in an academic written sub-corpus. There are, therefore, seven academic written sub-corpora embedded in one bigger academic written sub-corpus. These academic written sub-corpora were

intended to represent the TLU domain of academic written English used in various academic disciplines. Following this, high-frequency verb-noun collocations were systematically sampled from seven academic written sub-corpora in the process of target collocation identification which is described in the next section. The characteristics of academic written sub-corpora created in BNC are summarised in Table 3.1.

Table 3.1
Characteristics of academic written sub-corpora in BNC

Name of subcorpora	No. of text files	No. of words
1. Applied Science sub-corpus	37 texts	1,742,312
2. Art sub-corpus	58 texts	1,297,379
3. Belief & Thought sub-corpus	16 texts	614,981
4. Commerce & Finance sub-corpus	15 texts	463,786
5. Natural & Pure Sciences sub-corpus	60 texts	1,754,916
6. Social Science sub-corpus	239 texts	7,194,435
7. World Affairs sub-corpus	72 texts	2,710,219
Academic written sub-corpus	497 texts	15,778,028

3.4 Sampling TLU verb-noun collocations

In terms of TLU verb-noun collocations sampling, different scholars have different approaches to identifying lexical or phraseological units in corpora. The present study adopted a frequent word-based approach where TLU verb-noun collocations in seven academic written sub-corpora in BNC were systematically extracted based on both frequency statistics and human judgement in the academic written sub-corpus of BNC. A frequent word-based approach has been widely adopted by previous studies as it provides many advantages for second language studies. Figure 3.2 shows the TLU verb-noun collocations sampling procedure from academic written sub-corpora in BNC. The procedure of TLU verb-noun collocations sampling was done through the Lancaster BNCweb search engine (see Figure 3.3). A frequent word-based approach began by setting the criteria in order to locate high-frequency words based on the criteria set.

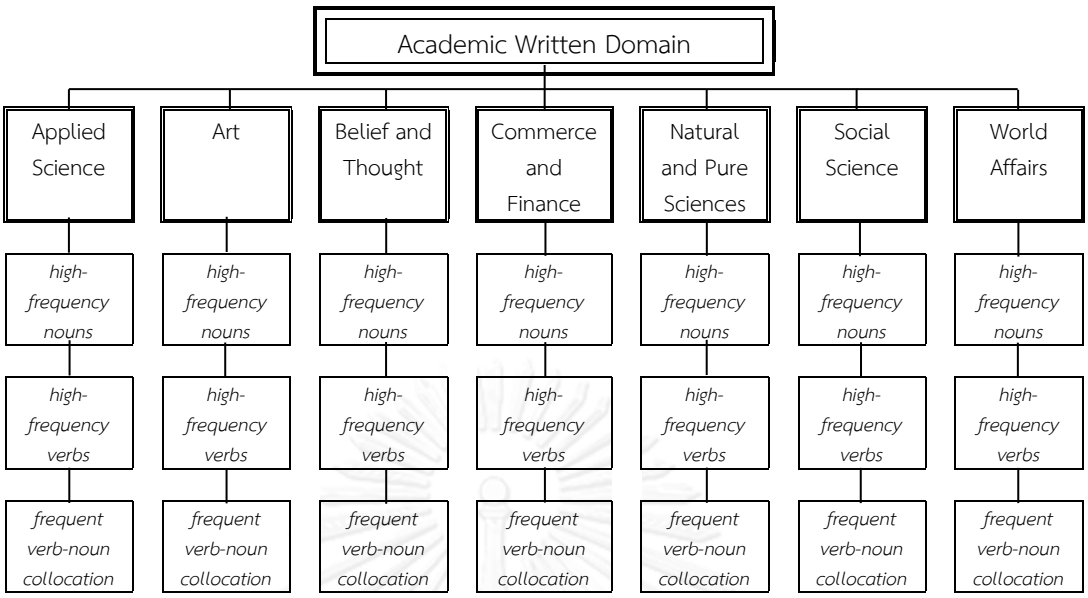

Figure 3.2. Diagram of corpus sampling procedure

Collocation parameters:			
Information:	collocations	Statistics:	Log-likelihood
Collocation window span:	5 Left - 5 Right	Basis:	subcorpus
Freq(node, collocate) at least:	1	Freq(collocate) at least:	5
Filter results by:	Specific collocate:	and/or tag: any verb	Submit changed parameters Go!

Figure 3.3. Screenshot of search engine in the Lancaster BNCweb service

3.4.1 Locating high-frequency nouns

It is important to note that the targeted nouns that form verb-noun collocations function as the object of the verb in sentences and were selected based on the high frequency criterion using Log-likelihood statistics. First of all, high-frequency nouns in seven academic written sub-corpora were searched and arranged in order of statistical significance using Log-likelihood statistics. Only most eight high-frequency nouns that were not proper nouns, functioned as objects, and appropriate for identifying their collocate verbs were selected.

By doing so, 56 high-frequency nouns found in seven academic written sub-corpora were selected and included in the final list of high-frequency nouns which were later used for further identifying their high-frequency collocate verbs in seven academic written sub-corpora. Take for example, in the sub-corpus of Applied Sciences, the word "*study*" was selected as one of eight high-frequency nouns found

in the sub-corpus and its raw frequency of occurrence is 4,201. The noun "*study*" was then searched for its high-frequency verbs in the sub-corpus of Applied Sciences.

3.4.2 Locating high-frequency verbs

Having been identified a list of 56 high-frequency nouns found in seven academic written sub-corpora based on frequency using on Log-likelihood statistics, 56 high-frequency nouns were searched for their high-frequency collocating verbs. It should be noted that high-frequency verbs pairs of high-frequency verb-noun combination must be transitive verbs that require objective nouns. Each selected high-frequency noun was then searched for its high-frequency collocate verb in the sub-corpus where such noun was previously identified.

So far several criteria based on human judgement and statistics were used to select high-frequency nouns and their high-frequency collocate verbs that frequently co-occurred in seven academic written sub-corpora. Like inclusion of high-frequency nouns, high-frequency collocate verbs were selected on the basis of Log-likelihood statistics, the collocation window span of 5 nodes, and the frequency of at least 5 occurrences. The window span of 5 nodes means the five-word distance before and after a high-frequency noun where a high-frequency collocate verb may appear in collocation with a high-frequency noun. Most high-frequency collocate verbs were arranged in order of statistical significance using Log-likelihood statistics.

In the sub-corpus of Applied Sciences, for example, the verb "approve' was found to most frequently co-occur with the high-frequency noun "study" and therefore "*approve"* and "*study"* were regarded as one pair of high-frequency verb-noun collocation frequently found in academic written English, especially in the field of Applied Sciences. Of all 56 verb-noun collocations identified, six pairs of collocations were excluded as they were the same verb-noun colocations. Therefore, only 50 pairs of verb-noun collocations were used to develop ACCT items. Table 3.2 shows the initial distribution of an "*approve-study*" collocation in the sub-corpus of Applied Sciences without taking into account its variations or varied forms.

Table 3.2

Initial distribution of approve-study combination in the sub-corpus of Applied Sciences

Collocation	Total No. in current subcorpus	Expected collocate frequency	Observed collocate frequency	In No. of texts	Log-likelihood value
approved study	84	0.941	52	6	356.364

3.4.3 Identifying variation of verb-noun collocations

After high-frequency nouns and collocate verbs were obtained and combined to form 50 pairs of high-frequency verb-noun collocations, a query-syntax was used to search the total raw frequency of any variations of each high-frequency verb-noun collocation in seven academic written sub-corpora where each high-frequency verb-noun collocation was found. The syntax-based searching was used to identify any variations of a particular verb-noun collocation. By using the syntax-based searching, the total raw frequency of each particular collocation can be computed. For example, in the case of the "*approve-study*" collocation, a syntax was used to search for any variations of the "*approve-study*" item in the academic written sub-corpus of Applied Sciences. The variation distribution of the "*approve-study*" collocation is presented in Table 3.3.

Table 3.3

Distribution of approve-study collocation variations in the sub-corpus of Applied Sciences

Word	Expected collocate frequency	Observed collocate frequency	Dispersion over text files	Log-likelihood value	Frequency per million words
approved study	0.654	52	7/37	359.0024	32.72
approved studies	0.410	7	5/37	26.662	

To sum up, of all 56 high-frequency verb-noun collocations identified, six of them were excluded as they were the same collocations, resulting in a total of 50 high-frequency verb-noun collocations commonly found in seven academic areas of academic written English. The syntax-based searching was used to identify the total exact raw frequency of each high-frequency collocation, including variations or forms,

so as to ensure the high-frequency of 50 final verb-noun collocations. In the next step, concordance lines (sentences) of each identified verb-noun collocation were searched and proper sentences were initially manually selected for composing test items on the ACCT. All these are described in detail in the process of item response development.

3.5 Item response development

3.5.1 Item format selection

With regard to item response development, the current collocation test was developed based upon test specification as presented in Appendix A and conceptual framework as reviewed and documented in chapter 2. A five-option multiple-choice format was chosen for the current collocation test because it is appropriate to measure a receptive aspect of collocational competence. A multiple-choice format is a receptive or selective response item in that test-takers choose from a set of responses in lieu of producing a response and therefore elicit test-takers' recognition knowledge which is a receptive aspect of lexical competency. As such, previous research used a multiple-choice format to elicit receptive dimension of vocabulary and collocational knowledge (Gyllstad, 2005, 2007; Jaen, 2007; Keshavarz & Salimi, 2007; Webb, Newton & Chang, 2013; Webb & Kagomoto, 2009).

3.5.2 Text input selection and evaluation

Sentences for composing five-option multiple-choice items were chosen from the concordance lines in each of seven written academic sub-corpora where each identified high-frequency verb-noun collocation was found in order to provide sufficient context. Details of collocation identification were already described in the collocation sampling framework. Sentences were selected only from the concordance lines that were appropriate in length and contained sufficient context. Upon manual examination and evaluation of text complexity and communality, 50 sentences were found appropriate for composing multiple choice items.

All 50 sentences include passive and active declarative sentences and ranged from 11 words to 28 words. The majority (86.6%) of the total 725 running words appeared in Longman 3,000 keywords. It can thus be claimed that collocations and language use in the present collocation test are common enough to facilitate test-takers in doing the test. Moreover, Flesch-Kincaid readability indices indicted that the

text inputs were not too complex to distract test-takers from demonstrating collocational competence. It is also worth pointing out that some words in pairs of academic collocations on the ACCT do not appear in the Academic World List, developed by Coxhead (2000) as the criteria for extracting academic collocations and academic words are different. Having said that, however, sampling high-frequency collocations from the academic written domain in BNC can ensure that collocations on the ACCT are maximally representative of those appearing in the academic settings. Text input evaluation results are presented in Appendix B.

3.5.3 Multiple-choice item construction

After 50 appropriate sentences were identified based on manual examination (researcher judgement), the readability index, and Longman 3,000 keywords were used to compose the test, they were then used to develop ACCT item stems or questions. After that ACCT item options or alternatives were developed, correct options were then developed using verbs that collocate with nouns in 50 pairs of high-frequency verb-noun collocations and distractors were developed based on the researcher's experience and literature review.

Figure 3.4 shows a single multiple-choice item used in the current ACCT. It consists of three aspects. One aspect is a step, which is a problem in the form of an incomplete sentence. Another aspect is one best correct choice, which is an intended answer to the problem. The other is four incorrect choices, which are distractors to the best correct choice. The following example illustrates the five-option multiple-choice item format of the ACCT with the verb replaced by a gap in the stem or sentence. In the example of a multiple-choice item 21, test-takers are asked to select the best answer that completes the sentence with the most appropriate meaning for academic written context.

21)	In 1986, as part of its wider proposals for the reform of local government finance, the government declared its intention to _____ a new grant system. *(a stem or question)*				
	a. renovate *(a distractor)*	b. integrate *(a distractor)*	c. introduce *(a correct choice)*	d. invent *(a distractor)*	e. install *(a distractor)*

Figure 3.4. Components of a multiple-choice item format of ACCT items

In sections that follow, I present details of test evaluation and revision as well as test trialling and quality evaluation. After composing the initial ACCT which consists of 50 multiple choice items, the next step was to evaluate the ACCT by three experts. Two experts are non-native teachers of English who are specialised in language testing and assessment and the other one is a native speaker of English. After experts evaluated the ACCT, then the ACCT was piloted, analysed, and revised before it was administered to actual participants in the actual testing situation.

3.6 Test evaluation and revision

The initial 50 ACCT items were evaluated by three experts in order to see whether the stem or question presents a single clear sentence and a sufficient context for a verb-noun collocation, whether the correct answer is clearly the best among plausible incorrect alternatives, and whether the incorrect alternatives are overall plausible enough to distract low-proficiency examinees away from the correct answer. Two experts were non-native teachers of English who are specialised in language assessment and the other expert was a native speaker of English. Fifty items were then revised based on expert evaluation and suggestions. The initial 50-item ACCT evaluated by three experts and revised according to expert suggestions. The test evaluation form is presented in Appendix H and the expert item evaluation result is presented in Appendix D.

3.7 Test trialling and quality evaluation

Fifty ACCT items were administered to a sample of 30 EFL graduate students in Chulalongkorn University who had similar characteristics but, of course, were not part of the actual participants in the real testing. Thirty EFL graduate students were classified into low, mid and high-proficiency levels based on Chulalongkorn University Test of English Proficiency (CU-TEP), TOEFL iBT, and IELTS scores they used to apply for the university. Responses from 50 multiple-choice Items were analysed based on CTT using TAP software (version 12.9.23). Items that had difficulty index roughly within the range of between 0.2 and 0.9 and discrimination index at least 0.2 were included to compose the final ACCT. The final piloted version of the ACCT consisted of 30 items with difficulty range between 0.2 and 0.9, discrimination range of at least 2.0, and a coefficient alpha test internal consistency of 0.85. Table 3.4 shows item statistics of 30 selected items from pilot study.

Table 3.4

Item statistics of 30 selected items from pilot study

New Item	Verb-noun collocation	Academic domain	No. of correct	*p*	r_{pb}
1	find a way	Social Sciences	19	0.63	0.42
2	cite a case	Social Sciences	17	0.57	0.45
3	leave school	Social Sciences	27	0.90	0.19
4	enforce a law	Social Sciences	24	0.80	0.22
5	make an award	Social Sciences	15	0.50	0.30
6	cover an area	Social Sciences	19	0.63	0.39
7	see figure	Natural Sciences	21	0.70	0.70
8	obtain a result	Natural Sciences	17	0.57	0.53
9	provide an example	Natural Sciences	22	0.73	0.47
10	improve health	Natural Sciences	21	0.70	0.36
11	conduct a study	Natural Sciences	22	0.73	0.50
12	have an idea	Belief and Thought	13	0.43	0.42
13	make sense	Belief and Thought	27	0.90	0.21
14	justify beliefs	Belief and Thought	24	0.80	0.20
15	hold the view	Belief and Thought	12	0.40	0.76
16	account for the fact	Belief and Thought	11	0.37	0.49
17	play a part	World Affaire	9	0.30	0.41
18	pursue a policy	World Affaire	8	0.27	0.58
19	fight the war	World Affaire	11	0.37	0.30
20	exercise power	World Affaire	8	0.27	0.53
21	introduce a system	World Affaire	11	0.37	0.66
22	apply a rule	Commerce/Finance	20	0.67	0.69
23	carry on a business	Commerce/Finance	13	0.43	0.19
24	appoint an expert	Commerce/Finance	21	0.70	0.28
25	terminate a contract	Commerce/Finance	18	0.60	0.52
26	use a word	Arts	27	0.90	0.39
27	do work	Arts	18	0.60	0.31
28	read text	Arts	6	0.20	0.22
29	have a disease	Applied Sciences	12	0.40	0.40
30	treat a group	Applied Sciences	17	0.57	0.39

* Note: *p* = difficulty, r_{pb} = point-biserial correlation

In addition to analysis of high-frequency collocations in BNC, all texts and collocations used on the 30-item ACCT were once again evaluated based on the New General Service List (NGSL), which is developed after the original General Service List (OGSL) and is claimed as the most important words for second language learners of English. The purpose of this comparative analysis was to ascertain to what extent texts and collocations on the ACCT are found in NGSL. If the percentage is high, it can then be confident that texts and individual words in pairs of collocations on the ACCT are frequent enough and thus are very likely to be encountered by EFL learners. Table 3.5 shows results of evaluation of text inputs on the ACCT. Based on text analysis using the Online Graded Text Editor, there were 582 running words on the ACCT which accounted for 97.42% in NGSL and 84.02% in OGSL. Of all 30 targeted verb-noun collocations, as much as 98.81% of 30 targeted verb-noun collocations were found in NGSL and 80.95% were found in OGSL. All these confirmed that text and collocation inputs used in the ACCT were highly frequent.

Table 3.5

Summary of text evaluation of final 30 ACCT item questions

	Sentences	Word count	% in GLS	
			NGSL	OGSL
1)	This means that lecturers and tutors will have to find ways of connecting with their students' outlooks.	16	87.50	93.75
2)	The following additional cases were cited in argument in the Court of Appeal.	13	100	84.62
3)	After leaving school, most of his friends moved away to university.	11	100	100
4)	It was not concerned with the position of local authorities which have the function of enforcing the law in their districts in the public interest.	25	96	88
5)	Meanwhile an award of £1 was made to full-time workers; part-time workers got nothing.	15	86.67	93.33
6)	These will cover areas such as equal opportunities, multi-cultural education, cross-curricular themes, competences and dimensions and special needs.	20	90	75
7)	The subject has been reviewed (White et al, 1981) and will be briefly described here (see Figure 5.1).	16	100	87.50
8)	We do not have space for a full description of all the experimental techniques used in obtaining the results discussed in this book.	23	95.65	86.96
9)	The history of theories of electricity provides an example of the changing fortunes of rival research programmes.	17	100	76.47

Table 3.5

Summary of text evaluation of final 30 ACCT item questions

Sentences	Word count	% in GLS	
		NGSL	OGSL
10) Prescribing is one possible treatment option; others include counselling, educating patients on self-limiting illnesses, and changes in lifestyle to improve health.	22	95.45	72.27
11) We conducted a two-year study to assess the effectiveness of the family smoking education and my projects in influencing smoking behaviour.	22	95.45	68.18
12) There is no general nature in common to those things, and any idea we have is never general or abstract, but always of some particular thing.	26	100	88.46
13) It is by virtue of such rules that we can make sense of the idea that we are objectively correct to call the new sensation a pain.	27	92.59	85.19
14) All agree that some of our beliefs are justified by their relation to other beliefs.	15	100	93.33
15) And philosophers talk of ‘sensations’ in this connection because of views they hold about perception.	14	92.86	85.71
16) A similar mechanism may perhaps account for the fact that some group-living animals drive sick or injured individuals out of the group.	23	100	78.26
17) It is evident that the larger and more popular temples may have played a considerable part in the economy of any province.	22	95.45	81.82
18) Those groups have brought pressure to bear on government to provide resources or pursue policies to the benefit of their members.	21	100	76.19
19) That unemployment fell as a result of war is an undeniable fact, but it was not the primary reason for the decision to fight the war.	26	100	88.46
20) Research would inevitably concentrate on informal relations and social structures through which power is exercised.	15	86.67	73.33
21) In 1986, as part of its wider proposals for the reform of local government finance, the government declared its intention to introduce a new grant system.	25	100	84.00
22) It has been said that these rules will be applied less stringently to a commercial contract than to other types of document.	22	100	86.39
23) Their power to admit and expel members has the important consequence of granting and revoking authority to carry on investment business.	21	95.24	76.19
24) If the parties agree on a procedure and the expert does not, the parties should appoint another expert.	18	100	83.33

Table 3.5

Summary of text evaluation of final 30 ACCT item questions

Sentences	Word count	% in GLS	
		NGSL	OGSL
25) Under a contract of sale, breach of condition by the seller allows the buyer to reject the goods and terminate the contract.	22	95.45	77.27
26) Here are some words which are commonly used in essay.	10	90	100
27) Very little work has been done in accounting for the development of an individual dramatic character in pragmatic or discourse terms.	21	90.48	90.48
28) Many of these texts can be read as elaborate commentaries on the nature of writing and reading.	17	88.24	88.24
29) Twenty three children had more severe but intermittent symptoms and nine had severe disease throughout the year.	17	100	88.24
30) The control group was treated with an oral triple therapy regimen which had previously been evaluated in a pilot study.	20	95	85
30 sentences in total	582	97.42	84.02
30 pairs of collocation in total	84	98.81	80.95

3.8 Chapter summary

Up to this point, details of the ACCT development have been thoroughly discussed. What has been documented in this chapter provide some theoretical and empirical evidence in support of assumptions related to domain inference, evaluation inference, generalisation inference, and explanation inference outlined in the ACCT interpretive argument. I finish off this chapter with a brief summary of the ACCT development process in Table 3.6. In the next chapter, I present the methodology used in the present study and much empirical evidence is introduced in the following chapter.

Table 3.6

Summary of process of the ACCT development

Stage	Procedure	Description	Outcome
1. Defining test purposes, context, and TLU domain	Defining the purposes of the ACCT	The purposes of the ACCT are to provide scores which can be interpreted as reflecting collocationnal competence and used as a norm-referenced test for placement or screening decision.	Clearly-defined ACCT purposes for collocationnal competence assessment and placement or screening decision-making
	Identifying test-takers' characteristics and testing setting	Test-takers are EFL graduate students with different proficiency levels and the setting is university or higher-education setting.	Clearly-identified testing context including EFL graduate students and university or higher-education setting
	Specifying the TLU domain of interest	The TLU domain of interest is the academic written English.	The TLU domain of interest is academic written English
2. Selecting TLU corpus of academic written English	Selecting a corpus representative of the TLU domain of academic written English	The British National Corpus (BNC) was chosen as it contains a wealth of textual data and information about the frequency and distribution of words and phrases in many different registers of English.	BNC representing the TLU domain of academic written English
3. Constructing academic written sub-corpora	Building a new academic written sub-corpus through the Lancaster BNCweb service	An academic written domain in BNC was located to create a new sub-corpus representing the academic written discourse.	The new academic written sub-corpus
	Building new seven academic written sub-corpora	Seven academic areas in an academic written sub-corpus were located to create seven academic written sub-corpora representing texts from varying academic disciplines	New seven academic written sub-corpora representing seven academic disciplines

Table 3.6

Summary of process of the ACCT development

Stage	Procedure	Description	Outcome
4. Sampling TLU verb-noun collocations	Identifying high-frequency nouns to create a list of key nouns based on Log-likelihood statistics	Only eight high-frequency and appropriate nouns in each of seven academic written sub-corpora were included for further identifying their collocate verbs.	A list of 56 high-frequency nouns from seven academic written sub-corpora
	Identifying high-frequency collocate verbs of 50 frequent verb-noun collocations based on Log-likelihood statistics	Each of 56 high-frequency nouns was searched for its high-frequency collocate verb. Only verb that requires an objective noun was selected to form a pair of high-frequency verb-noun collocation. Six pairs of collocations were excluded as they are the same verb-noun colocations.	A list of 50 high-frequency verb-noun collocations frequently found in seven academic sub-corpora representing academic written English
	Identifying variations of high-frequency verb-noun collocations in seven academic written sub-corpora	Each verb-noun collocation was searched using a query-syntax in the Lancaster BNCweb search-engine to identify any variations or forms of a particular verb-noun collocation.	A list of the total raw frequency of variations or various forms of each verb-noun collocation found in seven academic written sub-corpora
5. Item response development	Selection item response format	A five-option multiple-choice format was chosen for the current collocation test as it is appropriate to measure a receptive aspect of collocational competence.	A multiple-choice format for the test
	Selecting sentences containing targeted verb-noun collocations to form	Concordance lines containing sentences where each selected verb-noun collocation appeared were	A list of manually selected sentences containing 50 pairs of high-frequency verb-

Table 3.6

Summary of process of the ACCT development

Stage	Procedure	Description	Outcome
	stems or questions	created and then only clear sentences were manually selected to form item questions.	noun collocations
	Evaluating the commonality and complexity of texts by readability index and by comparing with Longman 3,000 keywords	Manually selected sentences were evaluated by comparing with Longman 3,000 keywords and using Flesch-Kincaid method of readability index, including the Flesch-Kincaid Grade level and the Flesch-Kincaid Reading Ease score of over	A list of proper sentences screened by readability index and Longman 3,000 keywords
	Developing ACCT item questions	Sentences, evaluated based on both the readability index and Longman 3,000 keywords were used to compose the test.	The initial 50-item ACCT
	Developing ACCT item options	Correct options were developed using verbs that collocate with nouns in 50 pairs of high-frequency verb-noun collocations and distractors were developed based on the researcher's experience.	
6. Test evaluation and revision	Evaluating the initial ACCT by three experts	Fifty ACCT items were reviewed by three experts. Two experts were non-native teachers of English who are specialised in language assessment and the other expert was a native speaker of English. Fifty items were then revised based on expert evaluation and suggestions.	The initial 50-item ACCT evaluated by three experts and revised according to expert suggestions

Table 3.6

Summary of process of the ACCT development

Stage	Procedure	Description	Outcome
7. Test trialling and quality evaluation	Piloting 50 ACCT items	Fifty ACCT items were administered to 30 graduate students with low, moderate and high English proficiency. Items were then analysed using TAP item analysis software. Items that had difficulty index of between 0.2 and 0.9 and discrimination index at least 0.2 were included to compose the final ACCT.	The final piloted version of 30-item ACCT with difficulty range between 0.2 and 0.9, discrimination range of at least 2.0, and a coefficient alpha test internal consistency of 0.85

CHAPTER 4
METHODOLOGY

In this chapter, I describe the methodology employed in the present study. The methodology is concerned with the analysis of empirical data in order to substantiate the assumptions underlying the warrants of the inferences in the interpretative argument for the ACCT. Chapter 4 begins with the presentation of the demographic characteristics of participants. Following this, research instruments, data collection procedure, and data analysis procedure are presented. This chapter ends with the chapter summary.

4.1 Participants

The participants were 193 EFL graduate students, purposively sampled to represent EFL graduate students with low, mid, and high levels of English proficiency and from a variety of academic fields at Chulalongkorn University, Thailand. A sample size of over 100 test-takers is proven to generate stable parameter estimates for the Rasch model analysis (W. H. Chen et al., 2014; Linacre, 1994). An initial number of participants were 199 but six of them were excluded as they did not provide complete data and information needed for analysis in this study. Of the six excluded students, one did not complete too many ACCT items and the others did not report any standardised English test scores. As such, the data were obtained solely from 193 EFL graduate students.

The participants were grouped into low, mid, and high levels of English proficiency based on Chulalongkorn University Test of English Proficiency (CU-TEP), TOEFL iBT, and IELTS scores they used to apply for the university. Of all 193 students, 84 (43.5%) were classified as beginner EFL learners, 59 (30.65%) were classified as intermediate EFL learners, and 50 (25.9%) were classified as advanced EFL learners. The criterion for classifying English proficiency levels is presented in Table 4.1 and demographic characteristics of EFL graduate students are presented in Table 4.2

Table 4.1

Criterion for classifying English proficiency levels

Proficiency Level	CU-TEP	TOEFL iBT	IELTS
Low-proficiency	0 - 449	0 - 44	0.0 - 4.5
Mid-proficiency	450-579	45 - 91	5.0 - 6.0
High-proficiency	580-677	92 - 120	6.5 - 9.0

Table 4.2

Demographic characteristics of 193 EFL graduate students

Demographic Characteristics	Proficiency level							
	Low		Mid		High		Total	
	n	*%*	*n*	*%*	*n*	*%*	*n*	*%*
Gender								
Male	32	49.2	22	33.8	11	16.9	65	33.7
Female	52	40.6	37	28.9	39	30.5	128	66.3
Study level								
Master	84	49.1	49	28.7	38	22.2	171	88.6
Doctor	0	0.0	10	45.5	12	54.5	22	11.4
Native language								
Thai	82	46.3	52	29.4	43	24.3	177	91.7
Chinese	1	12.5	3	37.5	4	50.0	8	4.1
Vietnamese	0	0.0	2	50.0	2	50.0	4	2.1
Lao	1	50.0	1	50.0	0	0.0	2	1.0
Hindi	0	0.0	0	0.0	1	100	1	0.5
Cambodian	0	0.0	1	100.0	0	0.0	1	0.5
Total	84	43.5	59	30.6	50	25.9	193	100

4.2 Instruments

4.2.1 Academic Collocational Competence Test

The ACCT (see Appendix C) is a paper-delivered multiple-choice test and was developed by the authors to measure the ability to recognise verb-noun collocations used in academic written English. It was developed based on high-frequency verb-noun collocations from BNC. Colocations and test inputs were extracted through the Lancaster BNCweb Server. A five-option multiple-choice item format was chosen for

the current collocation test as it is appropriate to measure a receptive collocational competence. The final version of the ACCT consists of 30 items and the time allowed for testing is 30 minutes.

Figure 4.1 shows an example of the ACCT Item 21. Test questions are incomplete sentences. Beneath each sentence, there are five verbs, marked a, b, c, d, and e. Examinees had to choose one verb that best collocates with a noun in the sentence with the most appropriate meaning for the academic context. Item stems were chosen from the concordance lines in seven written academic domains where each targeted collocation was found. Only sentences appropriate in length and sufficient in context were selected to form item questions. Table 3 summarises the process of the ACCT development.

21) In 1986, as part of its wider proposals for the reform of local government finance, the government declared its intention to _____ a new grant **system.**

a. renovate b. integrate c. introduce d. invent e. install

Figure 4.1. Example of an ACCT Item 21

4.2.2 Academic Vocabulary Level Test

The Academic Vocabulary Level Test (AVLT), developed by Schmitt et al. (2001), was used as a measure of vocabulary size knowledge. Only the academic section of the Vocabulary Levels Test Version 2 was used in this study. The academic section of the Vocabulary Levels Test consists of 30 items in 10 clusters (see Appendix C). It is based on Academic Word List (Coxhead, 2000). The Vocabulary Levels Test was designed as a discrete, selective vocabulary test with the words presented in isolation. As shown in Figure 4.2, the items are presented in groups of three words, together with six possible definitions. Examinees select definitions of the six words on the left that match each of the target three words on the right. The Vocabulary Levels Test was designed to provide an estimate of vocabulary size for second language (L2) learners of academic English. The rationale behind the test derives from studies which have shown that vocabulary size is directly related to the ability to use English in a variety of ways.

1 area	
2 contract	_____ written agreement
3 definition	_____ way of doing something
4 evidence	_____ reason for believing something is or is not true
5 method	
6 role	

Figure 4.2. Example of an AVLT bundle

4.2.3 Test reflection questionnaire

The test reflection survey (see Appendix C) is adopted and translated from Voss (2012) and is included at the end of the ACCT. The test reflection survey aims to elicit responses from the test-takers in relation to their awareness of engaging in the use of academic language during test administration and the relationship between academic language on the test and that in textbooks used at the university. This test reflection questionnaire includes three questions in English with equivalent Thai translations used to gather both quantitative and qualitative data. The questionnaire was translated into Thai by the researcher. The first two questions on the questionnaire are multiple-choice items with three options: “Yes”, “No”, or “I don‘t know”. The third question is an open-ended response format. The test reflection questions are: 1) were you thinking about academic English as you took the test? 2) do you think the English in this test is similar to academic English used in university textbooks?, and 3) explain how the English in this test is similar to or different from English used in university textbooks.

4.3 Data collection procedure

The ACCT was administered to students from December 2013 to January 2014, together with the 30-item Academic Vocabulary Levels Test Version 2 (Schmitt et al., 2001) and the 3-item test reflection questionnaire. Time allowed for the tests was 60 minutes. The tests were counterbalanced and administered during certain class periods. I asked for approval from teachers responsible for the classes and asked for cooperation from volunteer students. I delivered the ACCT, explained the test instruction both in Thai and in English, and monitored students.

4.4 Data analysis procedure

Analysis of empirical data was carried out to provide evidence in response to research questions addressed earlier. Empirical data analysis involved descriptive statistics, Rasch measurement analysis, correlation analysis, one-way analysis of variance, cut-score analysis, classification error analysis, and test reflection survey analysis. In this section, I describe the equipment and software employed in data analysis, the data preparation, and several analytical methods used in the present study.

4.4.1 Equipment and software

For the quantitative analysis, data from the ACCT scores, Rasch competency estimates, test reflection survey were processed in Microsoft Excel 2010 and IBM SPSS statistics version 22 for PC. Descriptive statistics, correlation analysis, analysis of variance, and test reflection data analysis were performed through IBM SPSS statistics. Winsteps software 3.81.0 was used to perform several applications of the Rasch measurement analysis and WinBUGS software 1.4.3 was used to estimate person competency estimates or theta for cut score setting and classification error estimation. With respect to qualitative analysis, the test reflection responses were transcribed and coded using Microsoft Excel 2010. Table 4.3 summarises analytical methods, data sources, and software used to analyse empirical data in the present study.

Table 4.3
Summary of analytical methods, data sources, and software

Analytical methods	Data sources	Programme/software	Inference
1) Data preparation and screening	- ACCT scores - AVLT scores - Test reflection data	- IBM SPSS Statistics version 22 - Microsoft Excel 2010	Evaluation
2) Descriptive statistics	- ACCT scores	- IBM SPSS Statistics version 22	Evaluation
3) Rasch model analysis	- ACCT scores - Collocational competence logits	- Winsteps 3.81.0	Domain to utilisation

Table 4.3

Summary of analytical methods, data sources, and software

Analytical methods	Data sources	Programme/software	Inference
4) Correlation analysis	- ACCT scores - AVLT scores - Collocational competence logits - Vocabulary knowledge logits	- IBM SPSS Statistics version 22	Explanation
5) Analysis of variance	- ACCT scores - AVLT scores	- IBM SPSS Statistics version 22	Extrapolation
6) Cut-score setting	- Collocational competence logits	- WinBUGS 1.4.3 - Winsteps 3.81.0 - IBM SPSS Statistics version 22 - Microsoft Excel 2010	Utilisation
7) Classification error estimation	- Collocational competence logits	- IBM SPSS Statistics version 22	Utilisation
8) Test reflection analysis	- Quantitative and qualitative responses from the questionnaire	- IBM SPSS Statistics version 22 - Microsoft Excel 2010	Explanation

4.4.2 Data preparation and screening

After the data were collected, data analysis was processed into Microsoft Excel 2010 and IBM SPSS statistics version 22 for PC for further analyses. Data processing was double-checked and screening for missing and completeness of the data. Initially, 199 EFL graduate students took the ACCT. Six examinees were excluded as they did not provide complete data that could be analysed in this study. Of the six students excluded, one student did not complete too many ACCI items and the others did not report any standardised English test scores. As a result, the data set was obtained only from 193 EFL graduate students.

4.4.3 Descriptive statistics

Descriptive statistics were performed using IBM SPSS Statistics version 22 in order to simplify the organisation and presentation of data. A frequency distribution graph was carried out using a stacked histogram to see whether the shape of a frequency distribution of the ACCT scores is symmetrical or skewed. Skewness and kurtosis statistics were used to describe the univariate normality of the ACCT score distribution. A skewness statistic informs the degree of asymmetry the data distribution shape. If the distribution has a tail going out to the left, the data distribution is negatively skewed. If the distribution has a tail going out to the right, it is positively skewed. A skewness value of zero indicates the distribution is symmetric (Howell, 2013). A positive skewness value greater than 1 or 2 indicates a positively skewed distribution, while a negative skewness value greater than -1 or -2 indicates a negatively skewed distribution.

Another measure of the degree of asymmetry the data distribution shape is a kurtosis. A normal distribution is called mesokurtic, meaning that the distribution tails are neither too thin nor too thick, and there are neither too many nor too few scores concentrated in the center. If a kurtosis value the curve becomes flatter and is called platykurtic. If the curve becomes more peaked with thicker tails, such a curve is called leptokurtic. A kurtosis value of zero indicates a normal shape of the score distribution (Howell, 2013). A negative value greater than -1 or -2 indicates a distribution which is a leptokurtic shape, whereas a positive kurtosis greater than 1 or 2 indicates a platykurtic shape. For psychometric purposes, skewness and kurtosis values should fall between -2 and 2 to indicate an acceptable normal distribution of the data.

A measure of central tendency included the mean and it informs the single value that identifies the center of the distribution and best represents the entire set of scores. A measure of variability included the range and the standard deviation. It informs whether the ACCT scores are spread out or clustered together. The standard deviation describes whether the scores are clustered closely around the mean or are widely scattered. Descriptive statistics helps determine whether the ACCT score data is normally distributed, which is a requirement for a norm-referenced score interpretation and other parametric inferential statistics.

4.4.4 Rasch measurement analysis

Rasch measurement analysis was performed using Winsteps software (Version 3.81.0). Several applications of the dichotomous Rasch model were performed including unidimensionality investigation, internal consistency estimation, item measure calibration, person-item variable map, person-item babble map, multiple-choice distractor functioning, differential test functioning, and differential item functioning. These Rasch-based analyses are described in detail in the subsequent section.

4.4.5 Unidimensionality investigation

Unidimensionality investigation informs whether the ACCT items measure a unitary, single latent collocational competence under measure and whether local independence assumption is met. Local independence checks the probability that an individual examinee's correct response to a particular ACCT item is not influenced by any other ACCT items on the ACCT. Unidimensionality analysis was investigated based on PCAR, item fit indices, and the point-measure correlation. To signify a significant unidimensional collocational competence in the score data, the first factor needs to be accounted for at least 20% of the variance for item parameters to be stable (Reckase, 1979) and items should exhibit acceptable fit statistics. When unidimensionality was achieved, it was assumed that local independence was also met (Bond & Fox, 2007; Boone et al., 2014; Hambleton et al., 1991).

Unidimensionality can also be checked using the point-measure correlation. The point-measure correlation (otherwise called a point-biserial correlation or a discrimination index) indicates the degree to which the scores on a particular item are consistent with the average score across the remaining items. To indicate as such, Wolfe and Smith (2007b) the point-measure correlation coefficient should be positive and the observed value should be close to the expected value. Furthermore, the point-measure correlation should exceed .3 to be more appropriate for a norm-referenced evaluation and should be positive and relatively equal to secure Rasch equal discrimination requirement. Results from unidimensionality analysis serve as empirical evidence supporting the assumptions underlying the warrants of the domain, evaluation, and explanation inferences.

4.4.6 Internal consistency estimation

Internal consistency indices of person and item measures were estimated through item reliability, item separation, item strata, person reliability, person

separation, and person strata. Person separation reliability, separation index and strata indicate how well items on the ACCT are able to separate examinees in this group on a continuum of the underlying collocational competence being measured. In other words, it indicates whether the person competency estimates are adequately dispersed along a competency continuum. Internal consistency was checked through item reliability, item separation, item strata, person reliability, person separation, and person strata.

The Person reliability, separation and strata coefficients indicate how well items on the ACCT are able to separate this sample of examinees on a continuum of the underlying collocational competence being measured. In other words, it indicates whether person collocation competency estimates are adequately dispersed along the competency hierarchy scale (Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Schumacker & Smith, 2007; Wolfe & Smith, 2007b). The person reliability indicates how well the ACCT is capable of separating examinees on a continuum of the underlying collocational competence being measured. It is correspondent to the traditional KR-20 or coefficient alpha internal consistency reliability in the classical test theory. A high person reliability coefficient is over 0.8. KR-20 coefficient alpha, the person reliability suffers from ceiling effects and thus to avoid this ceiling effect, a person separation index can supplement the person reliability. The person separation indicates the dispersion of person competency measures in standard error units. The higher the person separation value, the more dispersed person competency measures on the scale.

The person separation coefficient of 2 is equivalent to the person reliability coefficient of 0.80. Therefore, a good person separation should be at least 0.2. Another index is the person strata which indicates the number of statistically distinct levels (separated by at least 3 SEM) of student competency that the 30-item ACCT discriminated. The person strata should be at least 2 to indicate that the ACCT items distinguished ACCT a group of 193 EFL graduate examinees into low and high competence (Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Schumacker & Smith, 2007; Wolfe & Smith, 2007b). The item reliability indicates how well the group of examinees is capable of separating item difficulty estimates on a continuum of the underlying collocational competence being measured. Like the person reliability, a minimum criterion for good item reliability is 0.8. Item reliability also suffers from ceiling effects and hence the item separation index can be used to support the item reliability. The item separation indicates the dispersion of item difficulty measures in standard error units. The higher the item separation value, the

more dispersed item difficulty measures are. As with the person separation index, a minimum value for item separation index is 0.2.

Another index is the item strata which indicates the number of statistically distinct levels (separated by at least 3 SEM) of ACCT item difficulties that this group of 193 EFL examinees distinguished. The item strata should be at least 2 to indicate that this group of 193 EFL examinees distinguished discriminate ACCT item difficulties into two strata or difficulty levels, easy and difficult items (Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Schumacker & Smith, 2007; Wolfe & Smith, 2007b). Results gained from internal consistency indices were used to support many assumptions of the inferences. All the indices supported the assumptions underlying the warrant of the generalisation inference. The Rasch item strata yielded empirical evidence supporting the assumptions underlying the warrant of the domain inference and the Rasch person strata provide empirical evidence for the assumptions underlying the warrant of the extrapolation inference.

4.4.7 Item measure calibration

Item difficulty, standard error of estimates, item fit indices, and point-measure correlation were performed and presented in the same table. To evaluate the fit of the items to the Rasch model, Infit and Outfit statistics based on the unweighted mean-squared fit indices (Mnsq) and the unweighted standardised mean-squared fit indices (Zstd) were checked. There are two types of misfitting items: underfitting and overfitting. Underfitting items demonstrate that the ACCT items may not primarily assess a unidimensional construct of collocational competence while overfitting items show that the ACCT items may be redundant. Misfitting items should therefore be considered for deletion to eliminate noise or data redundancy from the analysis and in particular underfitting items are of grave concern. However, deleting misfitting items based solely on item fit criteria without taking into account content representativeness or coverage might cause the instrument to fail to capture core aspects of the construct, causing construct underrepresentation (Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b).

To indicate the fit of items to the expected Rasch model, the ideal Mnsq value is 1 but the acceptable Mnsq value ranges from 0.5 to 1.5. Items having Mnsq values outside the acceptable range are considered as misfit to the ideal Rasch model. Items displaying Mnsq values less than 0.5 and greater than 1.5 are regarded as overfit and underfit respectively. To avoid the problem concerning the Type I error rates, influenced by sample size and test length, the Mnsq is transformed to the

Zstd. The ideal Zstd value is 0. For a sample size of less than 1000 examinees, the acceptable Zstd value ranges from -2 to 2. Items having Zstd values outside the satisfactory range are considered as misfit to the expected Rasch model. Items having Zstd values less than -2 and greater than 2 are considered as overfit and underfit respectively (Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b).

Overfit items, which have Mnsq and Zstd values less than 0.5 and -2 respectively, are caused by measurement problems such as redundancy and on the one hand underfit items, which have Mnsq and Zstd values greater than 1.5 and 2 respectively, are resulted from such measurement problems related to unexpected multidimensionality or unpredictable responses such as lucky guessing and careless responses (Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b). Smith (2005) suggested that where the proportion of overfitting items is less than 5%, item difficulty and person ability estimates are not affected substantially and Linacre (2012) proposed that Mnsq values between 1.5 and 2.0 may be unproductive for construction of measurement but they are not degrading the Rasch model. Information gained from item fit indices was used to support the unidimensionality analysis which in turn serves as empirical evidence in support of the assumptions underlying the warrants of the domain description, evaluation, and explanation inferences. The criteria for assessing item fit are also applied to assess person fit.

4.4.8 Person-item variable map investigation

The person-item variable map is a visual diagram showing the distribution of student ccollocational competency and item difficulty, both calibrated on the comparable, common logit scale. The mean of item difficulty and student competency is usually set to 0 logits or measures. The mean student competency is compared with the mean item difficulty to see if ACCT items, on average, are difficult or easy for this group of EFL graduate examinees. The distribution of student competency is supposed to be matched with item difficulty distribution when norm-reference interpretations are of interest (Linacre, 2012). It is possible to compare the locations of students on the left side of the map with the locations of items on the right side of the map to see whether there are noticeable gaps in the item distribution that does not have items that precisely measure students who are at that level of competency.

The person-item variable map shows a visual distribution of student collocational competencies and ACCT item difficulties. Both student and item

estimates should be widely dispersed to adequately represent the collocational competence and TLU content and should be well matched with each other for the precision of estimates. The student competency distribution also informs if the ACCT did measure effectively what it was claimed to measure (Baghaei, 2008; Linacre, 2012). Information from the person-item variable map is used as empirical evidence in support of the assumptions underlying the warrants of the domain, evaluation, generalisation, explanation and extrapolation inferences.

4.4.9 Person-item babble map investigation

The person-item babble map is a visual vertical and horizontal diagram displaying the locations of student variable collocational competencies and ACCT item difficulties on the latent collocational competence measure. Viewed from a vertical line, it visually informs the information related the precision (reliability) and standard error of measurement of student and item estimates. The precision and standard error of measurement of student and item estimates can be observed by the babble size (Linacre, 2012). The bigger the babble is, the higher the error, and hence the lower the precision of the estimates. Looked from a horizontal line, the map depict graphically the information on the accuracy of item and student estimates.

The accuracy of person and item estimates is expressed in terms of how far items and persons, represented by babbles, are from the acceptable Outfit Mnsq zone on the horizontal axis. The farther the babble symbols from the acceptable Outfit Mnsq zone, the lesser the items and students fit the expected Rasch model, and thereby the lower the accuracy of the estimates (Linacre, 2012). Information gained from the person-item babble map serve as empirical evidence strengthening the assumptions underpinning the warrants of the generalisation and explanation inferences.

4.4.10 Multiple-choice distractor functioning analysis

The multiple-choice distractor analysis indicates the extent to which the responses to the distractors are consistent with the intended cognitive process around which distractors are constructed. Ideally, good distractors should attract equally small proportions of testees but in reality distractor are not equally attractive or selected in practice. To function as intended, it is recommended that each distractor should be selected by 5% of the respondents. The proportion of

respondents selecting each distractor should not outnumber that of the respondents selecting a correct choice. The proportion can be observed though the p-value index. It is also expected that distractors should attract lower-ability examinees and correct choices should be selected by higher-ability examinees. This indication can be observed though the average ability measure index.

The average ability measures of examinees for distractors should not exceed the average ability measure of a correct option in a particular item. The average ability measure index can be considered in conjunction with the distractor-measure correlation (analogous to the point-measure correlation). Distractors should have negative correlation values to indicate that examinees with lower-ability estimates are attracted by these distractors. If one of the distractor function properties is not met in a test item, then such item does not function as intended and needs to be revised (Wolfe & Smith, 2007b). Results from analysis of multiple-choice distractor functioning provide empirical evidence in support of the assumption underlying the warrant of the explanation inference.

4.4.11 Differential test functioning analysis

Differential test functioning (DTF) was performed to investigate the measurement invariance property of the ACCT. It indicates whether all items on the ACCT function the same way for male and female students. DTF can be investigated by separating males and females and then estimating the difficulty of each item for male and female subgroups. The DTF scatterplot shows the comparison of two sets of item difficulty between females on the y-axis and males on the x-axis. The dashed line is a trend line or a line of commonality through the mean of two sets of items. The blue and red curves demarcate approximate 95% confidence bands. Items that fall outside of this confidence bands are not invariant or are easier or more difficult for a particular group, meaning that their difficulty estimates vary according to gender. A DTF scatterplot was also used to give a picture of potential ACCT bias on individual items (Linacre, 2012). Results from analysis of DTF are used as empirical evidence in support of the assumption underlying the warrant of the generalisation inference.

4.4.12 Differential item functioning analysis

DIF takes the form of uniform and non-uniform. Uniform DIF exists when two subgroups of examinees (males and females) perform differently on a test item,

while non-uniform DIF exists when the difference in performance varies with ability level. The type of the DIF analysis in this study is a uniform DIF. A uniform-DIF exists when a particular item has different difficulty estimates (logits or measures) for all competency levels of males and females. DIF can also be checked through a DIF scatter plot. Two sets of item difficulty logit for male and female groups are plotted by estimating difficulty measure for each item for each group while holding other item difficulty and student competency estimates constant. A huge gap between males and females for a particular item indicates that a difficulty estimate is different for males and females. It should be noted that Rasch-based DIF analysis is based on preconditions: unidimensionality and local independence.

Dimensionality and DIF analyses differ in that dimensionality analysis provides information regarding secondary dimensions that are relevant to all examinees, whereas DIF analysis provides information about conditional differences in response probabilities using defined variables (such as gender) that dimensionality analysis does not examine. A Welch *t*-test ($t > 1.96$) and a *p*-value threshold of .05 ($p < .05$) were used to inform significant DIF and if difficulty estimate difference (DIF contrast) between males and females exceed 0.5 logit, it is considered as exhibiting a critical huge DIF size (Linacre, 2012). Results from DIF analysis serve as empirical evidence in support of the assumption underlying the warrant of the generalisation inference.

4.4.13 Correlation analysis

Correlation analysis was performed using IBM SPSS Statistics version 22. The data included the ACCT scores and AVLT scores as well as person collocational competence logits and person vocabulary size knowledge logits. The Pearson Product-Moment correlation was used to find the relationship between the ACCT and the AVLT. A correlation coefficient of greater than .8 indicates a high degree of the relationship. Results from correlation analysis are used as empirical evidence in support of the assumption underlying the warrant of the explanation inference.

4.4.14 Analysis of variance

Analysis of variance was performed using IBM SPSS Statistics version 22. The ACCT scores were used as a dependent variable and the proficiency groups were used as a factor or an independent variable in this one-way independent ANOVA. ANOVA assumptions of normality and homogeneity of variance were investigated through descriptive statistics and the Levene test respectively. If the data are

homogeneous and based on equal sample sizes, a Tukey HSD is used for post-hoc comparisons. If the data are heterogeneous and based on somewhat different sample sizes, a Games-Howell test is better used for post-hoc comparisons (Howell, 2008). A boxplot diagram was also created to show the ACCT score distributions for three proficiency levels. Results from analysis of variance provide empirical evidence for the assumption underlying the warrant of the extrapolation inference.

4.4.15 Test reflection survey analysis

Test reflection survey was investigated using IBM SPSS Statistics version 22 and content analysis. Data from test reflection survey involved both quantitative data from questions 1 and 2 as well as qualitative data from question 3. Responses to Questions 1 and 2 of the survey were analysed using a table chart and a chi-square test for independent. Responses to Question 3 were coded to support the results from Questions 2. Results from analysis of test reflection survey responses serve as empirical evidence supporting the assumption underlying the warrant of the explanation inference.

4.4.16 Cut-score establishment

Cut-score study was based on performance-based standard setting using a contrasting-group method (Livingston & Zieky, 1982) where empirical data were used as the foundation for determining cut scores. In this study, students' ACCT scores and competency logits from three proficiency levels were used as empirical data for locating cut-score thresholds. Cut scores are values or thresholds that demarcate the pass or failure, or competency levels. Frequency distributions of ACCT scores and collocational competency logits are generated for each of the three proficiency groups. Trendlines for each distribution are created and the intersection points of the trendlines between low and mid-proficiency group distributions and between mid and high-proficiency group distributions are used to set the cut scores for classifying examinees into low, mid, and high competency levels. Results from the cut score study were used as empirical evidence substantiating the assumptions underlying the warrant of the utilisation inference.

4.4.17 Classification error estimation

Classification error estimation was performed using IBM SPSS Statistics version 22 and WinBUGS 1.4.3. The cut scores identified in the cut score study are used see

to what extent the established cut scores produce negative and positive false classification. In other words, to what extent these cut scores are accurate and consistent in classifying examinees. The classification error estimation is based on two approaches, One estimation method was based on Livingston and Zieky (1982) and the other was based on a Bayesian approach. The data used for Livingston and Zieky's approach include ACCT scores and Rasch competency logits and the data used for a Bayesian approach are Bayesian Rasch competency logits. Student competency measures or logits are estimated by Winsteps and WinBUGS. Results from analysis of classification error were used as empirical evidence in support of the assumption underlying the warrant of the utilisation inference.

4.5 Chapter summary

What I have presented previously in this chapter is concerned primarily with research methodology, ranging from presentation of backgrounds regarding participant characteristics, research instruments, data collection procedure to data analysis procedure. A final data set was based on 193 EFL graduate students and was collected using the ACCT, the AVLT and the test reflection questionnaire. The data gained from these three instruments were empirical data which were analysed to provide empirical evidence in favour of the assumptions in the interpretive argument. The information provided in this chapter, in particular data analyse procedure, indeed guides the way in which the results of data analysis are presented and discussed in the next chapter. The next chapter has to do with the presentation of results obtained from empirical data analyses, which were introduced in this chapter.

CHAPTER 5
RESULTS AND DISCUSSION

In this chapter, I present the results from empirical data analysis to support the interpretive argument. The results provide empirical evidence which serves as backing for each assumption underlying the warrant of inferences in the validity argument. This chapter begins by presenting results from descriptive statistics of the total scores on the ACCT. It then presents empirical results from several applications of the Rasch measurement model. Following this, results from correlation analysis, analysis of variance, test reflection analysis, cut-score analysis, and classification error analysis are presented. All these necessarily provide evidentiary support for or challenge the assumptions underlying the inferences in the interpretive argument for the ACCT. This chapter ends with a short summary of the contents presented in this chapter.

5.1 Descriptive statistics

As presented in Table 5.1, descriptive statistics of the ACCT scores from 30 items and 193 EFL graduate students revealed that the ACCT score data were normally distributed. The mean score of the students was 14.13 with the standard deviation of 6.85, meaning that there was variability in the ACCT scores. The values for kurtosis and skewness did not exceed the range of -2 and +2. The skewness value of .37 indicated a slightly negatively skewed distribution and this was due to the fact that low-proficiency students outnumbered mid and high-proficiency students. The kurtosis value of -1.092 indicated a relatively flat distribution. Figure 5.1 shows a stacked histogram displaying the normal distribution of 30 ACCT items and 193 EFL graduate students. Therefore, the ACCT scores are appropriated for a norm-referenced placement decision. It should be noted, however, that the Rasch model does not assume that the data approximate a normal distribution.

Table 5.1

Descriptive statistics of the ACCT scores

Group	N	M	SD	Range	Min	Max	SK	KU
Low	84	8.47	2.69	13.00	3.00	16.00	.23	-.39
Moderate	59	15.06	5.13	21.00	4.00	25.00	-.05	-.82
High	50	22.56	3.49	15.00	14.00	29.00	-.28	-.12
Total	193	14.13	6.85	26.00	3.00	29.00	.37	-1.09

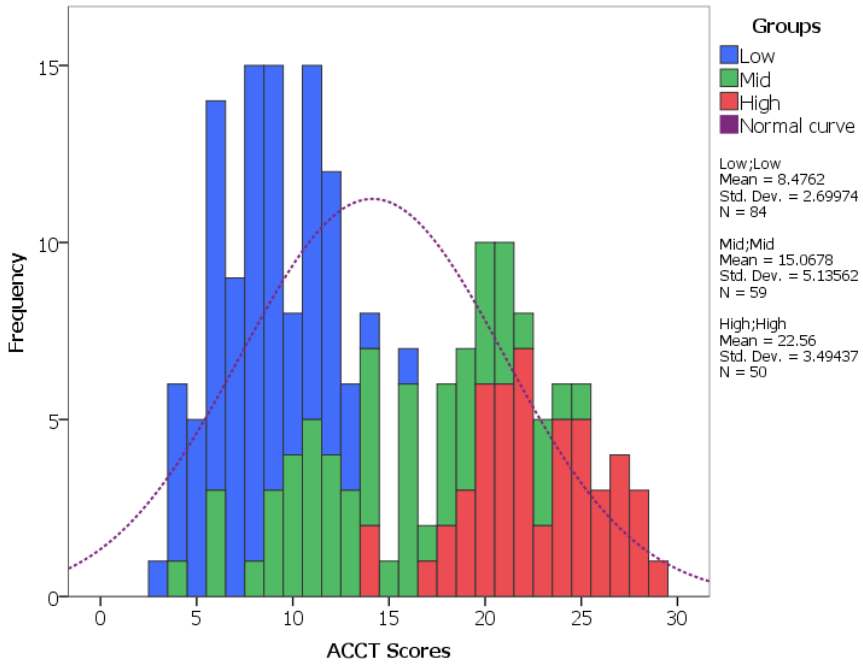

Figure 5.1. Stacked histogram showing the score distribution of 30 ACCT items and 193 EFL eaminees

5.2 Rasch measurement analysis

5.2.1 Unidimensionality

Unidimensionality and local independence was investigated through analysis of linearsed Rasch residuals (PCAR), item fit statistics, and point-measure correlation. Table 5.2 summarises indices from PCAR. PCAR was used to see if responses on the ACCT items exhibited a significant unidimension of the focal collocational competence. PCAR showed that the empirical data explained 32.0% of the observed variance in the data, which is very close to the expected Rasch model variance of 31.5%, meaning that that the computation of the Rasch difficulty estimates was successful (Linacre, 2012). The amount of the variance explained by different

components in the data was 32% with 14.9% explained by persons and 17.1% explained by items. The unexplained variance of the first contrast was 6.1 with the eigenvalue of 2.7. Reckase (1979) suggested that the variance explained by the focal factor should be greater than 20% to ensure the substantive unidimensional construct. Linacre (2012) recommended that the unexplained variance of the first contrast should not exceed 5% and the first contrast eigenvalue should not exceed 3 in eigenvalue unit.

Table 5.2

Summary of principle component analysis of standardised Rasch residual

Sources of Variance	Eigenvalue Units	Empirical Data		Rash Model
Total raw variance in observations	44.1	100%		100%
Raw variance explained by measures	14.1	32.0%		31.5%
Raw variance explained by persons	6.6	14.9%		14.7%
Raw Variance explained by items	7.5	17.1%		16.9%
Raw unexplained variance in total	30.0	68.0%	100%	68.5%
Unexplned variance in 1st contrast	2.7	6.1%	9.0%	
Unexplned variance in 2nd contrast	1.9	4.3%	6.4%	
Unexplned variance in 3rd contrast	1.7	3.8%	5.5%	
Unexplned variance in 4th contrast	1.6	3.6%	5.3%	
Unexplned variance in 5th contrast	1.5	3.4%	5.0%	

Figure 5.2 displays loading patterns of ACCT items on the first hypothesised contrast in the linearised residuals. The horizontal line represents item difficulty measures and the vertical line represents contrast loading. The ACCT items are represented as alphabetic letters. Items landing beyond the zero-loading region (above the dotted line) are labelled by capital letters, while items landing under this region are labelled by small letters. Overall, the ACCT items did not form distinguishable patterns or clusters. The ACCT items spread out in different zones of the plot, signifying unidimensionality of the measured collocational competence construct (Bond & Fox, 2007; Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b).

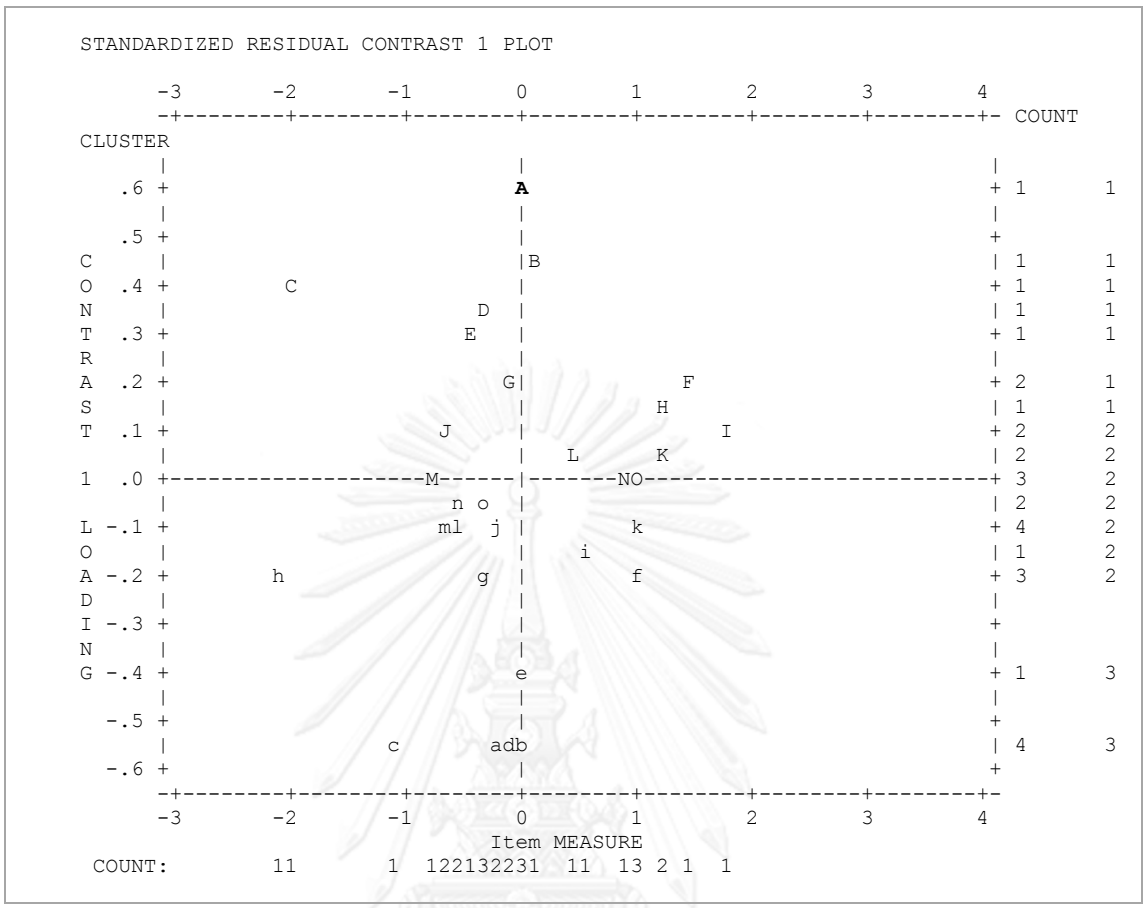

Figure 5.2. Scree plot of the standardised residual contrast in the ACCT

Since the variance of the focal collocational construct was explained by more than 20%, the first contrast eigenvalue was less than 3, and the ACCT items did not form distinguishable patterns or clusters, it was assumed that the focal collocational construct was substantively unidimensional. This was further substantiated by the evidence that 29 items possessed good fit indices and had positive point-measure correlations beyond 0.3 (see Table 5.4). On balance, PCAR, item fit statistics, and the point-measure correlation ensured the substantive unidimensionality of the measured collocational construct. Due to the present unidimensionality, it was also assumed that local independence was achieved, meaning that an individual response to a particular item is not influenced by his/her response to any other items (Bond & Fox, 2007; Embretson & Reise, 2000; Hambleton et al., 1991; van der Linden & Hambleton, 1997)

5.2.2 Internal consistency

Table 5.3 summaries the internal consistency indices of item and person measures. For 30 ACCT item, the item reliability was 0.96, indicating that 193 EFL graduate examinees very well spread ACCT item difficulties or ACCT item difficulties were widely dispersed on the item difficulty hierarchy. The item separation was 4.90, indicating that 30 ACCT items were separated into around five difficulty categories. This means that 193 EFL graduate examinees could statistically differentiate more difficult items from easier items. The item strata was 6.86, meaning that 193 EFL graduate examinees could statistically stratify 30 ACCT items into almost seven item difficulty levels. This indicates that the 30-item ACCT could sufficiently measure a wide range of EFL graduate's receptive collocational competence over a long period of time.

For 193 EFL graduate examinees, the person reliability was 0.86 and the coefficient alpha was .89, meaning that 30 ACCT items well differentiated 193 EFL graduate students in terms of receptive collocational competency; that is, student collocational competencies were well dispersed on the collocational competence hierarchy. The person separation was 2.48, indicating that student collocational competencies were classified into at least approximately two competency levels on the collocational competency hierarchy. In other words, 30 ACCT items could statistically distinguish higher-competency students from lower-competency students. The person strata index was 3.64, demonstrating that 30 ACCT items could statistically stratify 193 EFL graduate examinees into at least approximately three collocational competence levels. By and large, it suffices to say that the 30-item ACCT contained sufficient items to reliably measure this sample of 193 EFL graduate students with varying levels of receptive collocational competence.

Table 5.3

Summary of internal consistency indices

	Internal consistency indices		
Object	Reliability	Separation	Strata
Item	0.96	4.90	6.86
Person	0.87	2.48	3.64

5.2.3 Item measure calibration

Table 5.4 shows item fit statistics of the calibration of 30 multiple-choice ACCT items. Item difficulty estimates were presented in the third column. Overall, the range of item difficulty was 1.75 logits (Item 2) to -2.14 logits (Item 3). The mean item difficulty was zero (M = 0.00, SD = 0.91) and the mean standard error of estimate (S.E.) was very low (M = 0.18, SD = 0.01), meaning that the ACCT was not difficult or easy. Rasch item fit statistics showed that on a macro level, the data were fit very well to the expected Rasch model as evident by the mean Infit Mnsq (M = 1.0, SD = 0.21) and the mean Outfit Mnsq (M = 1.03, SD = 0.03). On a micro level, out of 30 items, 29 items had Infit Mnsq statistics between .5 and 1.5 and only Item 19 had an Infit Mnsq value of 1.7. Based on Outfit Mnsq statistics, 26 items had Outfit Mnsq statistics between .5 and 1.5. Items 2, 13, and 28 had an Outfit Mnsq statistic of slightly over 1.5 and Item 19 had an Outfit Mnsq statistic of 2.0, which was critically underfit to the expected Rasch model. Item 19 was most underfit and was therefore deleted prior to recalibrating the new data set. After reanalysing the new data set, Outfit Mnsq statistics of Item 2, 13, 28, remained a bit underfit and Item 23 turned out underfit to the expected Rasch model.

However, Infit Mnsq values of Items 2, 13, 23, and 28 fell within 0.5 and 1.5 and their Outfit Mnsq values were slightly beyond 1.5. If these items were excluded, the remaining items may not well represent the collocational competence construct (Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b). In this regard, I decided to keep Items 2, 13, 23, and 28 on the ACCT. The point-measure correlation in the last column indicated that all items displayed relatively equal, positive correlations and over 0.3 except for Item 19 (0.02), indicating that up to 29 ACCT items measured the same collocational construct in the same direction. All items had positive correlation coefficients and 29 items had observed correlation coefficients close to expected correlation coefficients. Only Item 19 displayed the lowest correlation close to zero and its observed and expected correlation values were noticeably different. Item fit indices confirmed that the data fit the Rasch model and thereby it can then be confident that any estimates of persons and items provided meaningful measurement properties (Iramaneerat et al., 2008) contributing to sound empirical evidence.

Table 5.4

Item estimates of 30 ACCT items based on 193 EFL graduate examinees

Item No.	Difficulty Estimates			Fit Estimates				PTM Correlation	
				Infit		Outfit			
	Score	b	S.E.	Mnsq	Zstd	Mnsq	Zstd	Obs	Exp
01. find a way	106	-0.44	0.17	1.16	2.25	1.21	1.72	0.37	0.48
02. cite a case	37	1.75	0.21	1.18	1.49	1.55	2.10	0.31	0.45
03. leave school	160	-2.14	0.20	0.95	-0.41	0.78	-0.71	0.38	0.33
04. enforce a law	100	-0.27	0.17	0.76	-3.72	0.68	-3.22	0.66	0.49
05. make an award	96	-0.16	0.17	1.13	1.69	1.13	1.14	0.41	0.50
06. cover an area	116	-0.71	0.17	0.90	-1.54	0.77	-1.79	0.55	0.47
07. see figure	92	-0.05	0.17	0.65	-5.31	0.59	-4.44	0.73	0.5
08. obtain a result	103	-0.36	0.17	1.11	1.55	1.06	0.54	0.42	0.49
09. provide an example	117	-0.74	0.17	1.02	0.32	1.12	0.90	0.43	0.46
10. improve health	113	-0.63	0.17	1.08	1.14	1.20	1.48	0.41	0.47
11. conduct a study	96	-0.16	0.17	0.61	-6.23	0.54	-5.06	0.76	0.50
12. have an idea	71	0.55	0.17	0.89	-1.29	0.85	-1.24	0.58	0.51
13. make sense	156	-1.98	0.20	1.15	1.39	1.56	1.87	0.19	0.35
14. justify belief	111	-0.58	0.17	0.94	-0.94	0.87	-1.05	0.52	0.47
15. hold the view	61	0.86	0.18	0.95	-0.55	0.97	-0.17	0.53	0.50
16. account for the fact	57	1.00	0.18	0.92	-0.81	0.90	-0.64	0.55	0.50
17. play a part	56	1.03	0.18	0.92	-0.84	0.84	-1.07	0.56	0.50
18. pursue a policy	56	1.03	0.18	1.00	0.06	0.99	0.02	0.49	0.50
19. fight the war	89	0.03	0.17	1.70	7.71	2.00	7.22	0.02	0.50
20. exercise power	51	1.20	0.19	0.89	-1.13	1.13	0.76	0.53	0.49
21. introduce a system	49	1.27	0.19	0.97	-0.26	0.99	0.00	0.49	0.48
22. apply a rule	103	-0.36	0.17	0.98	-0.31	0.94	-0.47	0.51	0.49
23. carry on a business	85	0.15	0.17	1.38	4.38	1.43	3.48	0.26	0.50
24. appoint an expert	89	0.03	0.17	0.73	-3.90	0.70	-3.10	0.68	0.50
25. terminate a contract	97	-0.19	0.17	0.91	-1.20	0.87	-1.16	0.55	0.49
26. use a word	129	-1.09	0.17	0.76	-3.66	0.63	-2.58	0.61	0.44
27. do work	103	-0.36	0.17	0.93	-0.97	0.88	-1.08	0.54	0.49
28. read text	45	1.42	0.20	1.20	1.86	1.59	2.64	0.31	0.48
29. have a disease	75	0.43	0.17	1.15	1.78	1.13	1.13	0.41	0.51
30. treat a group	110	-0.55	0.17	0.98	-0.31	0.92	-0.64	0.50	0.48
Mean	91.0	0.00	0.18	1.00	-0.30	1.03	-0.10		
S.D.	30.2	0.91	0.01	0.21	2.70	0.33	2.40		

* Note: b = Item difficulty measure

Figure 5.3 shows item characteristic curves (ICCs) of 30 multiple-choice items on the ACCT. The ICC displays a monotonically increasing relationship between the measures relative to item difficulty and the probability of a correct response on a test item. As a whole, all ACCT items exhibited similar slop curves or similar discrimination indices or relative equal point-measure correlations yet different difficulty measures. This is correspondent with the hypothesised Rasch model where discrimination indices of all items are held constant and item guessing indices are set to zero, and item difficulty varies according to person ability. The overall pattern of ICCs of 30 multiple-choice items is substantially similar. Point-measure correlations of the ACCT items were not varied, supporting equal slop curves or similar levels of item discrimination.

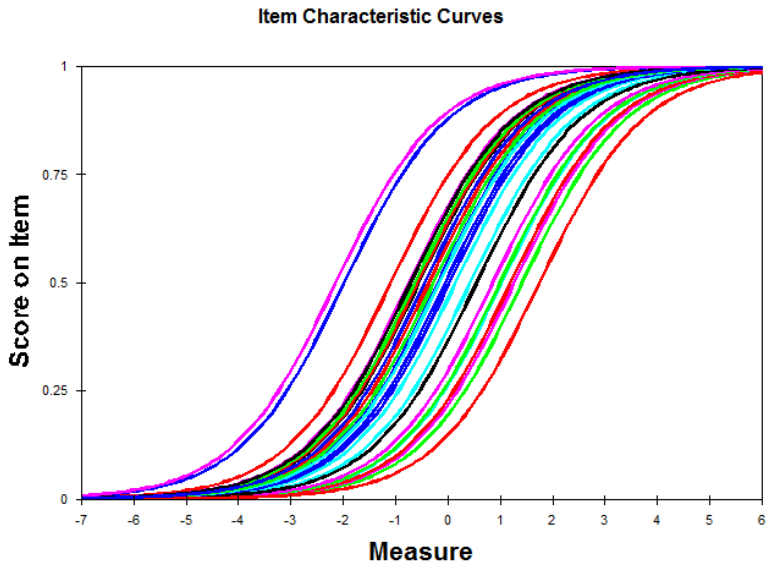

Figure 5.3. Item characteristic curves of 30 ACCT items

Figure 5.4 displays an ICC of Item 19. The red curve in the middle is the ICC as expected by the Rasch model and is positioned based on the difficulty of Item 19. ICC demonstrates the Rasch-model probability of how test-takers at different ability measures along the latent variable on an x axis would on average respond correctly to the item on a y-axis. The blue line with ‘‘x’’ is the empirical ICC based on the data. The blue line was expected to position in accord with the red line. Item 19 seemed problematic because observed responses to Item 19 were outside the confidence intervals (black lines) which were defined by 1.96 standard deviations, and the empirical ICC of Item 19 demonstrated inconsistent downward and upward

trends, indicating that there might be a second sub-dimension in this item and the empirical ICC of Item 19 did not correspond to the Rasch expectation in that more competent examinees have higher probability of answer easy items correctly, while the less competent have lower probability of answer easy items correctly. Special attention to Item 19 was paid in the subsequent analyses of multiple-choice distractor functioning and uniform DIF.

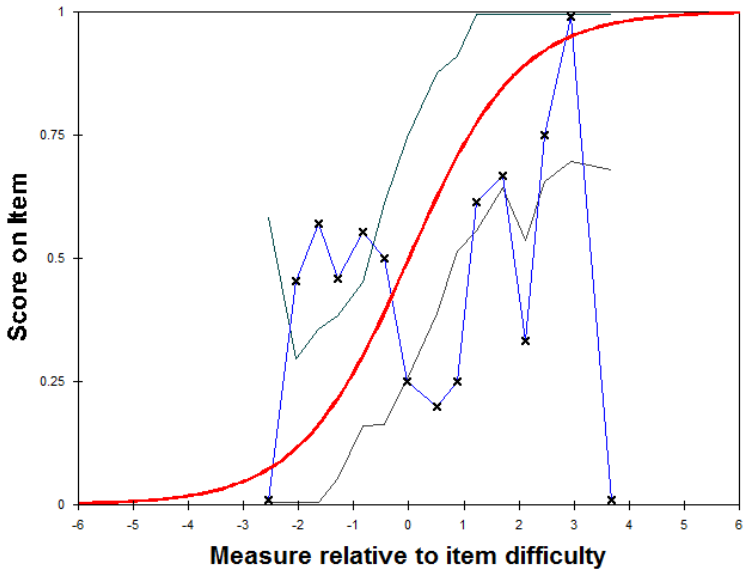

Figure 5.4. Item characteristic curve of ACCT Item 19

5.2.4 Person-item variable map

Figure 5.5 displays the person-item variable map providing a graphic summary of the results of scaling ACCT score data to the dichotomous rasch model. The Winsteps calibrated student collocational competency and item difficulty onto the common log-odds unit or logit scale, resulting in a single frame of reference for analysing and interpreting the results as well as comparing estimates between student collocational competency and item difficulty. The first column of the map shows the common measure scale, onto which collocational competency and item difficulty were calibrated. The common measure scale ranged from 4 logit at the top down to -3 logit at the bottom. The second column of the map shows the student measures of collocational competency. Students are labelled “H”, “M”, and “L”, representing students with high, mid and low English proficiency, classified initially. More competent students were located toward the top of map while less competent students were placed toward the bottom.

The third column displays the measures of item difficulty. As with the order of person competency hierarchy, the hardest items appear at the top, whereas the easiest items appear at the bottom. On the common measure scale, students who were mapped higher than item difficulty locations had higher probability of answering such items correctly. By and large, the mean of the ACCT difficulty was slightly higher than that of the student collocational competency. As shown in the map, students had the competency mean (M) at -.12 logit with standard deviation (S) of 1.27 and the range from 3.72 to -2.51 logits. Student competencies were widely spanned and equally spaced, meaning that students were well differentiated by ACCT items. Student competency distribution shows a relatively positively skewed, symmetrical and platykurtic distribution. This is probably due to the fact that low-proficiency examinees outnumbered other proficiency examinees in the sample group. Item difficulties were well spanned yet unequally spaced, with two huge gaps appearing at the top and the bottom of the scale. The ACCT items had the difficulty mean at 0 logit with a standard deviation of 0.91 and the range from 1.75 to -2.14 logits. Considering two existing gaps, item difficulties were well dispersed on the item difficulty scale, indicating that ACCT items were well differentiated by students and relatively representative of the measured collocational competence.

Although the ACCT items did not perfectly target or precisely assess all EFL graduate students, it still precisely estimated most students with a rather wide range of collocational competence. This was due to two huge gaps in ACCT item distribution. These gaps implied that the ACCT lacked items that precisely estimated some students with competencies beyond 2 logit and around 1.6 logit, thereby decreasing certain degree of construct representativeness. More items need to be added on the ACCT to fill these gaps for a well-matched person-item distribution, thereby enhancing the precision of competence estimates and the representativeness of the collocational construct (Baghaei, 2008; Linacre, 2012).

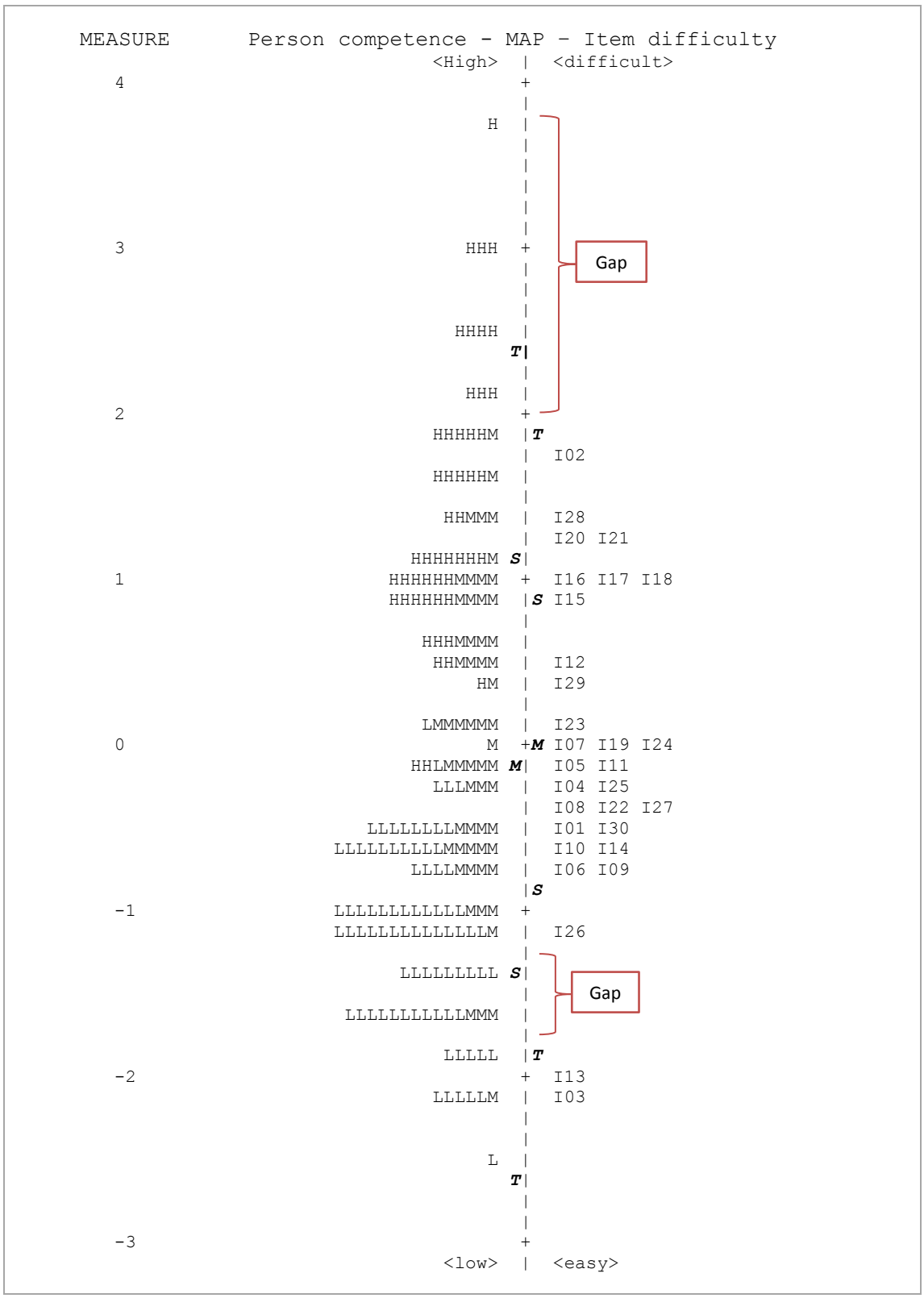

Figure 5.5. Person-item variable map of 193 EFL examinees and 30 ACCT items

5.2.5 Person-item babble map

Figures 5.6 and 5.7 present the person-item babble map portraying visual information regarding the precision and accuracy of person competency and item difficulty estimates. The babbles in darker colour represent 193 person measures, whereas the babbles in lighter colour represent 30 ACCT item measures. The precision of estimates is examined through standard error of measurement, while the accuracy of estimates is examined through the model fit. The person-item babble map alights each of person measures (darker-colour babbles) and item measures (lighter-colour babbles) vertically onto the same standardised interval logit scale. The scale has equal distances or units and ranges from +5 at the top, 0 in the middle, and down to -4 at the bottom. Higher values represent more competent persons and more difficult items, whereas lower values represent less competent persons and less difficult items. The standard error of measurement of person and item estimates is expressed by the size of the babbles, the larger the babbles, the greater the errors, and hence the lower the precision of estimates. The accuracy of person and item estimates is expressed in terms of how far item and person measures are from the acceptable Outfit/Infit Mnsq zone on the horizontal axis. The farther the babbles are from the acceptable Outfit/Infit Mnsq zone, the lesser fit the person and items to the expected Rasch model, and hence the lower the accuracy of estimates.

Item and person measures are horizontally located onto the standardised logit scale, ranging roughly between +4 and -4. Items and persons that acceptably fit the expected Rasch model are located within the Outfit/Infit Mnsq range of 0.5 to 1.5. The present study focused on item fit investigation. Items falling outside of this zone on the left are considered as overfit items, indicating that the responses are too predictable, whereas items falling outside of this zone on the right are considered as underfit items, indicating that responses to these items are too unpredictable. As portrayed in the maps, Item 19 was located the farthest from the acceptable Outfit/Infit Mnsq zone and thereby most underfit to the expected Rasch model, meaning that responses to the item are too unpredictable and may measure some related sub-dimensions that are irrelevant to the focal construct of the collocational competence. This indicates an indication of construct-irrelevant variance. For precise measurement, item difficulties should measure a single unidimensional construct, and spread out widely on the item difficulty hierarchy (Baghaei, 2008; Iramaneerat et al., 2008; Linacre, 2012; Schumacker, 2004; Wolfe & Smith, 2007b). The person-item babble map can be interpreted in conjunction with item measure statistics and the person-item variable map.

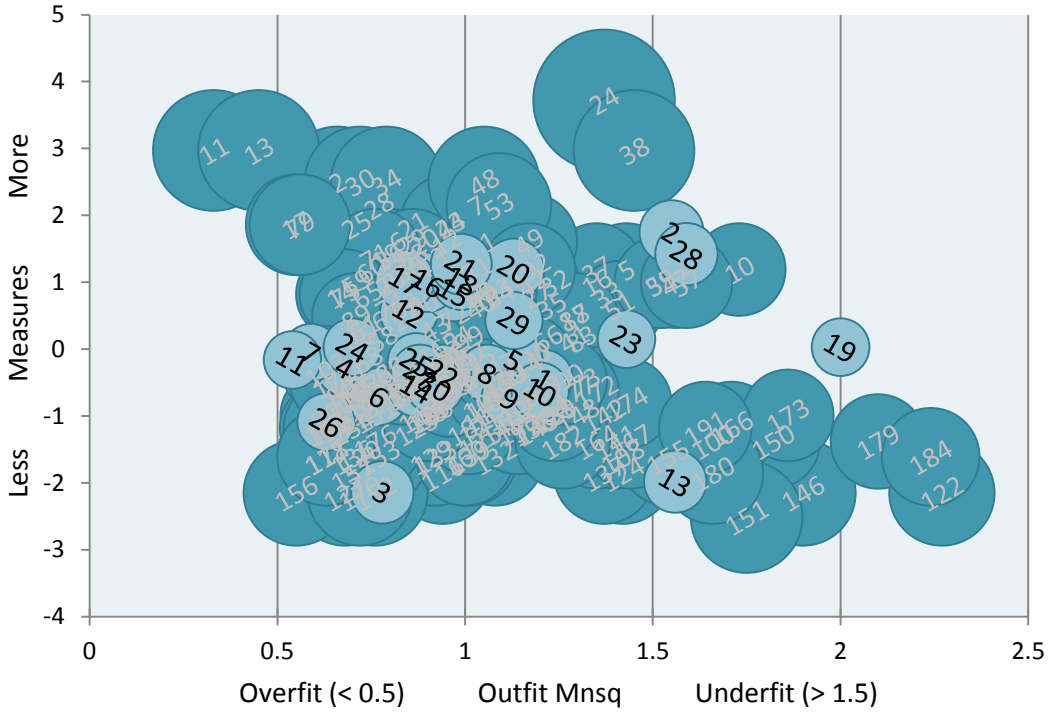

Figure 5.6. Person-item babble map by Outfit Mnsq

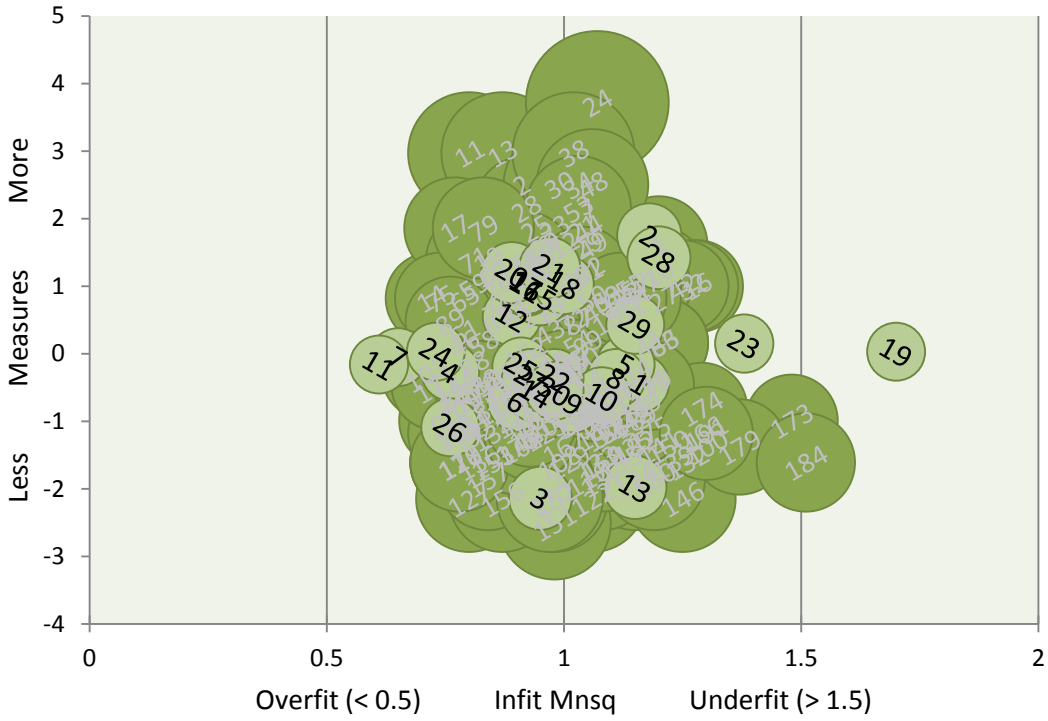

Figure 5.7. Person-item babble map by Infit Mnsq

5.2.6 Multiple-choice distractor functioning

Table 5.5 presents a summary of distractor statistics of Item 19, which was most underfit to the Rasch model. Of all 30 items, only Item 19 did not function in an intended way. The average competency measures of distractors exceeded the average competency measure of a correct choice (d), marked with an asterisk above its average competency measure. This means that the student proportion responding to a correct choice was not greater than that of those choosing other distractors. These distractors needed to be checked as they did not well elicit responses that were consistent with the intended cognitive process. The distractors in other items functioned well in consistence with the intended cognitive process as their average competency measure values were all lower than were the correct options.

Table 5.5

Summary of multiple-choice distractor functioning statistics of ACCT Item 19

			Response		Average	S.E	Oufit	PT
Item	Choice	Score	Count	%	Ability	Mean	Mnsq	Measure
19	e	0	22	11	-.54	.22	1.1	-.12
	c	0	13	7	-.42	.29	1.3	-.06
	a	0	31	16	-.04	.21	1.9	.03
	b	0	38	20	.10	.21	3.4	.09
	d	1	89	46	-.09*	.14	1.9	.02

* Average ability does not ascend with category score

5.2.7 Differential test functioning

To check invariant measurement on the test level, I performed gender-based DTF analysis. Figure 5.8 displays a DTF scatterplot of item difficulties for 65 males and 128 females. The majority of items were placed within two control lines except for Items 25, 29 and 30 which were slightly noticeably located outside control lines. These items were further examined through a uniform-DIF analysis on the item level. As displayed in Figure 5.8, there were only some outlier items deviated from the commonality line and only a few items located slightly outside the two 95% control confidence lines. These deviated items may be resulted from other measurement errors, in the usual Rasch estimation procedure. It is sound to say then that on the item level, the item estimates on the test level were not significantly invariant across gender subgroups.

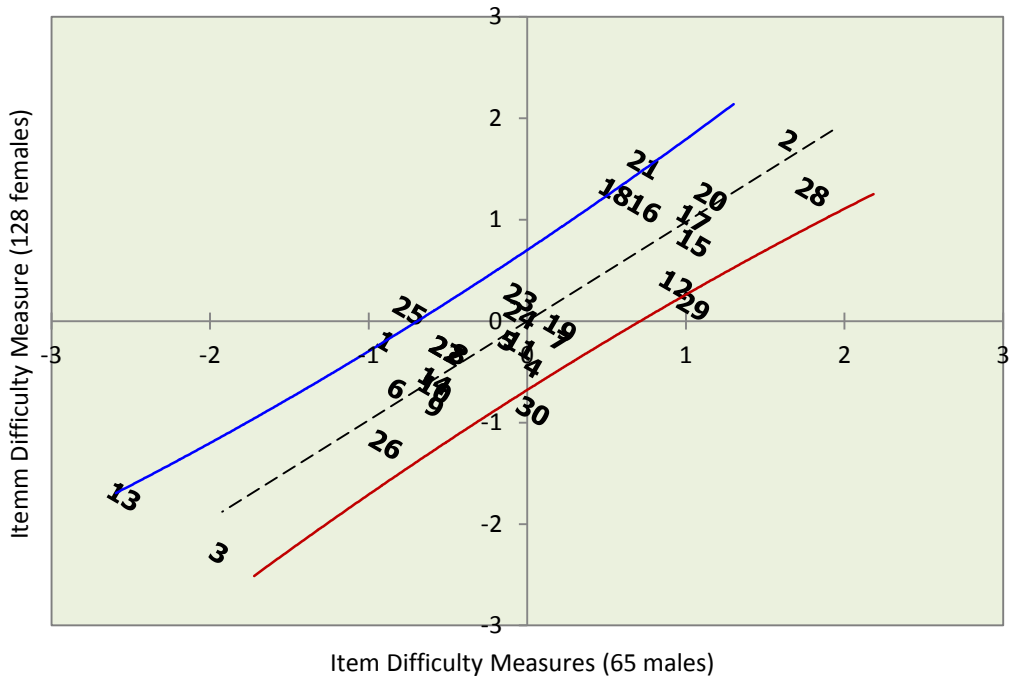

Figure 5.8. Differential test functioning by gender

5.2.8 Uniform differential item functioning

Table 5.6 displays results of a uniform-DIF analysis of 30 ACCT items. The second and third columns display the item difficulty estimates (DIF estimates) for male (M) and female (F) subgroups respectively and the forth column shows the difference (DIF size) between item difficulty for males and for females on the same item. The fifth and sixth columns show standard error (DIF S.E.) of difficulty estimates for the male and the female respectively. Rasch Welch t-value and a p-value for testing the significant DIF contrast are displayed in the right-most columns. Based on Rasch-Welch DIF test, ACCT Items 1, 21, 25, 29 and 30 exhibited significant DIF with a t-value greater than 2 and a p-value less than .05.

Based on Mantel-Haenszel DIF test, nonetheless, only ACCT Items, 25, 29, and 30 exhibited significant DIF. The DIF size for all items was also greater than 0.5 and consequently these DIF items were critical and needed further investigation of DIF causes (Linacre, 2012). ACCT Items 1, 21, 25 favoured a male subgroup or seemed easier for males, whereas ACCT Items 29 and 30 favoured a female subgroup or appeared easier for females. There were therefore five Significant DIF items found on the ACCT and further study needs to be conducted to uncover their causes.

Table 5.6

Uniform differential item functioning of 30 ACCT items by gender

Item	DIF estimates			S.E. of estimates			Rasch-Welch			Mantel-Haenszel	
No.	Male	Female	Size	Male	Female	Joint	*t*	*d.f.*	*p*	X^2	*p*
01	-0.89	-0.19	-0.71	0.28	0.21	0.35	-2.04	151	*0.04	1.37	0.24
02	1.65	1.79	-0.14	0.37	0.25	0.45	-0.32	145	0.74	0.01	0.90
03	-1.94	-2.27	0.33	0.32	0.27	0.42	0.79	159	0.42	0.01	0.90
04	0.04	-0.44	0.48	0.28	0.21	0.35	1.37	150	0.17	1.47	0.22
05	-0.12	-0.19	0.07	0.28	0.21	0.35	0.19	151	0.84	0.02	0.88
06	-0.82	-0.66	-0.16	0.28	0.21	0.35	-0.45	152	0.65	0.99	0.31
07	0.21	-0.19	0.39	0.29	0.21	0.36	1.10	149	0.27	0.00	0.98
08	-0.43	-0.31	-0.12	0.28	0.21	0.35	-0.34	152	0.73	0.02	0.88
09	-0.58	-0.83	0.25	0.28	0.21	0.35	0.72	153	0.47	1.06	0.30
10	-0.58	-0.66	0.07	0.28	0.21	0.35	0.21	152	0.83	0.00	0.96
11	-0.04	-0.23	0.19	0.28	0.21	0.35	0.54	151	0.58	0.05	0.81
12	0.94	0.38	0.56	0.32	0.21	0.38	1.47	144	0.14	1.16	0.28
13	-2.54	-1.71	-0.82	0.38	0.24	0.44	-1.86	141	0.06	0.23	0.62
14	-0.58	-0.58	0.00	0.28	0.21	0.35	0.00	152	1.00	0.00	0.94
15	1.05	0.78	0.26	0.33	0.22	0.39	0.67	144	0.50	0.00	0.95
16	0.74	1.12	-0.38	0.31	0.22	0.38	-1.00	150	0.31	0.49	0.48
17	1.03	1.03	0.00	0.33	0.22	0.40	0.00	146	1.00	0.00	0.95
18	0.55	1.28	-0.72	0.30	0.23	0.38	-1.91	153	0.05	2.01	0.15
19	0.21	-0.06	0.26	0.29	0.21	0.36	0.74	149	0.46	3.48	0.06
20	1.15	1.23	-0.07	0.33	0.23	0.40	-0.18	146	0.85	0.24	0.62
21	0.74	1.55	-0.81	0.31	0.24	0.39	-2.07	154	*0.04	3.52	0.06
22	-0.51	-0.27	-0.24	0.28	0.21	0.35	-0.68	152	0.49	0.25	0.61
23	-0.04	0.24	-0.28	0.28	0.21	0.35	-0.81	151	0.42	0.03	0.85
24	-0.04	0.07	-0.11	0.28	0.21	0.35	-0.31	151	0.75	1.52	0.21
25	-0.74	0.11	-0.85	0.28	0.21	0.35	-2.46	152	*0.01	7.92	*0.00
26	-0.89	-1.2	0.31	0.28	0.22	0.35	0.87	155	0.38	0.12	0.71
27	-0.51	-0.27	-0.24	0.28	0.21	0.35	-0.68	152	0.49	0.67	0.41
28	1.79	1.28	0.51	0.38	0.23	0.45	1.14	138	0.25	1.00	0.31
29	1.05	0.16	0.89	0.33	0.21	0.39	2.29	142	*0.02	4.78	*0.02
30	0.04	-0.88	0.92	0.28	0.21	0.35	2.6	152	*0.01	7.97	*0.00

**p <.05*

Figure 5.9 shows a uniform DIF plot of 30 ACCT items and 193 EFL graduate examinees and it can be interpreted in conjunction with Table 5.6. The uniform-DIF plot can be used as an informative tool for informing not only potential DIF on a micro item level but also potential DTF on a macro test level. The blue line with diamond-shaped points represents the item difficulty for the female subgroup, and the red line with square-plot points represents the item difficulty for the male subgroup. The black dashed line with dot points demonstrates the average item difficulty between male and female subgroups. The points on the blue and red lines are expected to be close to points on the dashed line in order to show that a particular item is not differentially more difficult or easier for the male or female subgroups.

For most ACCT items, item difficulty difference between males and females is not sizeable except for Items 1, 21, 25, 29 and 30 which exhibited noticeable gaps. It was apparent that difficulty estimates of ACCT Items 1, 21, 25, 29 and 30 varied significantly substantially across male and female subgroups. Considering these five uniform DIF items, the ACCT appeared not to exhibit a sizeable proportion of gender-based uniform DIF. It should also be noted that significant DIF items may not necessarily undermine the test or actually indicate biased items (Boone et al., 2014; Hambleton et al., 1991) and deleting DIF items for certain subgroups does not ensure that the test would be unbiased for other subgroups since the conclusion of bias goes beyond empirical data. DIF is preliminarily used to describe empirical evidence found in an investigation of item bias (Hambleton et al., 1991).

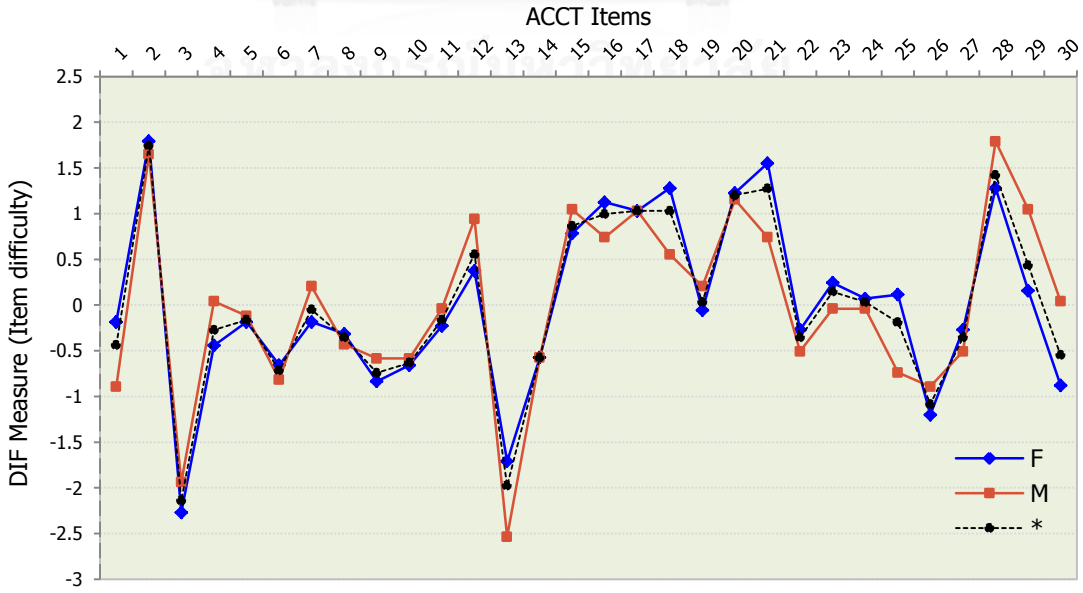

Figure 5.9. Uniform differential item functioning by gender

5.3 Correlation analysis

To provide empirical evidence supporting the assumption of the warrant underlying the explanation inference, a correlation analysis was performed to explore whether there was a strong and significant relation between ACCT scores and AVLT scores and between ACCT competence and vocabulary size knowledge. Scores on the ACCT and AVLT and person estimates on both tests were used to scrutinise the relationship between ACCT and AVLT. Figure 5.10 displays a scatterplot showing the relationship between ACCT scores and AVLT scores. The Pearson product-moment correlation indicated that there was a positive strong relationship between ACCT scores and AVLT scores (r = .74, p < .001). This indicated that students who did well on the ACCT had also done well on the AVLT.

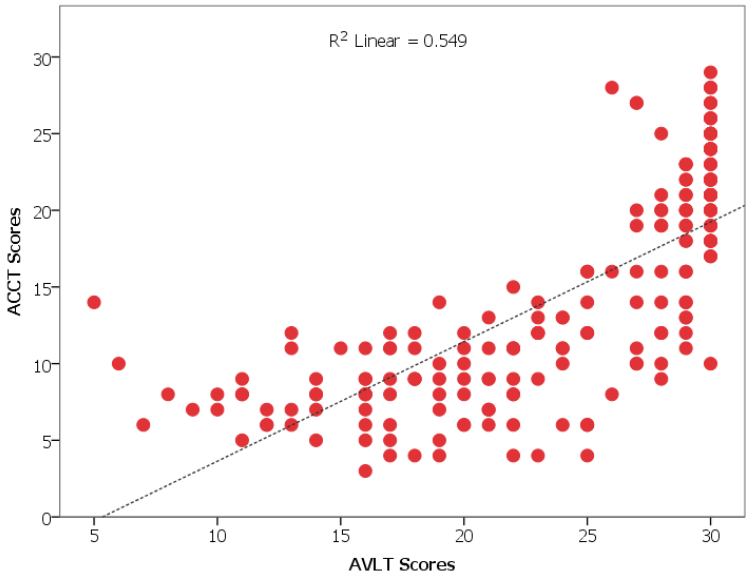

Figure 5.10. Scatterplot showing the relationship between ACCT scores and AVLT scores

Figure 5.11 shows a scatterplot showing the relationship between collocational competence measures and vocabulary size measures. The Pearson product-moment correlation revealed that there was a positive strong relationship between person collocational competence and vocabulary size knowledge (r = .79, p = .001).

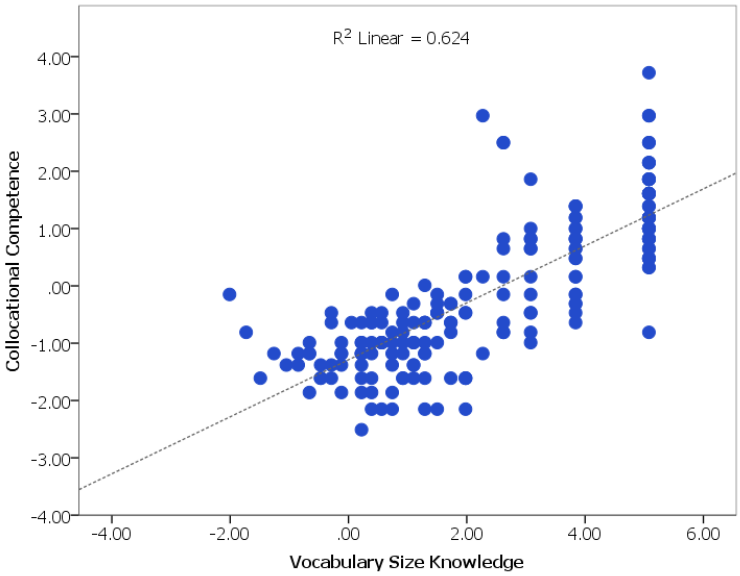

Figure 5.11. Scatterplot showing the relationship between collocational competence and vocabulary size knowledge

5.4 Analysis of variance

To provide empirical evidence in support of the extrapolation inference's assumption that the ACCT scores can distinguish among three proficiency groups of examinees, a one-way independent ANOVA was performed to test whether there were statistically significant differences in ACCT scores amongst three proficiency groups. The ACCT score was used as a dependent variable and a proficiency group was used as an independent variable. As shown in Table 5.7, a one-way independence ANOVA showed that there were statistically significant differences in ACCT scores amongst three proficiency groups, $F(2, 190) = 218.650$, $p < .001$. The Levene's test for homogeneity of variance shows that the variances are not homogeneous and the number of participants in three groups was quite unequal. Due to unequal sample sizes and unequal variances, the Games-Howell post-hoc test was used to compare the differences among subgroups as it is designed to deal particularly with such condition (Howell, 2008, 2013).

Table 5.7

Summary of descriptive statistics, homogeneity test of variance, and ANOVA

Proficiency	Descriptive Statistics			Homogeneity Test		ANOVA	
Groups	N	M	SD	Levene	*p*	*F*	*p*
Low	84	8.47	2.69	18.849	***.000	218.65	.000
Moderate	59	15.06	5.13				
High	50	22.56	3.49				
Total	193	14.13	6.85				

**p <.05. **p <.01. ***p <.001*

Table 5.8

Summary of the Games-Howell post-hoc test

					95% Confidence Interval	
Groups		Mean Difference (I-J)	Std. Error	Sig.	Lower	Upper
L	M	-6.591*	.7306	***.000	-8.33	-4.847
	H	-14.083*	.5753	***.000	-15.45	-12.711
M	L	6.591*	.7306	***.000	4.84	8.336
	H	-7.492*	.8314	***.000	-9.46	-5.514
H	L	14.083*	.5753	***.000	12.71	15.456
	M	7.492*	.8314	***.000	5.51	9.469

**p <.05. **p <.01. ***p <.001*

As summarised in Table 5.8, the Games-Howell post-hoc test indicated that the groups were significantly different from one another ($p < .001$). The ACCT scores of the high-proficiency group were significantly higher than those of the moderate-proficiency group and the ACCT scores of the moderate-proficiency group were significantly higher than those of low-proficiency students. Figure 5.12 shows a boxplot diagram showing ACCT score distributions for three proficiency levels using the dichotomous scoring scale. The three groups include high-proficiency students (n = 50), moderate-proficiency students (n = 59), and low-proficiency students (n = 84).

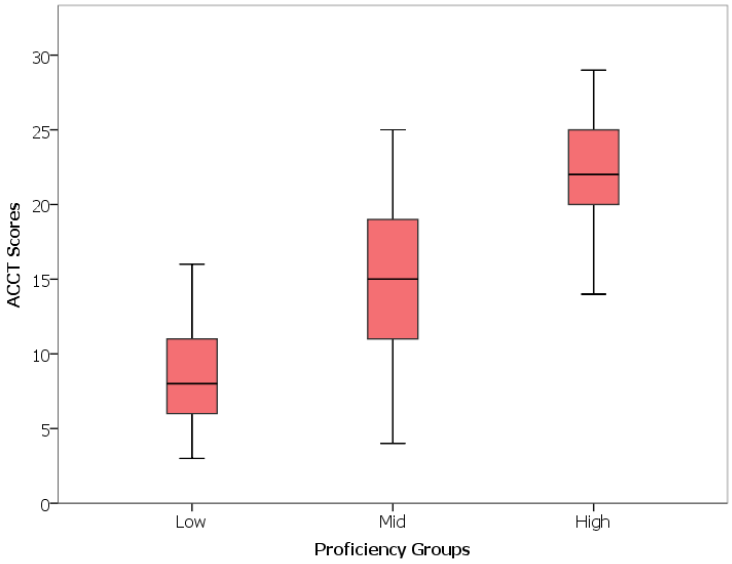

Figure 5.12. Boxplot diagram showing ACCT score distributions for three proficiency groups

5.5 Test reflection survey

The test reflection survey was intended to preliminarily elicit examinees' test-taking strategies with regard to meta-cognitive and cognitive strategies. A three-item test reflection questionnaire was adopted from Voss (2012) and also translated into Thai. It was delivered after the ACCT and AVLT. The first close-ended question asked as to whether examinees were thinking about academic English, or they were not thinking about academic, or they were not able to specify. The second close-ended question asked if examinees perceived that English texts on the ACCT was similar to academic English in university textbooks as they were responding to the ACCT items.

A chi-square test of independence was used based raw counts to examine the relation among the responses in three groups of examinees for both questions. The third open-ended question asked examinees to express their opinion regarding about similarities or differences between English on the ACCT and in university textbooks. Responses in Thai were also translated into English. Examinees responses were coded as either "able to compare" or "not able to compare." Those responses indicating the ability to compare were further coded as either "similar" or "different", or "both." Examples of examinee responses in each group were also presented for each category.

Table 5.9 shows a frequency and percentage of student responses to Question 1. All examinees responded to all options in Question 1. Of 193 examinees

responded, as many as 135 examinees (69.9%) selected a yes-option, 31 examinees (16.1%) chose a no-option, and 27 examinees (14%) ticked the last option "*I don't know.*" Figure 5.0 shows the percentage of examinees in three groups responding to three options in Question 1. As in Figure 5.13, among three options, the percentage of students choosing "yes" is the highest, meaning that the majority of examinees though that they were thinking though academic English while taking the ACCT. The chi-square test of independence revealed that the percentage of students reporting that they were thinking about academic English while taking the ACCT was significantly different among three proficiency groups, χ^2 (4, N = 193) = 10.035, p = .040. In other words, mid and high-proficiency groups were thinking about academic English more than low-proficiency group.

Table 5.9

Frequency counts and percentage of responses to test reflection survey question 1

	Responses			
Proficiency groups	Yes	No	I don't know	Totals
Low-proficiency group	49	18	17	84
Percentage within groups	*58.3%*	*21.4%*	*20.2%*	*100%*
Mid-proficiency group	47	6	6	59
Percentage within groups	*79.7%*	*10.2%*	*10.2%*	*100%*
High-proficiency group	39	7	4	50
Percentage within groups	*78.0%*	*14.0%*	*8.0%*	*100%*
Totals	135	31	27	193
Percentage within groups	*69.9%*	*16.1%*	*14.0%*	*100%*

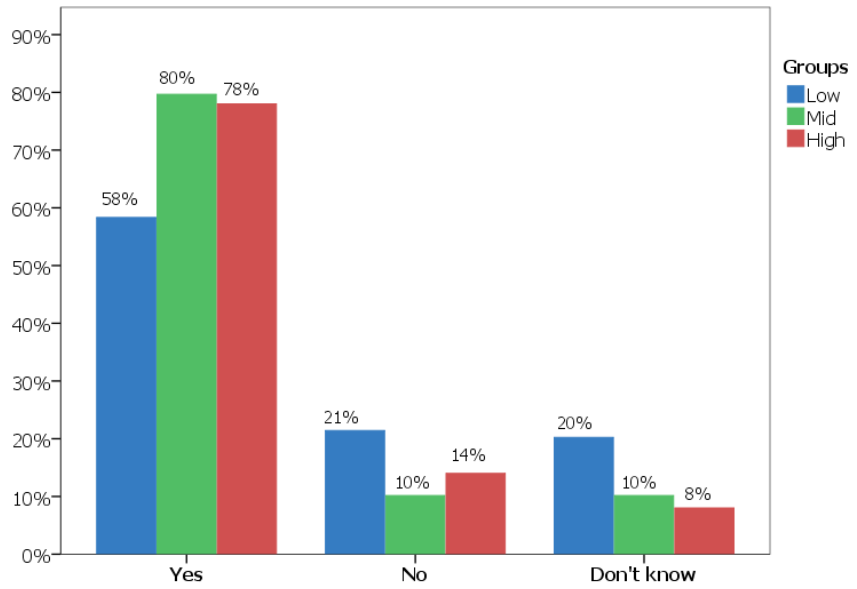

Figure 5.13. Table chart displaying the percentage of responses to test reflection survey question 1

Table 5.10 shows a frequency and percentage of student responses to Question 2. All examinees responded to all options in Question 2. Of 193 examinees responded, as many as 105 examinees (54%) chose a yes-option, 27 examinees (14%) selected a no-option, and 61 examinees (31%) ticked the last option "*I don't know.*" Figure 5.0 displayed the percentage of examinees in three groups responding to three options in Question 2. As in Figure 5.14, among three options, the percentage of students choosing "yes" is the highest, meaning that the majority of examinees though that English text on the ACCT was similar to academic English used in university textbooks. The chi-square test of independence revealed that the percentage of students reporting language in the ACCT was similar to academic English used in university textbooks was not significantly different among three proficiency groups, χ^2 (4, N = 193) = 7.365, p = .1180. This means that the majority of examinees perceived that English on the ACCT and in university textbooks were similar.

Table 5.10.

Frequency counts and percentage of responses to test reflection survey question 2

	Responses			
Proficiency groups	Yes	No	I don't know	Totals
Low-proficiency group	38	11	35	84
Percentage within groups	*45.2%*	*13.1%*	*41.7%*	*100%*
Mid-proficiency group	36	8	15	59
Percentage within groups	*61.0%*	*13.6%*	*25.4%*	*100%*
High-proficiency group	31	8	11	50
Percentage within groups	*62.0%*	*16.0%*	*22.0%*	*100%*
Totals	105	27	61	193
Percentage within groups	*54.4%*	*14.0%*	*31.6%*	*100%*

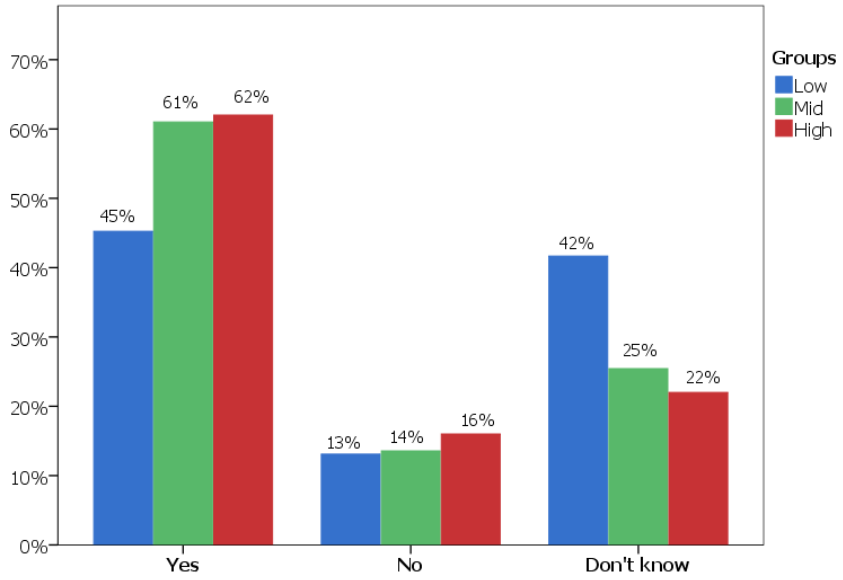

Figure 5.14. Table chart displaying the percentage of responses to test reflection survey question 2

Table 5.11 shows a frequency and percentage of student responses to Question 3. Of 193 examinees, 120 examinees (62%) responded to Question 3 while 73 examinees (37%) did not respond to Question 3. Responses to Question 3 provided the comments or reasons why test-takers perceived that the language on the ACCT was similar or different from that used in university textbook. Since some responses were written in Thai, an English translation version was also provided for each response in Thai.

Table 5.11.

Frequency and percentage of responses to test reflection survey question 3

	Responses		
Proficiency groups	Responded	Not responded	Totals
Low-proficiency group	42	42	84
Percentage within groups	*50.0%*	*50.0%*	*100%*
Mid-proficiency group	44	15	59
Percentage within groups	*74.6%*	*25.4%*	*100%*
High-proficiency group	34	16	50
Percentage within groups	68.0%	32.0%	100%
Totals	120	73	193
Percentage within groups	*62.2%*	*37.8%*	*100%*

Table 5.12 and Figure 5.15 present the percentage of examinees who were able and unable to compare and contrast English texts on the ACCT and in university textbooks. Overall, the majority of each proficiency group was able to compare and contrast the texts on the ACCT and in university textbooks. The highest percentage (89%) of high-proficiency group provided responses indicating the ability to compare and contrast the texts. Low-proficiency responses accounts for 71 per cent of low-proficiency group, whereas the mid-proficiency group provided the lowest percentage (65%) of responses. This at least indicates that most of examinees were familiar with academic English. High-proficiency examinees were mostly familiar with academic English and thus they may be more exposed to academic English.

Table 5.12

Frequency and percentage of responses that are able and unable to compare texts

	Responses		
Proficiency group	Able to compare	Unable to compare	Total
Low-proficiency group	30	12	42
Percentage within groups	*71.4%*	*28.6%*	*100%*
Mid-proficiency group	39	5	44
Percentage within groups	*88.6%*	*11.4%*	*100%*
High-proficiency group	22	12	34
Percentage within groups	*64.7%*	*35.3%*	*100.0%*
Totals	91	29	120
Percentage within groups	*75.8%*	*24.2%*	*100%*

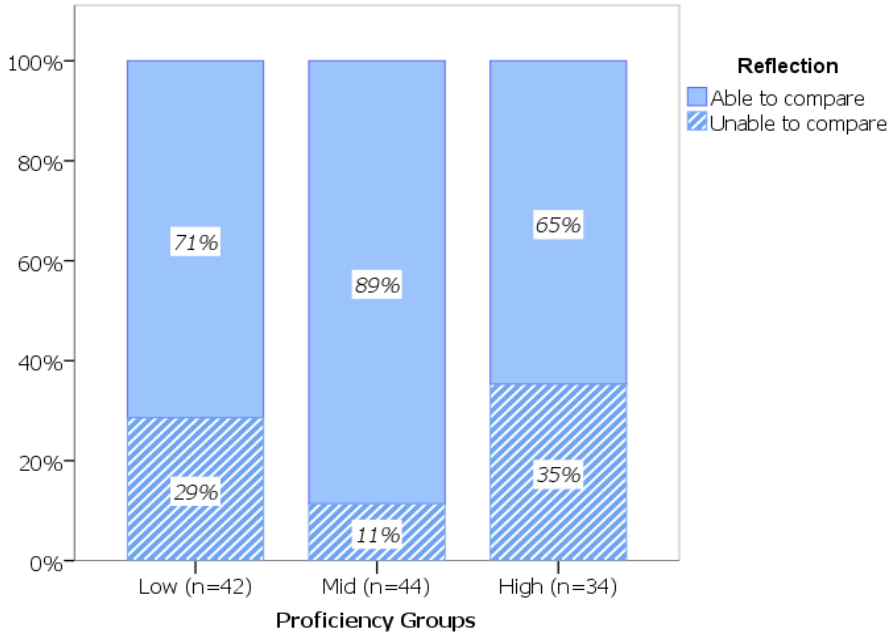

Figure 5.15. Table chart displaying the percentage of responses that are able and unable to compare texts

Table 5.13 and Figure 5.16 present the percentage of test-takers who were able to compare and contrast the English on the ACCT and in university textbooks. Interestingly, perception was quite different amongst three-proficiency groups. The majority of high-proficiency examinees (27%) and mid-proficiency examinees (27%) expressed that the texts on the ACCT and textbooks were similar. The largest percentage (50%) of responses indicating differences were found in the low-proficiency group, followed by the mid-proficiency group (40%) and the high-proficiency group (10%). The highest percentage (50%) of responses demonstrating both similarity and difference was found in the low-proficiency group, followed by the mid-proficiency group (38%) and the high-proficiency group (13%).

Table 5.13

Frequency and percentage of responses that are able to compare texts as similar different or both

Proficiency group	Responses			
	Similar	Different	Both	Total
Low-proficiency group	21	5	4	30
Percentage within groups	*70.0%*	*16.7%*	*13.3%*	*100%*
Mid-proficiency group	32	4	3	39
Percentage within groups	82.1%	10.3%	7.7%	100%
High-proficiency group	20	1	1	22
Percentage within groups	*90.9%*	*4.5%*	*4.5%*	*100%*
Totals	73	10	8	91
Percentage within groups	*80.2%*	*11.0%*	*8.8%*	*100%*

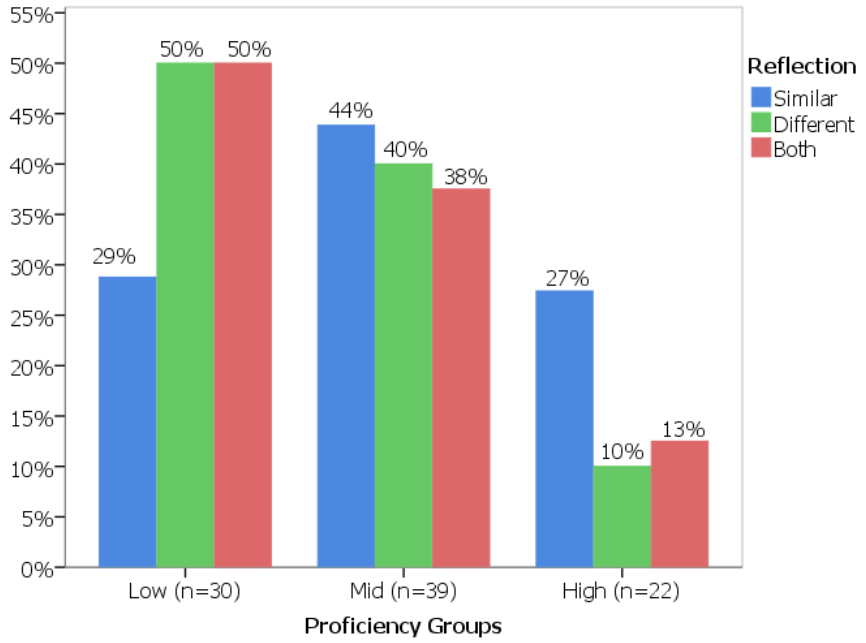

Figure 5.16. Table chart displaying the percentage of responses that are able to compare texts as similar different or both

Qualitative responses to Question 3 from three proficiency groups were coded to explore why examinees though the texts in both sources were similar or different or both. Responses from test-takers who though English on the test and in textbooks was similar were categorised into three main categories: (1) language features and use, (2) content context and discipline, and (3) textbooks and other

academic sources. Each category also has sub-categories. In terms of language features and use, examinees perceived that the texts on the ACCT were similar to those in textbooks in terms of sentence structure, words, and use style. Some respondents though that sentences on the ACCT were similar to those in textbooks in that the sentences were complex and long. Others expressed that words on the ACCT were similar to those in textbooks in the sense that words were specific to various fields of study. Still others though that the style of language on the ACCT was similar to that in textbooks in the sense that the style was formal and academic. Table 5.14 presents examples of responses indicating similarity in language features and use.

Table 5.14

Example of responses indicating similarity in academic sources

Proficiency groups	Responses
1. High-competency:	*ใช้ภาษาที่เป็นทางการและคำศัพท์วิชาการ มีการใช้ภาษาเขียนที่สูงและเข้าใจยากกว่างานเขียนทั่วไปซึ่งอาจเป็นเพราะบทความหรือ text ที่เขียนใน professional หรือ นักเรียนที่มีความรู้ในศาสตร์นั้นๆพอสมควร* *The language is formal and contains academic words. The language is written in a complicated way that it is more difficult to understand, compared with general written English. This may be because the texts are used in journals in specific disciplines that require background knowledge to better comprehend.*
2. High-competency:	*All words are often seen or appeared in the textbooks. If students have an opportunity to read a lot of textbooks in various fields, they will get familiar with these words even though they might not know or notice their meanings or the way to make sentence or match with others. So my answer is that test containing the similar sentence structure or words to the textbooks in the university.*
3. Mid-competency:	*ภาษาอังกฤษในแบบสอบนี้มีความคล้ายคลึงกับตำราวิชาการบางเล่ม ตรงที่ใช้ภาษาอย่างเป็นทางการ ใช้ศัพท์ที่ยาก ชั้นสูงแบบที่ไม่ค่อยเจอในหนังสือทั่วไป* *English on the test is similar to that in some textbooks in that the language is forma land has difficult words which are not commonly found in general books.*

Table 5.14

Example of responses indicating similarity in academic sources

Proficiency groups	Responses
4. Mid-competency:	*English on the ACCT is similar to English in university textbooks in terms of structure and vocabulary as university textbooks might be written by using academic collocations to describe academic information*
5. Mid-competency:	*ใช้ภาษาที่เป็นทางการ มีรูปแบบโครงสร้างค่อนข้างเฉพาะ เช่น passive voice คำศัพท์ที่ใช้มีรูปแบบตายตัว เช่น conduct research* *The language is formal and has specific grammatical structure, such as passive voice. Words are strictly used together, such as "conduct research."*
6. Mid-competency:	*The English in the test is similar to the English in university textbooks because those words used relates to academic issues. English in the test can be seen in many university textbooks. Those words used in the test are quite similar. It is not too complicated to remember or recognise. However, issue relating to collocation are still quite challengeable to test- takers*
7. Low-competency:	*คำศัพท์บางคำคล้ายกับคำศัพท์ในงานวิจัยต่างประเทศ* *Some words are similar to words in international journals.*
8. Low-competency:	*ค่อนข้างคล้ายกับที่เรียนมาบ้างแต่บางคำก็ไม่คุ้นเคย* *The language is quite similar but some words are not familiar.*

In respect of content context and discipline, examinees perceived that the texts on the ACCT are similar to those in textbooks in terms of the content, context, and discipline. Examinees perceived that content context and discipline are specific and varied. They perceived that some words are familiar to them but some are not, meaning that those words are frequently used in specific fields. These responses are related to those related to specific words. Table 5.15 shows example of responses indicating similarity in terms of content context and discipline.

Table 5.15

Example of responses indicating similarity in content, context, and discipline

Proficiency groups	Responses
1. High-competency:	*มีความคล้ายคลึงกันในแง่ของระดับภาษาวิชาการ ซึ่งมักพบตามตำราเรียนระดับมหาวิทยาลัย งานวิจัยต่าง ๆ สังเกตุได้จากวิธีเรียบเรียงประโยค เนื้อหาที่มีลักษณะเฉพาะด้าน และคำศัพท์เฉพาะทางในด้านต่างๆ เช่นด้านวิทยาศาสตร์ กฎหมาย การเมือง ซึ่งคิดว่าคำเหล่านี้หรือวิธีเขียนแบบนี้มักพบในตำราวิชาการบ่อยครั้ง* *The language is similar in terms of academic style, which is commonly found in university textbooks and research articles. This can be observed from sentence arrangement, specific content, and specific words in many fields such as sciences, law, and politics. These sentence patterns and words are commonly found in academic textbooks.*
2. High-competency:	*เจอในพวกบทความเฉพาะทาง เช่น กฎหมาย ซึ่งเป็นสำนวนที่ใช้เฉพาะด้าน* *It was found in journals from specific fields such as law. The expressions are used in specific fields.*
3. High-competency:	*It depends on which field the textbooks focus on. In the English field, most of the words in the test can be found in the textbooks but in other fields I don't think they are covered.*
4. Mid-competency:	*บางศัพท์สามารถพบได้บ่อยๆ แต่บางคำใน test นั้นคล้ายๆคำศัพท์เฉพาะด้านที่บางสาขาอาจได้เรียน พบได้ตาม textbooks ของเอกต่างๆ แตกต่างกันไป* *Some words are frequently found but some are used in specific fields and found in textbooks from different disciplines.*

In terms of textbooks and other academic, examinees perceived that not only are the texts on the ACCT similar to those in textbooks, but the texts are also similar to those used in other academic sources, such as articles, journals, teaching materials, and documents. Table 5.16 provides examples of responses indicating similarity in academic sources.

Table 5.16

Example of responses indicating similarity in academic sources

Proficiency groups	Responses
1. Mid-competency:	*คล้ายกับหนังสือเรียนและเอกสารงานวิจัยที่ได้อ่านโดยจะใช้คำศัพท์ที่ชัดเจนเข้าใจง่าย* *It is similar to textbook and research articles I read. The words are clear and easy to understand.*
2. Low-competency:	*คล้ายกันกับในงานวิจัยที่อ่านวารสารวิทยาศาสตร์บางคำศัพท์* *Some are similar to scientific journals.*
3. Low-competency:	*คำศัพท์ที่พบปรากฏอยู่ในเอกสารต่างๆ ทั้งตำราเรียนหรืองานวิจัย* *Words are found in documents, textbooks and research articles.*

As for examinees who perceived that the texts on the ACCT are different from those in textbooks, their responses can be categorised into two major categories: (1) language features and use and (2) content context and discipline, and each category also has sub-categories. With regard to language features and use, examinees expressed that the texts on the ACCT were more formal and the texts in textbooks were more familiar than those on the ACCT. Words on the ACCT were more varying and unfamiliar. Table 5.17 shows examples of responses demonstrating difference in language features and use.

Table 5.17

Examples of responses demonstrating difference in language features and use

Proficiency groups	Responses
1. High-competency:	*More formal*
2. Mid-competency:	*เป็นคำที่ปรากฏไม่บ่อยนักในตำราเรียน ในแบบสอบมีคำศัพท์ที่แปลกใหม่หลากหลายกว่า คำหลักๆ ที่ใช้เกินการเรียนรู้* *Words are not often found in textbooks. Words on the test are new, varied, and not necessary to learn.*
3. Low-competency:	*ต่างกัน ในตำราเรียนคำศัพท์ที่ใช้จะเป็นคำศัพท์ที่รู้จักมากกว่า สามารถเดาคำศัพท์ได้ง่ายกว่า* *It is different. In textbooks, words are more familiar and easier to guess meaning.*
4. Low-competency:	*ต่างกันเพราะศัพท์วิชาการเป็นศัพท์เฉพาะไม่ค่อยได้เจอในตำราเรียน*

Table 5.17

Examples of responses demonstrating difference in language features and use

Proficiency groups	Responses
	It is different. Academic words are specific and not often found in textbooks.

In terms of content context and discipline, some respondent though the texts on the ACCT were more specific and meaning of words were different. Table 5.18 presents examples of responses indicating difference in terms of content context and discipline.

Table 5.18

Examples of responses indicating difference in content, context, and discipline

Proficiency groups	Responses
1. Mid-competency:	*ในตำราเรียนไม่ specific เท่าในข้อสอบ* *English in textbooks is not as specific as English on the test.*
2. Mid-competency:	*แบบสอบน่าจะมีความจำเฉพาะเจาะจงและใช้ในบริบทงานค้นคว้าวิจัยเชิงวิชาการมากกว่าแบบคำศัพท์เชิงวิชาการครับผม* *The test requires specific memory and context related to academic research rather than academic words.*
3. Low-competency:	*ต่างกันที่ความหมายในการนำมาใช้ คำที่มาใช้ในรูปแบบประโยคอื่นๆ ทำให้ความหมายที่ได้ไม่ตรงกับตำรา* *It is different in terms of language use. When words are used in different sentences their meaning is different from that in textbooks.*

As for examinees who perceived that the texts on the ACCT are different from those in textbooks, their responses can be categorised into two major categories: (1) language features and use and (2) content context and discipline, and each category also has sub-categories. With respect to language features and use, students expressed that they had to use grammatical knowledge and other content knowledge in order to answer the ACCT. Some perceived that texts on the ACCT were not as formal as those in textbooks and some words were similar and some were not. Table 5.19 presents examples of responses indicating similarity and difference in language features and use.

Table 5.19

Examples of responses indicating similarity and difference in language features and use.

Proficiency groups	Responses
1. Mid-competency:	*ต่างกันในบางส่วนที่จะต้องใช้ความรู้ด้านไวยากรณ์ภาษาอังกฤษและบางส่วนจะต้องใช้ความรู้ด้านอื่นๆ ซึ่งเป็นไปตามประสบการณ์ของแต่ละบุคคลประกอบด้วย* *It is different at some point in that the test requires knowledge of grammar, different content knowledge, and personal experience.*
2. Mid-competency:	*น่าจะคล้าย เพียงแต่จะไม่เป็นวิชาการมากเกินไป มีการเขียนให้เข้าใจง่าย ใช้ได้ในชีวิตประจำวัน* *It may be similar but texts on the test are not too academic, easy to understand and used in daily life.*
3. Low-competency:	*คล้ายบ้างในบางคำแต่ส่วนใหญ่ไม่คล้าย* *Some are similar but most are not.*

Some respondent though words were related to different contents and fields and the texts were similar but the content were different. Table 5.20 presents examples of responses indicating similarity and difference in content context and discipline.

Table 5.20

Examples of responses indicating similarity and difference in content, context, and discipline.

Proficiency group	Responses
1. High-competency:	*ภาษาอังกฤษในแบบสอบเหมือนกับภาษาอังกฤษที่อ่านจากหนังสือทั่วไปในแต่ละวิชาชีพ แต่ไม่เหมือนกับภาษาที่ใช้ในตำราเรียน แต่จริงๆ แล้วค่อนข้างตัดสินใจยากว่าจะมาจากตำราเรียนได้หรือไม่เพราะแต่ละคำถามมีมา 1 ประโยค หากมีประโยคเสริมข้างหน้าหรือหลังอาจทำให้ตัดสินใจได้ชัดเจนมากขึ้น* *English on the test is similar to English in general books in different fields but is different from English in textbooks. In fact, it is difficult to judge if English on the test is excerpted from textbooks because each question has only one*

Table 5.20

Examples of responses indicating similarity and difference in content, context, and discipline.

Proficiency group	Responses
	sentence. If more sentences are provided, it is clearer to decide.
2. Mid-competency:	*มีคำศัพท์ที่มีความหมายเหมือนกันแต่ใช้แตกต่างกันในแต่ละบริบทถ้าไม่ใช้คำศัพท์ที่เข้ากับบริบทก็จะให้ความหมายไม่ชัดเจน* *Some words have similar meaning but are used in different contexts. If words are used in inappropriate contexts, their meaning is not clear.*
3. Low-competency:	*คิดว่าน่าจะมีบริบทในการใช้ที่แตกต่างกันแต่สามารถใช้ร่วมกันได้ในบางกรณี* *I think the context of language use is different but in some cases the language on the test and in textbooks can be used in a similar way.*
4. Low-competency:	*ภาษาอังกฤษในแบบสอบคล้ายกับภาษาอังกฤษเชิงวิชาการที่ใช้ในตำราเรียนในมหาวิทยาลัย แต่เนื้อหาคนละอย่าง* *English on the test is similar to that in textbooks used in university but the content is different.*

5.6 Cut-score establishment

In this study, cut-scores were established following a contrasting-groups method (Livingston & Zieky, 1982). Cut scores were set for two main decisions. The primary decision is to place students into three competency levels and the secodnary decision is to screen examinees as pass or fail, or as remedy or non-remedy or other appropriate binary decisions. In this study, ACCT scores and collocational competency logits (henceforth referred to as theta) estimated by Winsteps were used to establish two sets of thresholds for placement and screening purposes. Cut scores are values or thresholds that demarcate the pass or failure, or competency levels.

In order to locate the cut scores for classifying examinees into low, mid and high-competency levels, Frequency distributions of the ACCT theta and the ACCT scores were generated for each of the prior three proficiency groups and then three trendlines (similar to normal curves) were plotted for three frequency distributions

using Microsoft Excel 2010. The intersection between trendlines of frequency distributions of mid and high-proficiency groups was demarcated by the black dashed line to determine the first cut score between low and mid-competency levels. The intersection between trendlines of frequency distributions of mid and high-proficiency groups was demarcated by the black dashed line to determine the second cut score between mid and high-competency levels.

Figure 5.17 present three theta-based frequency distributions for three proficiency groups. The trendlines of low and mid-proficiency groups were intersected at approximately 0.0 in competency logit scale and thus I decided to use a competency logit of 0.0 as the first cut score between low and mid competency levels. Not only is the first cut score intended primarily to place students into low and mid-competency levels, it was used additionally as the cut scores for screening as to which students should or should not take more English courses. Therefore, the first cut score is of critical threshold for the use of ACCT scores for screening decision.

To separate mid-competency students from high-competency students, the second cut score was determined at the point where the trendlines of mid and high-proficiency distributions were intersected. The intersection between the trendlines of mid and high-proficiency group distributions was very nearly at 0.9 on the competency logit scale and thereby I used a 0.9 logit as the second cut score for classifying examinees into mid and high-competency levels. The person-item variable map in Figure 5.19 was also used to give visual information on the theta-based cut scores and three competency bands.

The process of setting the cut-scores using the ACCT scores was exactly the sample as the ACCT theta-based cut-score setting process. The first cut score for classifying examinees into low and mid-competency groups was located at 14 on the ACCT score scale and the second cut score for classifying examinees into mid and high-competency groups was established at 19 on the ACCT score scale. After two set of cut scores were determined, they were further investigated to see to what extent these sets of cut scores accurately classified examinees into the competency levels that they were expected to be.

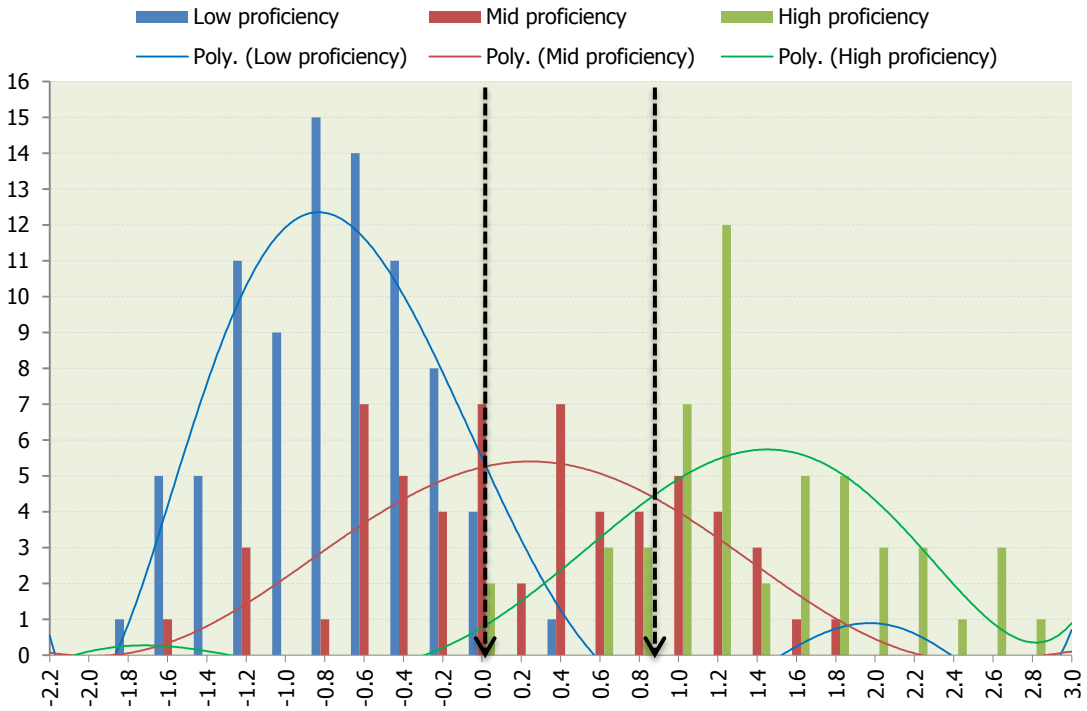

Figure 5.17. Intersected trendlines of three proficiency group distributions of collocational competence estimates

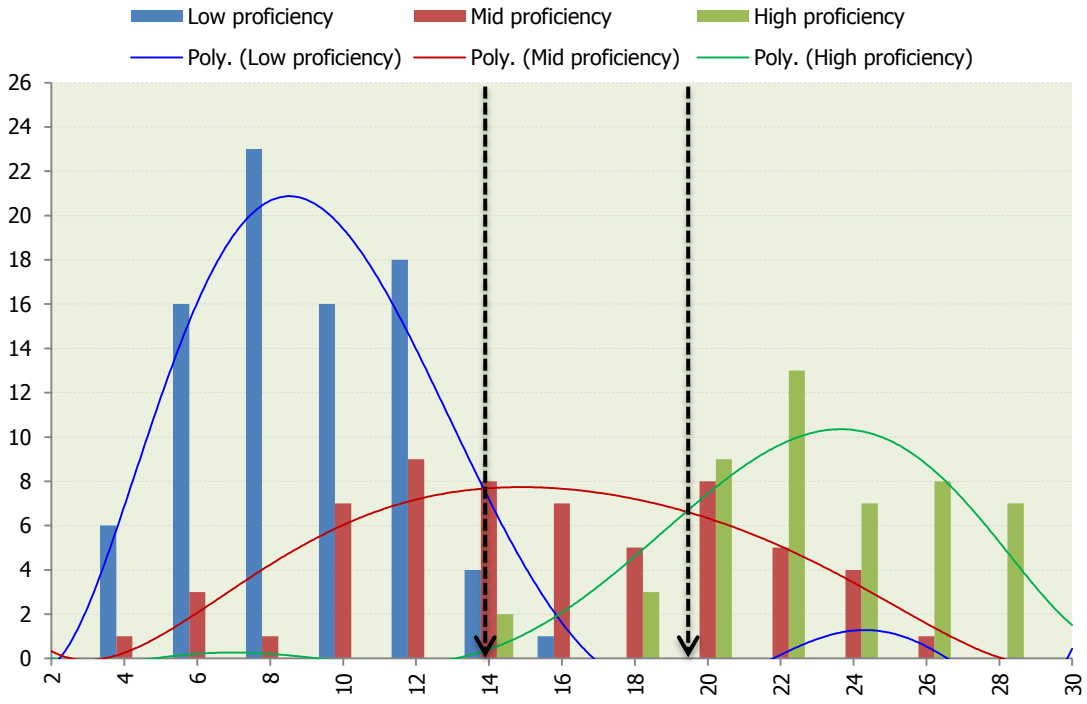

Figure 5.18. Intersected trendlines of three proficiency group distributions of the ACCT scores

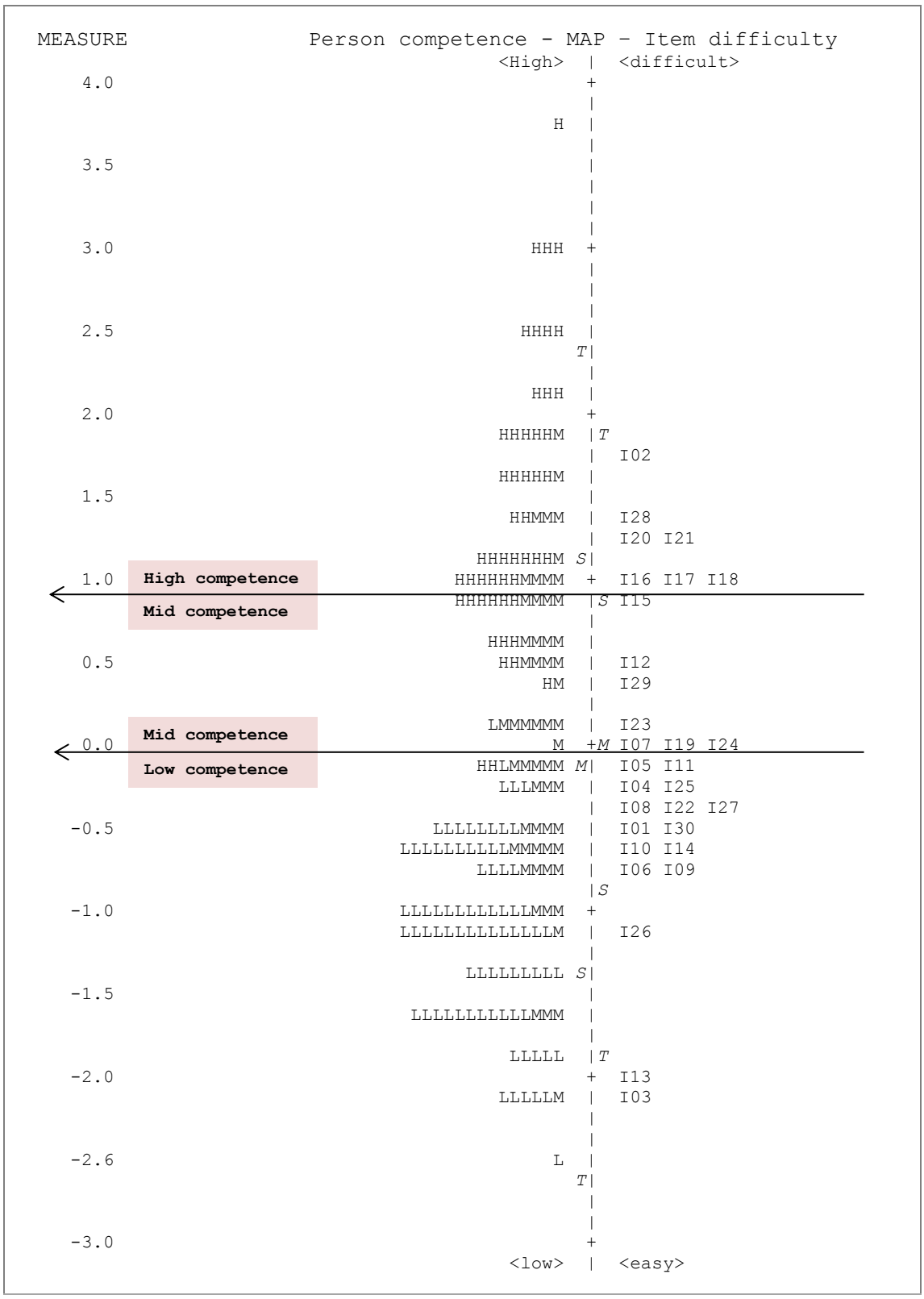

Figure 5.19. Person-item variable map showing the two cut scores for classifying three competency bands

It cannot be absolutely assumed that EFL graduate samples in this study would be placed perfectly accurately, yet it seems reasonable to assume that the three pre-classified proficiency groups in this sample would to a logical extent represent different competency levels in relation to the TLU construct of collocational competence. It is, thereby, interesting and informative to ascertain as to what extent the ACCT scores and collocational competence logits (theta) would classify or place EFL graduate test-takers into the pre-classified proficiency groups. Based on the contrasting-group cut score setting, the first cut-score thus was set at 14 on the ACCT and at 0.0 on a competency logit for screening or placing examinees into graduate programmes in university or other high-education settings with some appropriate English instruction. The second cut-score was set at 19 on the ACCT score and at 0.9 on a competency logit for screening or placing examinees into doctoral or English-medium international programmes in university or other high-education settings with some appropriate English instruction without necessarily requiring additional English instruction.

Table 5.21 illustrate the cut scores and descriptions for each competency level. EFL graduate examinees in the low-competency group would be expected to obtain scores below 14 or theta below 0.0. Test-taker performance at this level should indicate that they are not ready for their graduate studies in university or other high-education settings. Alternatively, they might be accepted to their graduate studies on the condition that they must take and pass appropriate English courses before graduation. EFL graduate examinees in the mid-competency group would be expected to receive scores between 14 and 18 or theta between 0.0 and 0.89. Test-taker performance at this level should indicate their readiness for master studies but would benefit from more suitable English courses while doing their studies in university or other high-education settings. EFL graduate examinees in the high-competency group would be expected to obtain scores between 19 and 30 or theta between 0.9 and above. Test-taker performance at this level should indicate their readiness for their doctoral studies or English-medium or international programmes in university or other high-education settings without necessarily taking additional English courses.

It should be noted that the description of competency levels in this study are intended to serve as a preliminary guideline for facilitating the interpretation of the use of the ACCT scores. More studies need to be conducted to find more evidence to elaborate and support the more appropriate description of competency bands for the ACCT.

Table 5.21

Cut scores and descriptions for each competency level

Cut scores	Cut theta	Competency levels	General descriptions
19 — 30	0.9 — 4.0	High-competency learners	EFL graduate test-takers in the high-competency level would be expected to obtain scores between 19 and 30 or theta between 0.9 and above. Performance at this level should indicate readiness for doctoral studies or English-medium international programmes in university or other high-education settings without necessarily taking additional English courses.
14 — 18	0.0 — .89	Mid-competency learners	EFL graduate test-takers in the mid-competency level would be expected to receive scores between 14 and 18 or theta between 0.0 and 0.89. Performance at this level should indicate readiness for master studies but would optionally need some suitable English courses while doing their studies in university or other high-education settings.
0 — 13	-4.0 — -0.1	Low-competency learners	EFL graduate test-takers in the low-competency level would be expected to obtain scores below 14 or theta below 0.0. Performance at this level should indicate that test-takers are not ready for graduate studies in university or other high-education settings. Alternatively, they might be accepted to graduate studies on but are required to take and pass appropriate English courses before graduation.

5.7 Classification error estimation

Classification accuracy is the degree to which a cut score can accurately classify examinees into different competency levels. Any misclassified examinee signifies a classification error. There are two types of false classification errors, a false positive error and a false negative error. A false positive error exists when an examinee is classified into a competency level higher than his or her true competency, while a false negative error, on the other hand, occurs when an examinee is classified into a competency level lower than his or her true competency (Cizek & Bunch, 2007). The classification error estimation in this study is based on Livingston approach and a Bayesian approach. As presented earlier, the cut scores were established at 0.0 and 0.9 logits on the competency scale and at 14 and 19 scores on the ACCT score scale. The 0.0 cut score was used to facilitate decision on screening students as pass or fail and as low or mid-competency levels, whereas the 0.9 cut score was used to facilitate decision on placing students into mid or high-competency levels.

For a Bayesian-based classification approach, the data were students' collocational competencies (theta) estimated using WinBUGS, which uses a Bayesian Rasch estimation method. WinBUGS estimated 1,000 competence thetas for each of the 193 EFL graduate examinees and then calculated the mean theta for each student to represent the actual ability estimate of each person. The mean theta of each student was then compared with two sets of theta-based cut scores (0.0 and 0.9) in order to determine which competency group he or she belongs to. After the competency level of each person was identified through comparison between the mean theta and cut scores, 1,000 thetas estimated for each person were compared with the cut scores to see the proportion of the thetas that are consistent or inconsistent with the competency level of each person as identified initially.

The proportion of misclassified thetas of all persons in each competency group was calculated as the percentage of classification error rate for a particular cut score, whereas the proportion of correctly classified thetas of all persons in each competency group was calculated as the percentage of classification consistency rate for a particular cut score. For example, if the cut score for passing a test is 0.0 and the mean theta of Examinee A is 0.1, Examinee A is classified as passing or as a low-competency examinee. Then, 1,000 theta of Examinee A were compared with the cut score 0.0 to see how many thetas out of 1,000 correctly or incorrectly classified examinee A into the expected low-competency level. The total proportion of correct and incorrect classification for each competency level was calculated as

the percentage of classification consistency and error respectively for the established cut scores. Therefore, the mean theta in a Bayesian approach was first used to determine the expected competency levels of examinees and 1,000 thetas estimated for each examinee were used to estimate the degree of classification consistency and error of each examinee for the established cut scores. Finally, the total proportion of correct and incorrect classification in each competency level was calculated as the percentage of consistent classification and error or false classification respectively for the located cut scores.

Table 5.22 shows classification accuracy and error based on a Bayesian approach. Two theta cut scores (0.0 and 0.9) were used to determine how consistently these cut scores classified examinees into three competency levels. Overall, by using two theta cut scores, most students were classified into their true competency level. In particular, low and high competency students were highly consistently classified using this set of cut scores, with classification consistency rate of 90% and 83.6% respectively. However, approximately 64% of mid-competency students were consistently classified using this set of cut scores as much as 24% of the true mid-competency students was positively misclassified into a high-competency level, while about 10% was negatively misplaced into a low-competency level.

Table 5.22

Classification accuracy and error for theta-based cut-scores using a Bayesian method

	Expected competency levels			
True proficiency groups	Low (< 0.0)	Mid (0.0 – 0.8)	High (≥ 0.9)	Consistency
Low (113)	90.1%	9.8%	0.1%	79.43
Mid (34)	10.9%	64.6%	24.5%	
High (46)	0.2%	16.1%	83.6%	

As for Livingston and Zieky's approach, a set of ACCT cut scores (14 and 19) and a set of theta-based cut scores (0.0 and 09) were all used to investigate to what extent these two sets of cut scores are accurate or erroneous in classifying examinees. The proportion of new expected competency groups was compared with that of initially classified proficiency groups to see the degree of correspondence and non-correspondence between the two classifications. The percentage of

correspondence indicates the degree of cut score classification accuracy, whereas the percentage of non-correspondence reflects the degree of cut score classification error. It should also be reminded that 193 EFL graduate students were grouped based on CU-TEP, TOEFL iBT, and IELTS scores they reported and therefore it cannot be assumed that students would be placed very accurately. However, the ACCT is expected to provide scores and competence estimates which distinguish students with different competency levels as they were groups initially.

Table 5.23 shows classification accuracy based on the ACCT scores. Overall, a set of the cut scores (14, 19) produced about 75% of classification accuracy and 25% of false classification error. This set of cut scores accurately classified low and high competency examinees, accounting for 98% and 90% of classification accuracy respectively. Mid-competency examinees were not accurately classified using this set of cut scores as a percentage of classification accuracy was as low as 29% and classification error was as much as 71%.

Table 5.23

Classification accuracy and error for score-based cut-scores

True proficiency groups	Expected competency			Error			
	Low (0-13)	Mid (14-18)	High (19-30)	False negative	False positive	Total	Accuracy
Low (n=84)	82	2	0	0%	2%	2%	98%
Mid (n=59)	24	17	18	40%	30%	71%	29%
High (n=50)	0	5	45	10%	0%	10%	90%
Total (n=193)				15%	10%	25%	75%

With regard to Rasch-based logit cut scores, Table 5.24 reveals that overall a set of logit cut scores (0.0, 0.9) yielded approximately 74% of classification accuracy and 26% of classification error, which was quite similar to ACCT cut-score classification. This set of cut scores most accurately classified or place low-competency-students, accounting for 99% of classification accuracy. There was only a 1% chance that this set of cut scores might misplace true low-competency students into a mid-competency level. The high-competency group was classified as accurately as 72% with a negative false classification error of 28%. As with ACCT cut-score classification, mid-competency students were not accurately classified using this set of cut scores as the accuracy percentage was only 35%. Up to 65% of false

classification error was computed for the mid-competency group, meaning that 39 (65%) students in a mid-proficiency group were misplaced; that is, 29 (49%) students were negatively misplaced into a low-competency level, while 10 (16%) students were positively misplaced into a high-competency level.

Table 5.24

Classification accuracy and error for theta-based cut-scores

Proficiency groups	Expected competency			Error			Accuracy
	Low (< 0)	Mid (0-0.8)	High (≥ 0.9)	False negative	False positive	Total	
Low (n=84)	83	1	0	0%	1%	1%	99%
Mid (n=59)	29	20	10	49%	16%	65%	35%
High (n=50)	2	12	36	28%	0%	28%	72%
Total (n=193)				21%	5%	26%	74%

All things considered, the rate of classification accuracy and error for score-based cut-scores and theta-based cuts-cores was not significantly different. These sets of cut scores accurately classified examinees into low and high-competency groups. However, mid-competency students were most erroneously classified using both sets of cut scores. Based on Livingston and Zieky's approach, the largest number of examinees in low and high-competency groups corresponds with initial low and high-proficiency groups. However, high non-correspondence was found between the mid-competency group and the initial low-proficiency group. By using these sets of cut scores for the mid-competency group, there were about 65% and 71% that examinees would be misplaced into competency levels higher or lower than the actual mid-competency level. However, it is up to the test users to adjust the cut scores that can classify test-takers as accurately and consistently as possible for particular purposes and decisions.

5.8 Chapter summary

To conclude, this chapter presented the results and discussion related primarily to quantitative findings and secondarily to qualitative findings. Both types of findings were based on several analyses of empirical data obtained from student responses on the ACCT, AVLT and the test reflection questionnaire. Most of empirical data analyses in this chapter are based on several applications of a Rasch

measurement model. The Rasch measurement analysis provided sound and sufficient sources of evidence in support of the assumptions in the interpretive argument. Results obtained from descriptive statistics, analysis of variance, correlation analysis, test reflection survey, cut score study, and classification error analysis also well supported the assumptions in the interpretive argument. All the results served as empirical evidence in support of the assumption underlying the warrant of the inferences in the interpretive argument initially stated.

In the next chapter that follows, I present the conclusion which begins with a brief overview of what has been presented from chapter 1 to chapter 5. Following this is a presentation of the evaluation of the evidence collected to support the interpretive argument which contributes to a lesser or greater degree to the construction of the validity argument of the ACCT in the second stage of the argument-based approach. Building a validity argument for the ACCT is indeed at the heart of chapter 6. Before ending the chapter, concise answers to research questions are presented and the implications of the current study are proposed thereafter. Chapter 6 ends with a discussion of caveats of this study and recommendations for future research.

CHAPTER 6
CONCLUSION

The primary purpose of the present study was to apply the argument-based approach (Kane, 1992, 2006, 2011, 2013) to serve as the framework for developing and validating the ACCT for EFL graduate students. The argument-based approach consists of two argument building stages, the interpretive argument development and validity argument stages. The development of the ACCT interpretive argument built on the TOEFL interpretive argument framework (Carol A Chapelle et al., 2008) and Voss (2012)'s interpretive argument framework. The ACCT was developed as a norm-referenced measure of academic collocational competence aimed specifically to facilitate decisions on screening EFL graduate students or them into proper academic English courses in university or other higher-education institutions. High-frequency verb-noun collocations were manually selected using a corpus-based approach and all test materials were obtained from BNC which is representative of the TLU of academic written English of interest. A corpus-based analysis was carried out through Lancaster BNCweb Server. High-frequency nouns were identified first and then high-frequency verbs that collocate with those nouns were selected to form pairs of restricted verb-noun collocations.

Test items were developed using a best-answer five-option multiple-choice format and were marked using a dichotomous scoring method. The key for the dichotomous scoring method was based on verbs that collocate with nouns in pairs of restricted verb-noun collocations. Participants also took the Academic Vocabulary Level Test (Schmitt et al., 2001) used as a measure of receptive vocabulary knowledge. Participants were surveyed using a test reflection questionnaire (Voss, 2012), used to elicit information with regard to examinees' perception on comparing and contrasting academic language English on the test with English in university textbooks. Results from analyses of both quantitative and qualitative data provide evidence in support of the interpretive argument for the ACCT. Each of the seven inferences had a warrant based on underlying assumptions that necessitated theoretical and empirical backing derived from the review of relevant literature and empirical analyses of both quantitative and qualitative data collected after the administration of the research instruments. Both theoretical and empirical backing could either substantiate or rebut the ACCT interpretive argument specified in the first stage of the argument-based approach.

In this chapter, I present in detail the evaluation of the evidence collected in support of the ACCT interpretive argument with a view to building the validity argument for the ACCT in the second stage of the argument-based approach. Following this, I finish off this chapter with the guiding answers to research questions, proposal of implications of this study, discussion of limitations and suggestions for future research, as well as summary of the contents in this chapter.

6.1 Development of the ACCT validity argument

As pointed out earlier, Kane's argument-based approach focuses on the validation of the interpretation and use of test scores. To achieve this, Kane proposed two stages of argument construction. First, the interpretive argument is developed to specify the interpretation and use of test scores, which in turn direct the way in which the test is to be developed and validated. The interpretive argument of the ACCT followed the TOEFL interpretive argument (Carol A Chapelle et al., 2008). Second, the validity argument is constructed to determine to what extent the score interpretation and use is valid or feasible based on evaluation of backing gathered in support of the interpretive argument. The same backing can also support other assumptions in different inferences if they are dependent. In this section, I present an evaluation of backing supporting seven inferences in the ACCT interpretive argument in order to build the validity argument for the ACCT. It should be noted, nevertheless, that evidence supporting the utilisation and consequence inference was not extensively investigated since more evidence supporting these inferences could be studied after the ACCT is used for quite an extended period.

6.1.1 Evaluating the domain inference

The domain inference was aimed to connect performance in the academic English domain with observation on the ACCT. The warrant of this inference is that student performances on the ACCT reveal the collocational competence relevant to and representative of the TLU domain in university or other higher-education settings. This warrant was found plausible due to the collected backing supporting its underlying assumptions. The first assumption is that collocations on the ACCT are representative of the TLU domain of the academic written discourse. This assumption was supported by analyses of TLU domain and corpus. The TLU domain was investigated through the analysis of academic written English from seven

academic disciplines in the academic written discourse of BNC, which is claimed to relate and represent academic written English. The analysis of corpus was presented in the test development in chapter 3.

The second assumption is that collocations on the ACCT are representative of the TLU domain of academic written discourse. This was substantiated by a systematic sampling of collocations from BNC. Collocations on the ACCT were sampled from high-frequency verb-noun collocations the TLU domain of academic written discourse in BNC as presented in chapter 3. Another backing was gained from the person-item variable map. Rasch person-item variable map showed a relative wide distribution of the item difficulty hierarchy with only two noticeable gaps. The third assumption is that the ACCT can elicit student responses which reflect the collocational competence. This assumption was supported by test item response modelling, expert review of the test, and the Rasch model analysis results. Receptive collocational competence was operationalised with a multiple-choice item format which required examinees to select a proper verb in collocation with a noun as a node (headword) in the sentential context for each pair of targeted collocations. The use of multiple-choice task to measure receptive aspect of collocational competence was backed up by theoretical evidence documented from textbooks and previous studies related to psychological testing as well as vocabulary and collocation assessment, discussed in chapter 2.

Another backing was derived from expert review of the test. Three experts were asked prior to the test trialling to evaluate the appropriateness of item format in terms of the stems or questions, best-answer choices, and alternative choices or distractors. Another backing was resulted from empirical Rasch unidimensionality analysis. PCAR, item fit indices, and point-measure correlation confirmed a significant dominant collocational construct. A final backing was gained from the Rasch item strata of 6.8 which indicated the ACCT captured almost 7 levels of collocational competence. Table 6.1 summarises backing evidence in support of the assumptions underlying the warrant of the domain inference.

Table 6.1

Summary of backing evidence in support of the assumptions underlying the warrant of the domain inference

Warrant	Underlying assumptions	Backing evidence
Observations of performance on the ACCT reflect the collocational competence representing the TLU domain of academic written English in universities or other institutions of higher education	1) Performance on the ACCT reflects collocational competence which contributes partly to performance on the academic English writing task.	• TLU domain was clearly defined and the corpus representing the TLU domain was accordingly identified.
	2) Collocations on the ACCT are representative of the TLU domain of academic written discourse.	• Verb-noun collocations were systematically sampled from varying academic domains in BNC. • Rasch person-item variable map showed a relative wide distribution of the item difficulty hierarchy with only two noticeable gaps.
	3) The ACCT can elicit test-takers' performance reflecting collocational competence.	• Item response was developed based on literature review. • Test items were evaluated and revised according to experts. • Rasch unidimensionality analysis confirmed a significant dominant collocational construct. • Rasch item strata of 6.8 indicated the ACCT captured almost 7 levels of collocational competence.

6.1.2 Evaluating the evaluation inference

The evaluation inference has the warrant that observed performance on the ACCT is evaluated to provide observed scores reflective of the collocational competence. This warrant is underlined by three assumptions. The first assumption is that the scoring procedure is appropriate to elicit responses that serve as evidence of various collocational competence levels. This assumption was supported by data checking and screening. Data were double-checked and screened for accuracy and completeness of test-taker responses and response keying. Another backing was derived from scoring and rubric development. Selection of scoring method was based on literature review. The verbs in pairs of targeted verb-noun collocations sampled from BNC were used to develop the answer key for the dichotomous scoring method. Rubric criteria were also based on pairs of these sampled verb-noun collocations. Verbs in pairs of targeted collocations were used as correct options and marked as 1 full point. The Rasch dichotomous scaling and Rasch unidimensionality analysis also supported this assumption. The dichotomous Rasch model scaled observed scores into comparable measured scores, hence contributing to the standardisation of scoring process. Rasch Unidimensionality analysis based on PCAR, point-measure correlation, and item fit statistics confirmed that dichotomous item scoring is appropriate for eliciting the single collocational construct under measure.

The second assumption is that test administration condition is conducive for test-takers to maximally demonstrate collocational competence. This assumption was backed up through the trialling of the multiple-choice task which taped into performance of collocational competence through a discrete receptive, context-dependent task format. Time allowed for the test was sufficient for examinee to maximally demonstrate their collocational competence. Scores from piloted study were also evaluated based on CTT. The third assumption is that psychometric properties of the ACCT are appropriate for norm-referenced evaluation. This assumption was supported by the evidence that descriptive statistics indicated that the ACCT scores using the dichotomous scoring method were normally distributed, point-measure correlations of 29 ACCT items were over 0.3, the person-item variable map showed a relatively well match of person and item distributions, which is appropriate for norm-referenced interpretation. Table 6.2 summarises backing evidence in support of the assumptions underlying the warrant of the evaluation inference

Table 6.2

Summary of backing evidence in support of the assumptions underlying the warrant of the evaluation inference

Warrant	Underlying assumptions	Backing evidence
Observed performance on the ACCT is evaluated to provide observed scores reflective of the collocational competence.	1) The scoring procedure is appropriate to elicit responses that serve as evidence of various collocation competence levels.	• Data were double-checked and screening for response accuracy and completeness. • Scoring and rubric were developed based on literature review and sampled collocations from BNC. • Rasch dichotomous scaled responses into interval logits or measures. • Rasch unidimensionality analysis confirmed a significant dominant collocational construct.
	2) Test administration condition is conducive for test-takers to maximally demonstrate collocational competence.	• The ACCT was trialled and evaluated based on CTT in the pilot study. • Time allowed for the test was sufficient.
	3) Psychometric properties of the ACCT are appropriate for norm-referenced evaluation.	• Descriptive statistics showed a normal distribution of the ACCT score data. • Point-measure correlations were positive and over 0.3. • The distribution of person ability relatively matched the distribution of item difficulty.

6.1.3 Evaluating the generalisation inference

The generalization inference has the warrant that observed scores on the ACCT are estimates of expected scores which are congruent across items and invariant across gender. This warrant was supported by four assumptions. The first assumption that estimates of test-takers' performance can consistently distinguish among test-takers was substantiated by coefficient alpha reliability and Rasch internal consistency indices. Item reliability (0.96), item separation (4.9), and item strata (6.86), coefficient alpha (0.89), person reliability (0.86), person separation (2.48), and person strata (3.64) were far beyond the threshold criteria. Another backing was from the person-item variable map. The map showed graphically a relatively well-matched person-item distribution, indicating precise assessment of the ACCT for the examinees. The Rasch person-item babble map visually revealed that person and item measure were overall well mapped, indicating precise assessment of the ACCT items for the examinees. All these indicated that the ACCT consistently distinguished and precisely measured this sample of EFL graduate students.

The second assumption that psychometric properties of the ACCT items are invariant across males and females who had equal collocational competence levels was backed up by gender-based DTF and uniform DIF analyses. DTF and uniform DIF analyses indicated that overall the ACCT difficulty measure was not invariant across gender on the test level and only five ACCT items (1, 21, 25, 29 and 30) displayed significant and substantive uniform DIF on the item level. However, excluding these items might cause the instrument to fail to capture important aspects of the construct, causing construct underrepresentation (Schumacker, 2004; Wolfe & Smith, 2007b). The third assumption that the test specification of the ACCT is adequately detailed and consistent to develop equivalent task or test forms was supported by test development process in chapter 3 and development of test specification. Test development process and test specification (see Appendix A) were presented in the way that equivalent test tasks and test forms can replicate. The fourth assumption that the paper-based administration of the test is sufficiently uniform to produce consistent results was supported by task trialling and CTT-based evaluation. In the pilot study, the researcher explained the instruction and delivered the test in classroom. Table 6.3 summarises backing evidence in support of the assumptions underlying the warrant of the generalisation inference.

Table 6.3

Summary of backing evidence in support of the assumptions underlying the warrant of the generalisation inference

Warrant	Underlying assumptions	Backing evidence
Observed scores on the ACCT are estimates of expected scores which are congruent across items and invariant across gender.	1) Estimates of test-takers' performance can consistently distinguish among test-takers.	• Rasch internal consistency indices were high and thus indicated reliable, consistent assessment of the ACCT. • Rasch person-item variable map showed a relatively well-matched person-item distribution, indicating precise assessment of the ACCT for the examinees. • Rasch person-item babble map showed that person and item measures were well mapped, indicating precise assessment of the ACCT for the examinees.
	2) Psychometric properties of the ACCT item are invariant across males and females who have equal collocational competence levels.	• Rasch differential test functioning analysis confirmed an invariance measurement of the ACCT across gender. • Rasch differential item functioning analysis showed five significant gender-basd uniform DIF items on the ACCT.
	3) The test specification of the ACCT is adequately detailed and consistent to develop equivalent task or test forms.	• Test specification was clearly developed for replication.

Table 6.3

Summary of backing evidence in support of the assumptions underlying the warrant of the generalisation inference

Warrant	Underlying assumptions	Backing evidence
	4) The paper-based administration of the test is sufficiently uniform to produce consistent results	• The ACCT was trialled, monitored, and instructed.

6.1.4 Evaluating the explanation inference

The explanation inference is based on the warrant that expected scores are attributed to the collocational competence construct in the academic written discourse. This warrant is underlined by four assumptions. The first assumption that performance on the ACCT reflects test-takers' collocational competence was supported by construct definition, scoring and rubric development, and Rasch applications. Interactionist construct definition was thoroughly reviewed. Scoring and rubric were developed based literature review and targeted collocations from the corpus. The person-item babble map showed fit of most item and person measures, thereby indicating relevant assessment of the ACCT with regard to the latent construct of collocational competence. Overall, PCAR, a scree plot of the standardised residual contrast, point-measure correlation, and item fit statistics confirmed the substantive unidimensionality of collocational construct under measure. The person-item variable map also showed a relatively wide distribution of item distribution, hence indicating relatively representative assessment of the ACCT with regard to the latent construct of collocational competence.

The second assumption that the construct under measure is collocational competence which is defined as a restricted lexical collocation in academic written texts was supported by collocation definition, and several applications of Rasch measurement analysis. Collocation was defined based on a phraseologist approach reviewed in chapter 2. On the whole, Rasch unidimensionality analysis confirmed the substantive unidimensionality of collocational construct in question. The person-item variable map demonstrated that item difficulties were relatively widely dispersed on the item difficulty scale, indicating that ACCT items were relatively representative of the measured collocational competence in the TLU domain. The person-item babble

map showed relevant and somewhat representative assessment of the ACCT in relation to the latent construct of collocational competence.

The third assumption that scores on the ACCT correlate positively to other tests of English language proficiency related to the construct was supported by correlation analysis between ACCT scores and AVLT scores and correlation analysis between collocation competency measures and vocabulary knowledge measures. The Pearson product-moment correlation showed statistically significant good relationship between ACCT scores and AVLT scores (r = 0.74) and between collocation competency measures and vocabulary knowledge measures (r = 0.79). The fourth assumption that while doing the test, test-takers use cognitive process related to collocation use in academic language was supported by Rasch multiple-choice distractor analysis. The multiple-choice distractor analysis revealed that only a correct choice and distractors of Item 19 did not function in an intended way. Test reflection survey indicated that most examinees demonstrated relevant meta-cognitive strategies while doing the ACCT. Table 6.4 summarises backing evidence in support of the assumptions underlying the warrant of the explanation inference.

Table 6.4

Summary of backing evidence in support of the assumptions underlying the warrant of the explanation inference

Warrant	Underlying assumptions	Backing evidence
Expected scores are attributed to the collocational competence construct in the academic written discourse.	1) Performance on the ACCT reflects test-takers' collocational competence.	• Interactionist construct definition was thoroughly reviewed. • Scoring and rubric were developed based literature review and targeted collocations from BNC. • Rasch unidimensionality analysis confirmed a significant dominant collocational construct. • Rasch person-item variable map showed relatively representative assessment of the ACCT with only two gaps.

Table 6.4

Summary of backing evidence in support of the assumptions underlying the warrant of the explanation inference

Warrant	Underlying assumptions	Backing evidence
		• Rasch person-item babble map showed relevant assessment of the ACCT.
	2) The construct to be assessed is collocational competence which is defined as a restricted lexical collocation in academic written texts.	• Phraseologist collocation definition was thoroughly reviewed. • Rasch unidimensionality analysis confirmed a significant dominant collocational construct. • Rasch person-item variable map showed relatively representative assessment of the ACCT with only two huge gaps. • Rasch person-item babble map showed relevant assessment of the ACCT.
	3) Scores on the ACCT correlate positively to other tests of English language proficiency related to the construct	• Correlation analysis showed a relatively high correlation between ACCT scores and AVLT scores • Correlation analysis showed a relatively high correlation between ACCT theta and AVLT theta
	4) While doing the test, test-takers use cognitive process related to collocation use in academic language	• Rasch multiple-choice distractor analysis showed only Item 19 had malfunctioning distractors. • Test reflection survey showed that most examinees exercised their relevant cognitive strategies while doing the ACCT.

6.1.5 Evaluating the extrapolation inference

The extrapolation inference is based on the warrant that the collocational competence construct as measured by the ACCT accounts for relevant language performance in the academic discourse in university or other higher-education settings. This warrant is underlined by two assumptions. The first assumption that collocations on the ACCT reflect those that the test-takers will be exposed to in the context of the academic written discourse was supported by TLU domain and corpus analysis and Rasch person-item variable map. The TLU domain was investigated through the analysis of BNC, as discussed previously in the evaluation of the domain inference. The person-item variable map indicated although there were two huge gaps in the item difficulty distribution that did not have items targeted to some high and low-ability students, item difficulties were widely dispersed on the item difficulty scale, indicating that ACCT items were well differentiated by this group of students and relatively representative of the measured collocational competence.

The second assumption that scores on the ACCT distinguish among proficiency groups with and without experience and topical knowledge of academic language was supported by Rasch person strata, Rasch person-item variable map, and a one-way independent ANOVA. Overall, the assumption was well supported by the Rasch evidence. The person strata index (3.64) indicated that at least three distinct competency levels were differentiated by ACCT items. The Rasch person-item variable map showed that student collocational competencies were widely distributed along the person competency scale, meaning that ACCT items well targeted a wide range of student collocational competency. Furthermore, a one-way independent ANOVA showed a statistically significant difference in the ACCT scores amongst three proficiency groups. Table 6.5 summarises of backing in support of the assumptions underlying the warrant of the extrapolation inference

Table 6.5

Summary of backing evidence in support of the assumptions underlying the warrant of the extrapolation inference

Warrant	Underlying assumptions	Backing evidence
The collocational competence construct as measured by the ACCT accounts for relevant language performance in the academic discourse in university or other higher-education settings.	1) Collocations on the ACCT reflect those that the test-takers will be exposed to in the context of the academic written discourse.	• TLU domain was clearly defined and the corpus representing the TLU domain was accordingly identified. • Rasch person-item variable map showed relatively representative assessment of the ACCT with only two huge gaps.
	2) Scores on the ACCT distinguish among proficiency groups with and without experience and topical knowledge of academic language.	• Rasch person strata of 3.6 indicated the ACCT distinguished at least three competency levels. • Rasch person-item variable map showed a wide range of collocational competence. • Analysis of variance showed a significant difference between three proficiency groups.

6.1.6 Evaluating the utilisation inference

The utilisation inference is based on the warrant that Performance on the ACCT contributes to making appropriate norm-referenced decisions about placement in English language courses in universities or other institutions of higher education. Results from cut-score study and classification error study served as empirical evidence backing the two assumptions that the interpretation of the ACCT scores provides enough information which contributes to the decision making process and the ACCT scores are intended to be used to contribute to and facilitate student

placement decision in appropriate English language courses in universities or other institutions of higher education.

Cut score and classification studies showed that low and high-competency students were accurately classified but mid-competency students was not accurately classified as some of students in this level were potentially misclassified. Cut scores and classification accuracy may need to be further established and investigated in order to classify examinees as accurately and consistently as possible. More potential evidence supporting these assumptions may be derived from analysis of correlation between the ACCT scores and English course grades. Table 6.6 summarises backing evidence in support of the assumptions underlying the warrant of the utilisation inference.

Table 6.6

Summary of backing evidence in support of the assumptions underlying the warrant of the utilisation inference

Warrant	Underlying assumptions	Backing evidence
Performance on the ACCT contributes to making appropriate norm-referenced decisions about placement in English language courses in universities or other institutions of higher education	1) The interpretation of the ACCT scores provides enough information which contributes to the decision making process	• Contrasting group cut-score setting gave two sets of cut-scores based on scores and theta. • Classification error analysis showed little error of cut scores in low and high-competency groups but high error in the mid-competency group.
	2) The ACCT scores are intended to be used to contribute to and facilitate student placement decision in appropriate English language courses in universities or other institutions of higher education	• Contrasting group setting gave two sets of cut-scores based on scores and theta. • Classification error analysis showed little error of cut scores in low and high-competency groups but high error in the mid-competency group.

Table 6.6
Summary of backing evidence in support of the assumptions underlying the warrant of the utilisation inference

Warrant	Underlying assumptions	Backing evidence
		• Correlation study should be checked between ACCT scores and English class grades.

6.1.7 Evaluating the consequence inference

The consequence inference is based on the warrant that the interpretation and use of the ACCT scores are appropriate and advantageous for all test users and stakeholders. This warrant requires two assumptions that the construct of the ACCT raises awareness about the importance of collocations in academic English and the construct of the ACCT raises awareness of introducing the importance of collocations in English instruction and material developments. Empirical evidence supporting the consequence inference was not investigated in this study since it can be backed up by empirical evidence from future washback study and stakeholder survey. Table 6.7 summarises potential backing in support of the assumptions underlying the warrant of the consequence inference and Table 6.8 summarises all of the evidence collected in support of the ACCT validity argument in the present study.

Table 6.7
Summary of potential backing in support of the assumptions underlying the warrant of the consequence inference

Warrant	Underlying assumptions	Potential backing
The interpretation and use of the ACCT scores are appropriate and advantageous for all test users and stakeholders.	1) The construct of the ACCT raises awareness about the importance of collocations in academic English.	• Future washback study • Future stakeholder survey
	2) The construct of the ACCT raises awareness of introducing the importance of collocations in English instruction and material developments	• Future washback study • Future stakeholder survey

Table 6.8

Summary of evidence in support of the ACCT validity argument

Sources of validity evidence	Types of inferences						
	Domain	Evaluation	Generalisation	Explanation	Extrapolation	Utilisation	Consequence
1) TLU domain and corpus analysis	✓		✓		✓		
2) Systematic collocation sampling	✓						
3) Test specification development			✓				
4) Item response development	✓						
5) Scoring and rubric development		✓		✓			
6) Interactionist construct definition				✓			
7) Expert review of the test	✓						
8) Test trialling and evaluation		✓	✓				
9) Adequate testing time		✓					
10) Data preparation and screening		✓					
11) Descriptive statistics		✓					
12) Rasch dichotomous scaling		✓					
13) Rasch unidimensionality analysis	✓	✓		✓			
14) Rasch internal consistency indices			✓				
15) Rasch item strata index	✓						
16) Rasch person strata index					✓		
17) Rasch differential test functioning			✓				
18) Rasch differential item functioning			✓				
19) Rasch person-item variable map	✓	✓	✓	✓	✓		
20) Rasch person-item babble map			✓	✓			
21) Rasch multiple-choice distractor functioning				✓			
22) Correlation study				✓			
23) Analysis of variance					✓		
24) Test reflection survey				✓			
25) Cut score establishment						✓	
26) Classification error estimation						✓	
27) Washback study							
28) Stakeholder survey							

6.2 Structuring stages of evidence collection for the ACCT validity argument

It is important to keep in mind that some sources of evidence supported more than one inference as the inferences are interrelated and thus sound evidence from one inference also substantiated other inferences. Figure 6.1 displays the structure of evidence collection procedure in support of the ACCT validity argument. It is based on the stages of the TOEFL validity argument (Carol A Chapelle et al., 2008, p. 349). Each stage is supported by the empirical backing collected to support the inference from the domain inference at the bottom up to the consequent inference at the top. Some baking can support more than one inference as the inferences are interdependent.

The intended backing for the consequence inference was not investigated in this study and further study needs to bridge this discrepancy by taking into account, for example, washback study and stakeholder survey and documenting relevant rationales in order to gain more evidence in support of the consequence inference. It should be reminded that validation is an ongoing process since validity changes over time. Therefore, the interpretation and use of test scores should be modified and revised as occasions demand and as test users see fit. Once the interpretation and use of test scores are revised, then validity evidence need to be refreshed and re-accumulated to enhance the validity of score interpretation and use.

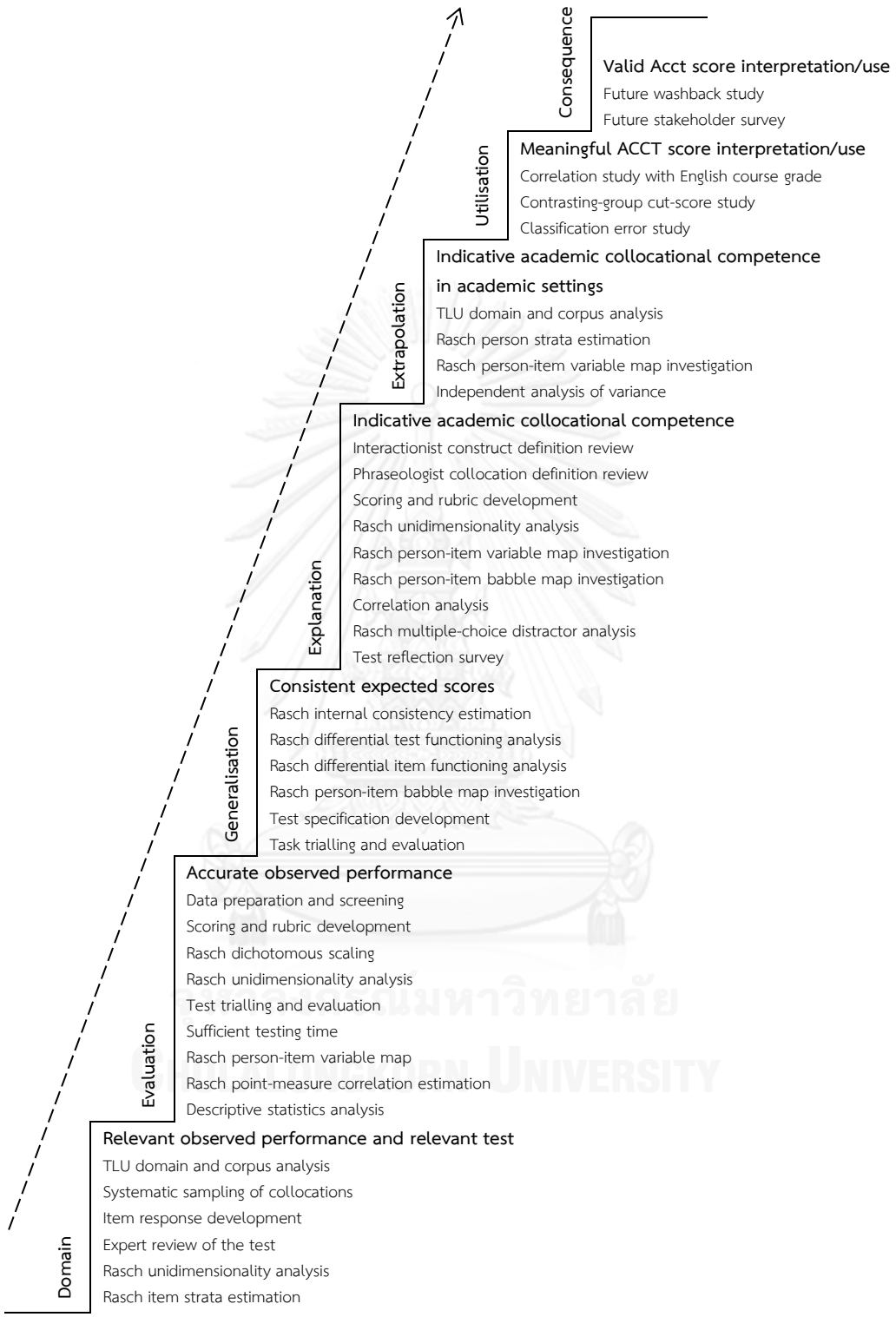

Figure 6.1. Stages of evidence collection in support of the ACCT validity argument

6.3 Guiding responses to research questions

In this section, I present the information in this thesis that guide responses to the research questions of the present study. The three research questions addressed in chapter 1 are: 1) to what degree are scores on the ACCT interpreted as an indicator of collocational competence of EFL university students and used for placement decision in English language courses in universities or other academic institutions at tertiary level?, 2) how does the argument-based approach to validation help develop the ACCT and validate the proposed interpretation and use of scores on the ACCT?, and 3) how does the Rasch psychometric model help validate psychometric properties of the ACCT?. The first question is presented first, then the presentation of responses to the second and third research questions are provided respectively.

6.3.1 Response to research question 1

Responses to this question are derived primarily from chapter 6 with a particular focus on the section presenting the construction of the ACCT validity argument. In can be concluded that overall the ACCT scores are reasonably interpreted as reflecting collocational competence of EFL university students and appropriately used for placement decision. As discussed earlier in this chapter, both theoretical and empirical sources of evidence were collected to support the proposed interpretation and use of the ACCT scores outlined in the interpretive argument. Theoretical evidence was collected in chapter 2 where relevant theory or a priori was reviewed and theoretical evidence also directed the way in which empirical evidence was gathered. Empirical evidence was collected in chapter 5 using three research instruments discussed in chapter 4. The previous chapter presented empirical quantitative and qualitative findings from empirical data analysis.

Each of theoretical and empirical evidence supports to a varying degree one or more assumptions underlying the warrants of the inferences in the ACCT interpretive argument, thereby contributing overall to the reasonable degree of the ACCT validity argument. Explanation of the degree of the ACCT validity argument was earlier presented in detail in the section on building the validity argument for the ACCT in this chapter. More evidence needs to be collected to enhance the degree of the ACCT validity argument or the validity of the interpretation and use of the ACCT scores.

6.3.2 Response to research question 2

Responses to this question are gained primarily from almost all chapters in this study, from chapter 2 to chapter 6. The role of the argument-based approach in developing the ACCT began after the purpose of the test development was defined. As earlier mentioned, the purpose of the test guides the development of both the ACCT and the ACCT interpretive argument which were carried out in a parallel fashion. The argument-based approach consists of two intertwined argument-building stages, the interpretive argument and the validity argument. It was the interpretive argument development stage that comes into play in test development. The interpretive argument framework that outlined the proposed interpretation and use of the ACCT scores serve as an overall guideline for developing and validating the ACCT. As the ACCT was being developed, the ACCT interpretive argument was revised and modified to fit the proposed interpretation and use and the current study. When the proposed interpretation and use became appropriate for this study, the test design was started following a priori relevant to the proposed interpretation and use in the ACCT interpretive argument. Evidence relevant to the validity of the proposed interpretation and use was accumulated when theory and priori was reviewed for test development.

In the development stage of the ACCT, the ACCT interpretive argument was also developed to represent the proposed interpretation and use of the ACCT scores and to correspond with the characteristics of the ACCT. When assumptions are found to be too complex, the ACCT and the interpretive argument were revised and modified to fit the context of the study and make it possible to back up the assumptions. This iterative process of development and revision of the ACCT and the interpretive argument proceeded until the ACCT and the ACCT interpretive argument was consistent and appropriate within the context of the present study. The proposed interpretation and use of the ACCT scores in the interpretive argument influenced the way in which decision was made in relation to the design of the ACCT during test development process. The development stage of the ACCT and the ACCT interpretive argument also produced evidence that supported the intended interpretations and use of the ACCT scores.

The development of the ACCT and the validation of the proposed interpretation and use of the ACCT scores took an extended effort, and the focus of the inquiry shifted over time. As such when the interpretive argument and inquiries were fleshed out, more evidence was required and collected as long as it helped support the development and validation of the ACCT. To sum up the interpretive

argument in the argument-based approach serves as the guideline for the development of the ACCT. Details of the ACCT interpretive argument was present in chapter 2 and detailed description of how the ACCT was developed was presented in chapter 5.

To validate the score interpretation and use of the ACCT, both interpretive argument and validity argument play a very important role in the validation. As pointed out by Kane (2013), these two arguments are likely to be intertwined in practice and are not neatly sequential. Once the ACCT interpretive argument was clearly and adequately developed and backed up by evidence, then the evaluation of the evidence collected to support the ACCT interpretive argument was conducted in the second stage, development of the ACCT validity argument. In the appraisal stage, the focus is placed upon the development of an adequate validity argument. The coherence and completeness of the ACCT interpretive argument for the proposed interpretation and use was evaluated in chapter 6.

Both theoretical and empirical evidence were evaluated to the coherence and completeness of the assumption underlying the warrants of the inferences in the ACCT interpretive argument. As mentioned previously validity is a matter of degree and thus collected evidence provide varying degree of the validity of the interpretation and use of the ACCT scores. The development of the ACCT and the validation of the proposed interpretation and use of the ACCT scores took an extended effort, and the focus of the inquiry shifted over time. As such when the interpretive argument and inquiries were fleshed out, more evidence was required and collected as long as it helped support the development and validation of the ACCT.

6.3.3 Response to research question 3

Responses to this question are obtained primarily from chapter 5 where results of an analysis of the unidimensional dichotomous Rasch model were presented and also from chapter 6 where a wealth of Rasch-based evidence was mapped onto the argument-based framework. The Rasch measurement approach was applied with the aim of accumulating empirical evidence reinforcing or rebutting the degree of the ACCT validity argument or the claimed interpretation and use of the ACCT scores. It is evident from this study that the Rasch measurement approach provided several pieces of empirical psychometric evidence that serve as sound and sufficient evidential backing in support of several assumptions underlying the

warrants of the inferences laid out in the ACCT interpretive argument. In short, the Rasch measurement approach did provide empirical evidence that made the interpretation and use of the ACCT scores compellingly feasible. This proves that the Rasch measurement approach serves as the cost-effective, time-saving psychometric tool for the contemporary validation of measurement instruments.

In this study, several applications of the Rasch measurement approach were mapped onto several assumptions underlying the inferences of the argument-based validation model. Assumptions underlying the domain inference were properly supported by Rasch-based evidence. The assumption that collocations on the ACCT were representative of the TLU domain of the academic written discourse was backed up Rasch evidence that the item strata index indicated that ACCT items were categorised into at least six difficulty levels, the point-measure correlation values were over zero and positive, the person-item variable map showed a relatively wide dispersion of item difficulty hierarchy though with two noticeable gaps. These reasonably ensured that collocations on the ACCT were representative of the TLU domain of the academic written discourse. Another assumption that that the ACCT can elicit student responses which reflect the collocational competence was made possible by that evidence that the Rasch unidimensionality analysis confirmed a significant dominant collocational construct since PCAR showed that the ACCT scores accounted for over a minimal criterion (20%) of the focal collocational construct, ACCT items showed positive point-measure correlations, and the item fit statistics revealed that 29 ACCT items well fit the Rasch model, meaning that the ACCT can compellingly elicit student responses which reflect the collocational competence under measure.

The assumption behind the evaluation inference was substantially supported by Rasch evidence. The assumption that the scoring procedure is appropriate to elicit responses that serve as evidence of various collocation competence levels was made feasible due to the fact that the dichotomous Rasch model scaled observed scores into comparable, interval data, hence contributing to the standardisation of scoring process. What is more, the Rasch unidimensionality analysis confirmed a significant dominant collocational construct and hence the scoring procedure was appropriate for eliciting the collocational competence construct. Another assumption that psychometric properties of the ACCT are appropriate for norm-referenced evaluation was supported by Rasch evidence all items had positive point-measure correlations and up to 29 items had point-measure correlations over 0.3. Another Rasch backing for this assumption was that the distribution of person competence

hierarchy was relatively well matched with that of item difficulty hierarchy, making it possible for the ACCT scores to be normatively evaluated. To sum up, Rasch dichotomous scaling, Rasch unidimensionality analysis, Rasch point-measure correlation, and Rasch person-item variable map reasonably supported the the feasibility of the evaluation inference.

Assumptions underlying the generalisation inference were reasonably substantiated by Rasch evidence. The assumption that estimates of test-takers' performance can consistently distinguish among test-takers was made possible by Rasch internal consistency indices, Rasch person-item variable map, and Rasch person-item babble map. Item reliability, separation, and strata and person reliability, separation, and strata were beyond acceptable criteria and thus reassure internal consistency indices were high. The person-item variable map showed a relatively well-matched person-item distribution, indicating precise, reliable assessment of the ACCT for the examinees. The person-item babble map showed that person and item measures were generally well-mapped, hence reassuring precise assessment of the ACCT for the examinees. Another assumption that estimates of test-takers' performance can consistently distinguish among test-takers was reasonably feasible by Rasch evidence that the DTF analysis showed a slight dispersion of variant items, indicating a consistent measurement of the ACCT across gender. The DIF analysis also uncovered that as many as 25 ACCT item possessed invariance difficulty indices across males and females while only five ACCT item (Items 1, 21, 25, 29 and 30) appeared to display uniform DIF or difficulty measure variance across gender. Overall the ACCT difficulty was invariant across gender subgroups yet only five items that had different difficulty measures for male and female subgroups. Therefore, Rasch-based DTF and uniform DIF ensured that psychometric properties of the ACCT item are invariant across males and females who have equal collocational competence levels.

Assumptions underlying the explanation inference were reasonably supported by Rasch evidence. The assumption that performance on the ACCT reflects test-takers' collocational competence was substantiated by the evidence that Rasch unidimensionality analysis confirmed a dominant unidimensional collocational construct measured by ACCT items and the person-item variable map confirmed that the ACCT items measured representative collocational construct by showing widely-dispersed and relatively well-matched distributions of student competencies and item difficulties in spite of two huge gaps in the item distribution that did not have items targeted to some high and low-proficiency students. Moreover, Rasch person-

item babble map ensure relevant assessment of the ACCT with regard to the latent construct of collocational competence by demonstrating almost all ACCT items were located within the acceptable zone and close to the latent construct scale. Rasch unidimensionality analysis and Rasch person-item variable and babble maps may help support another assumption that while doing the test, test-takers use cognitive process related to collocation use in academic language. This assumption was also made cogent by Rasch-based multiple-choice distractor analysis which revealed that only one ACCT item (Item 19) had a correct option and distractors that did not function in the way around which they were developed. All Rasch applications help ensure that assumptions underlying that explanation inference were sufficiently substantiated and thus feasible.

The assumption underlying the extrapolation inference was well supported by Rasch evidence. It was reasonably assumed that collocations on the ACCT reflect those that the test-takers will be exposed to in the context of the academic written discourse since Rasch person-item variable map showed a relatively wide dispersion of item difficulty on the construct variable scale despite two noticeable gaps, thereby signifying relatively representative assessment of the ACCT. It could convincingly be assumed as well that scores on the ACCT distinguish among proficiency groups with and without experience and topical knowledge of academic language. This was due to the Rasch evidence that since the person strata index revealed that approximately three distinct competency levels were differentiated by ACCT items and the person-item variable map indicated that students competencies were widely spanned and relatively equally spaced along the collocational competency hierarchy. Therefore, could be concluded that these Rasch applications reasonably supported the assumptions underpinning the extrapolation inference. In terms of the utilisation inference, the Rasch measurement model helped provide competency measures or theta which could be used as performance data for cut-score establishment and classification error analysis. In the Rasch model analysis, competency measures were converted from the ACCT scores and were on the interval logit scale; therefore, using competency measures for cut score and classification error analyses, or even other parametric statistics analyses could provide more meaningful measurement outcomes (Embretson & Reise, 2000; Iramaneerat et al., 2008). This study did not provide applications of the Rasch measurement approach to support the consequence inference since the consequence inference is beyond the scope of the current study. It will be of great

value and interest for further research to apply the Rasch psychometric model to seek empirical evidence that can be used in support of the consequence inference.

It is evident that a Rasch measurement approach provided sound and sufficient evidence strengthening the ACCT validity argument. Rasch indices and visual plots reasonably serve as essential psychometric properties of the ACCT, as already presented above. These psychometric properties are considered as empirical evidence backing the ACCT interpretive argument and strengthening the ACCT validity argument. This study indeed underscores the cost-effective, time-saving advantages that a Rasch measurement approach offers to test developers, test validators, and test validation frameworks, particularly Kane's argument-based approach.

6.4 Implications of the study

Findings from this study made several significant contributions to the study of collocation as well as the development and validation of language assessment instruments. Implications of the study are discussed in terms of theoretical, methodological, and pedagogical aspects.

On a theoretical front, the present study could provide the way of applying the argument-based approach to define construct definition of collocational competence and model the framework for validating the interpretation and use of language test scores. The findings shed novel light into modelling a more thorough framework for developing and validating language tests that could provide scores which is appropriately interpreted with inference to linguistic competence and used with regard to placement decision on placing test-takers into appropriate English language courses in universities or other higher-education institutions. Another theoretical implication is that this study presents a way of developing a test of specific collocation to assess specific collocational knowledge as an indicator of general English proficiency and as a construct of a measure for placement decision in academic English courses in university or other higher-education settings. Using a test of specific collocational knowledge provides more information on test-takers' language proficiency, which in turn contributes to a well-made testing decision.

On a methodological dimension, the current findings inform test developers of deploying the hybrid of a Rasch model and an argument-based model to develop measures of language knowledge and validate the score interpretation and use of language assessment instruments. While most of prior studies applied CTT to investigate item and test characteristics, far fewer studies used the Rasch model or

other IRT models to examine the psychometric quality of language tests. The findings from this study could draw more attention to applying the Rasch IRT model to assess collocational competence and other language abilities.

On a pedagogical account, the present study could raise the awareness of introducing collocations in English language instruction and materials in class since awareness is considered as an important aspect of language learning. If the awareness of the importance of teaching and learning collocations increases, this means that the use of collocation tests could potentially lead to the intended consequences in the form of positive washback.

6.5 Limitations and suggestions

Several limitations were recognised in the present study. Firstly, test-takers' cognitive process was not sufficiently investigated in this study since the current study used only the test reflection survey to tap into examinees' cognitive process. Therefore, further research should be conducted using, for example, think-aloud protocol and other verbal report methods to scrutinise cognitive and metacognitive process and test-taking strategies of test-takers with a view to providing empirical evidence in support of the ACCT validity argument. Secondly, the consequence inference was not examined in the present study due to that fact that more evidence for this inference can be gathered after the ACCT was used for a while. Nevertheless, evidence supporting other inferences could to a certain extent confirm positive consequences of the use of the ACCT scores. To ensure positive consequences of the use of the ACCT scores, further research is needed to investigate the impact or wasback of the ACCT utilisation and survey stakeholders' opinion on the utilisation of the ACCT scores.

Thirdly, the current study did not further examine the actual causes of gender-based uniform DIF items on the ACCT. DIF items can be caused by several factors and the fact that ACCT items exhibited significant DIF does necessarily mean that these DIF items are actually biased. What DIF can inform at this stage is that the DIF items on the ACCT had different psychometric properties in terms of difficulty measures for male and female EFL graduate test-takers. To uncover whether DIF items are indeed biased towards a particular gender subgroup, further research is called for to delve more deeply into the actual causes of DIF items on the ACCT, which would provide more evidence solidifying or challenging the ACCT validity argument. Another caveat of the study is that local independence assumption of the

Rasch measurement model was not adequately examined. Although it could be reasonably assumed that local independence held by virtue of the present unidimensional construct, other Rasch applications should further be employed to confirm that individual response to a particular ACCT item does not by any means influence his or her response to any other ACCT items.

A further caveat is that this study used a contrasting-group standard setting method to exemplify the way of establishing cut scores for classifying test-takers into different competency levels and a contrasting-group method was based on normal trendlines of the score and theta distributions of prior proficiency groups, classified based on CU-TEP, TOEFL, and IELTS scores reported by test-takers. Therefore, it could not be completely assumed that the sets of cut scores could completely accurately classy examinees but it should be positive to a reasonable extent that the ACCT scores would distinguish examinees into different competency levels. However, the cut scores can be adjusted depending on the test users' judgement and decision and standard setting methods for cut score establishment. It is thus of greater use that future research be carried out to address this limitation by using different standard setting methods in establishing cut scores for the ACCT and analysing their classification error, consistency, and accuracy. All these would yield empirical evidence for or against the utilisation and consequence inference of the ACCT scores.

One more limitation was identified pertaining to correlation study. This study investigated only the relationship between scores and ability logits both on the ACCT to provide preliminary empirical evidence for the explanation inference. Although a relatively strong relation was significantly found between the ACCT and the AVLT, further studies should be undertaken to explore that relationship between the ACCT and reading comprehension tests or other measures of linguistic constructs related to collocational competence, as guided by a priori. If the ACCT scores or ability logits were found to correlate with other measures of related constructs or non-testing behaviours, it can then be more confident that the ACCT provides scores which can be interpreted as reflecting collocational competence.

Finally, although a sample of 193 EFL graduate students was sufficient to provide stable estimates for the Rasch measurement analysis in the present study, future research should replicate this study with a larger sample size to provide more stable person and item estimates and with more characteristics for EFL test-takers, such as varying academic fields and undergraduate students, to enhance the generalisability of the ACCT use. Additionally, further research should employ different sampling approaches and take advantage of different corpora to obtain

collocations that represent as much as possible the TLU domain of academic written English. As much as evidence was collected to support the ACCT validity argument in this study, more evidence still needs to be gathered to maximise the validity argument of the ACCT in order to enhance the degree of the appropriateness of the interpretation and use of the ACCT scores.

Despite several limitations addressed previously, findings from this study do make a significant contribution to the theoretical and practical paradigm of language assessment and evaluation. It is highly recommended that the ACCT should be used as a supplementary test for the existing placement tests or used to provide information as part of decision-making process about screening or placing EFL graduate students into proper English courses in university or other high-education institutions. Information provided merely by the ACCTS may not entirely guarantee appropriate interpretation and use of the ACCT scores since no assessment instruments can perfectly measure psychological, unobserved constructs. For this reason, only through using multiple measurement instruments can test users be certain that decision is properly made as intended.

6.6 Chapter summary

This chapter is concerned primarily with development of the ACCT validity argument, which is the second stage of the argument-based approach to validation. I began by presenting a brief summary of the research purposes. I then presented how the validity argument of the ACCT was developed based on evaluation of the evidence collected to support the assumptions underlying the warrants of the inferences in the interpretative argument. Following this, brief and concise responses to research questions were presented as the guideline for answering research questions in this study. This chapter ends with a discussion of implications of the current study as well as the limitations and recommendations of the present study.

REFERENCES

Akbari, R. (2012). Validity in language testing. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *Second language assessment* (pp. 30–36). New York, NY: Cambridge University Press.

Akbarian, I. h. (2010). The relationship between vocabulary size and depth for ESP/EAP learners. *System, 38*(3), 391–401. doi: 10.1016/j.system.2010.06.013

Alsakran, R. A. (2011). *The productive and receptive knowledge of collocations by advanced Arabic-speaking ESL/EFL learners.* (1497925 M.A.), Colorado State University, Ann Arbor. ProQuest Dissertations & Theses Global database.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Aryadoust, V. (2009). Mapping the Rasch-based measurement onto the argument-based validity framework. *Rasch Measurement Transactions, 23*(1), 1192–1193.

Association, A. E. R., Association, A. P., & Education, N. C. o. M. i. (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice*. Oxford: Oxford University Press.

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*(1), 1145–1146.

Baghaei, P., & Amrahi, N. (2011). Validation of a multiple choice English vocabulary test with the Rasch model. *Journal of Language Teaching and Research, 2*(5). doi: 10.4304/jltr.2.5.1052-1060

Bahns, J. (1993). Lexical collocations: a contrastive view. *ELT journal, 47*(1), 56–63. doi: doi: 10.1093/elt/47.1.56

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101–114. doi: http://dx.doi.org/10.1016/0346-251X(93)90010-E

Baleghizadeh, S., & Golbin, M. (2010). The effect of vocabulary size on reading comprehension of Iranian EFL learners. *Linguistic and Literary Broad Research and Innovation, 1*(2), 33–47.

Bazzaz, F. E., & Samad, A. A. (2011). The use of verb noun collocations in writing stories among Iranian EFL learners. *English Language Teaching, 4*(3), 158–163. doi: 10.5539/elt.v4n3p158

Beglar, D. (2009). A Rasch-based validation of the vocabulary size test. *Language Testing, 27*(1), 101–118. doi: 10.1177/0265532209340194

Benson, M., Benson, E., & Ilson, R. F. (2010). *The BBI combinatory dictionary of English: Your guide to collocations and grammar*. Amsterdam: John Benjamins Publishing

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2013). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research, 18*(1), 54–74. doi: 10.1177/1362168813505389

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd Ed.). Mahwah, NJ: Lawrence Erlbaum.

Boone, W., Staver, J., & Yale, M. (2014). *Rasch analysis in the human sciences*: Springer.

Brennan, R. L. (2013). Commentary on "validating the interpretations and uses of test scores". *Journal of Educational Measurement, 50*(1), 74–83. doi: 10.1111/jedm.12001

Carr, N. T. (2011). *Designing and analyzing language tests*. Oxford: Oxford University Press.

Carter, R. (1998). *Vocabulary: Applied linguistic perspectives* (2nd ed.). London: Routledge.

Chan, T.-p., & Liou, H.-C. (2005). Effects of web-based concordancing instruction on EFL students' learning of verb–noun collocations. *Computer Assisted Language Learning, 18*(3), 231–251. doi: 10.1080/09588220500185769

Chapelle, C. A. (1998). Construct definition and validity inquiry in SLA research. In L. F. Bachman & A. D. Cohen (Eds.), *Interfaces between second language acquisition and language Testing research* (pp. 32–70). Cambridge: Cambridge University Press.

Chapelle, C. A. (2008). The TOEFL validity argument. In C. Chapelle, M. Enright & J. Jamieson (Eds.), *Building a validity argument for the Test of English as a Foreign Language* (pp. 319–352). London: Routledge.

Chapelle, C. A. (2012). Validity argument for language assessment: The framework is simple. *Language Testing, 29*(1), 19–27. doi: 10.1177/0265532211417211

Chapelle, C. A., Enright, M. K., & Jamieson, J. (2010). Does an argument-based approach to validity make a difference? *Educational Measurement: Issues and Practice, 29*(1), 3–13. doi: 10.1111/j.1745-3992.2009.00165.x

Chapelle, C. A., Enright, M. K., & Jamieson, J. M. (2008). *Building a validity argument for the Test of English as a Foreign Language*. London: Routledge.

Chen, K. Y. (2011). The impact of EFL students' vocabulary breadth of knowledge on literal reading comprehension. *Asian EFL Journal, 51*, 30–40.

Chen, W. H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., & Revicki, D. A. (2014). Is Rasch model analysis applicable in small sample size pilot studies for assessing item characteristics? An example using PROMIS pain behavior item bank data. *Quality of Life Research, 23*(2), 485–493. doi: 10.1007/s11136-013-0487-5

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89.

Cowie, A. P. (1998). *Phraseology: theory, analysis, and applications*. Oxford: Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. doi: 10.2307/3587951

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge: Cambridge University Press.

Daskalovska, N. (2013). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning*, 1–15. doi: 10.1080/09588221.2013.803982

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Douglas, D. (2000). *Assessing languages for specific purposes*. Cambridge: Cambridge University Press.

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes, 28*(3), 157–169. doi: 10.1016/j.esp.2009.02.002

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *International Review of Applied Linguistics in Language Teaching, 47*(2). doi: 10.1515/iral.2009.007

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Engelhard, J. G. (2013). *Invariant measurement: Using Rasch models in the social, behavioral, and health sciences*. New York, NY: Routledge

Firth, J. R. (1957). *A synopsis of linguistic theory, 1930–1955.* Oxford: Basil Blackwell.

Furr, R. M., & Bacharach, V. R. (2014). *Psychometrics: an introduction*. Thousand Oaks: Sage.

Ganji, M. (2012). On the effect of gender and years of instruction on franian EFL learners' collocational competence. *English Language Teaching, 5*(2), 123–133. doi: 10.5539/elt.v5n2p123

Gao, Y.-m., & Zhang, Y. (2009). A tentative corpus-based study of collocations acquisition by chinese english language learners. *Canadian Social Science, 1*(3), 105–112. doi: http://dx.doi.org/10.3968%2Fj.css.1923669720050103.016

Gitsaki, C. (1999). *Second language lexical acquisition: A study of the development of collocational knowledge*. San Francisco, CA: International Scholars Publications.

Goudarzi, Z., & Momi, M. R. (2012). The effect of input enhancement of collocations in reading on collocation learning and retention of EFL learners. *International Education Studies, 5*(3), 247–258. doi: DOI: 10.5539/ies.v5n3p247

Gyllstad, H. (2005). Words that go together well: Developing test formats for measuring learner knowledge of English collocations. *The Department of English in Lund: Working Papers in Linguistics, 5*, 1–31.

Gyllstad, H. (2007). *Testing English collocations: Developing receptive tests for use with advanced Swedish learners*. Lund: Lund University, Media-Tryck.

Haastrup, K., & Henriksen, B. (2000). Vocabulary acquisition: Acquiring depth of knowledge through network building. *International Journal of Applied Linguistics, 10*(2), 221–240.

Haladyna, T. M. (1994). *Developing and validating multiple-choice test items*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–333. doi: 10.1207/s15324818ame1503_5

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston Kluwer Nighoff

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hashemi, M., Azizinezhad, M., & Dravishi, S. (2012a). Collocation a neglected aspect in teaching and learning EFL. *Procedia - Social and Behavioral Sciences, 31*, 522–525. doi: 10.1016/j.sbspro.2011.12.097

Hashemi, M., Azizinezhad, M., & Dravishi, S. (2012b). The investigation of collocational errors in university students' writing majoring in English. *Procedia - Social and Behavioral Sciences, 31*, 555–558. doi: 10.1016/j.sbspro.2011.12.102

Hoey, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics, 19*(1), 24-44. doi: 10.1093/applin/19.1.24

Howell, D. C. (2008). Best practices in the analysis of variance. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 341–357). Thousand Oaks, CA: Sage.

Howell, D. C. (2013). *Statistical methods for psychology* (8th ed.). Belmont, CA: Wadsworth Cengage Learning.

Hsu, J.-y. T. (2007). Lexical collocations and their relation to the online writing of Taiwanese college English majors and non-English majors. *Electronic journal of foreign language teaching, 4*(2), 192–209.

Hsu, J.-y. T. (2010). The effects of collocation instruction on the reading comprehension and vocabulary learning of taiwanese college English majors. *Asian EFL Journal, 12*(1), 47–87.

Hsu, J.-y. T., & Chiu, C.-y. (2008). Lexical collocations and their relation to speaking proficiency of college EFL learners in Taiwan. *Asian EFL Journal, 10*(1), 181–204.

Iramaneerat, C., Smith, E. V., & Smith, R. M. (2008). An introduction to Rasch measurement. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 50–70). Thousand Oaks, CA: Sage.

Jaén, M. M. (2007). A corpus-driven design of a test for assessing the ESL collocational competence of University students. *IJES, International Journal of English Studies, 7*(2), 127–148.

Jenkins, J. (2007). *English as a lingua franca: Attitude and identity*. Oxford: Oxford University Press.

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

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education and Praeger.

Kane, M. T. (2011). Validating score interpretations and uses. *Language Testing, 29*(1), 3–17. doi: 10.1177/0265532211417210

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*(1), 1–73. doi: 10.1111/jedm.12000

Keshavarz, M. H., & Salimi, H. (2007). Collocational competence and cloze test performance: A study of Iranian EFL learners. *International Journal of Applied Linguistics, 17*(1), 81–92.

Kim, D. H. (2008). *A study on the use of lexical collocations of Korean heritage learners: Identifying the sources of errors.* (1464272 M.A.), University of Southern California, Ann Arbor. ProQuest Dissertations & Theses Global database.

Laufer, B. (2011). The contribution of dictionary use to the production and retention of collocations in a second language. *International Journal of Lexicography, 24*(1), 29–49.

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. doi: 10.1111/j.1467-9922.2010.00621.x

Le, H. (2011). *Developing a validity argument for the English Placement Listening Fall 2010 test at Iowa State University.* (1498791 M.A.), Iowa State University, Ann Arbor. ProQuest Dissertations & Theses Global database.

LeBaron Wallace, T. (2011). An argument-based approach to validity in evaluation. *Evaluation, 17*(3), 233–246. doi: 10.1177/1356389011410522

Lewis, M., & Conzett, J. (2000). *Teaching collocation: Further developments in the lexical approach*. Hove: Language Teaching.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement Transactions, 7*(4), 328.

Linacre, J. M. (2012). *A user's guide to Winsteps Ministeps Rasch-model computer programs.* Retrieved from http://www.winsteps.com/a/winsteps-manual.pdf

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton: Educational Testing Service.

Marco, L., & José, M. (2011). Exploring atypical verb+noun combinations in learner technical writing. *International Journal of English Studies, 11*(2), 77–96.

McKay, S. L., & McKay, S. (2002). *Teaching English as an international language: An introduction to the role of English as an international language and its implications for language teaching*. Oxford: Oxford University Press.

McNamara, T., & Knoch, U. (2012). The Rasch wars: The emergence of Rasch measurement in language testing. *Language Testing, 29*(4), 555–576. doi: 10.1177/0265532211430367

Menon, S., & Mukundan, J. (2012). Collocations of high frequency noun keywords in prescribed science textbooks. *International Education Studies, 5*(6), 149–160. doi: 10.5539/ies.v5n6p149

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. doi: 10.2307/1176219

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2008). *Measurement and assessment in teaching* (10th ed.). Upper Saddle River, NJ: Pearson.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.

Miyakoshi, T. (2009). *Investigating ESL learners' lexical collocations: The acquisition of verb + noun collocations by Japanese learners of English.* (3367916 Ph.D.), University of Hawai'i at Manoa, Ann Arbor. ProQuest Dissertations & Theses Global database.

Molina-Plaza, S., & de Gregorio-Godeo, E. (2010). Stretched verb collocations with give: their use and translation into Spanish using the BNC and CREA corpora. *ReCALL, 22*(02), 191–211. doi: 10.1017/s0958344010000078

Molinaro, N., Canal, P., Vespignani, F., Pesciarelli, F., & Cacciari, C. (2013). Are complex function words processed as semantically empty strings? A reading time and ERP study of collocational complex prepositions. *Language and Cognitive Processes, 28*(6), 762–788. doi: 10.1080/01690965.2012.665465

Nation, I. S. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242. doi: 10.1093/applin/24.2.223

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam: John Benjamins Publishing.

O'Dell, F., & McCarthy, M. (2009). *English collocations in use: Advanced*. Cambridge: Cambridge University Press.

Oller, J. W. (2012). Grounding the argument-based framework for validating score interpretations and uses. *Language Testing, 29*(1), 29–36. doi: 10.1177/0265532211417212

Pardo-Ballester, C. (2010). The validity argument of a web-based spanish listening exam: Test usefulness evaluation. *Language Assessment Quarterly, 7*(2), 137–159. doi: 10.1080/15434301003664188

Qian, D. D. (1999). Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 56*(2), 282–308. doi: 10.3138/cmlr.56.2.282

Qian, D. D. (2002). Investigating the relationship between vocabulary knowledge and academic reading performance: An assessment perspective. *Language Learning, 52*(3), 513–536. doi: 10.1111/1467-9922.00193

Rahimi, M., & Momeni, G. (2012). The effect of teaching collocations on English language proficiency. *Procedia - Social and Behavioral Sciences, 31*, 37–42. doi: http://dx.doi.org/10.1016/j.sbspro.2011.12.013

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Chicago, IL: University of Chicago Press.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Chicago, IL: University of Chicago Press

Read, J. (2000). *Assessing vocabulary*. Cambridge: University Press Cambridge.

Read, J. (2007). Second language vocabulary assessment: current practices and new directions. *International Journal of English Studies, 7*(2), 105–126. doi: 10.6018/ijes.7.2.49021

Read, J., & Chapelle, C. A. (2001). A framework for second language vocabulary assessment. *Language Testing, 18*(1), 1–32. doi: 10.1177/026553220101800101

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational and Behavioral Statistics, 4*(3), 207–230.

Reckase, M. D. (2009). *Multidimensional item response theory*. New York, NY: Springer.

Reynolds, C. R., Livingston, R. B., & Willson, V. L. (2008). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, NJ: Pearson

Sadeghi, K. (2009). Collocational differences between L1 and L2: Implications for EFL learners and teachers. *TESL Canada Journal, 26*(2), 100–124.

Schmitt, N. (2004a). *Formulaic sequences: Acquisition, processing, and use*. Amsterdam: John Benjamins.

Schmitt, N. (2004b). *Vocabulary in language teaching* (4th ed.). Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Basingstoke: Palgrave Macmillan.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing, 18*(1), 55–88. doi: 10.1177/026553220101800103

Schumacker, R. E. (2004). Rasch measurement using dichotomous scoring. *Journal of applied measurement, 5*(3), 328–349.

Schumacker, R. E., & Smith, E. V. (2007). A Rasch perspective. *Educational and Psychological Measurement, 67*(3), 394–409. doi: 10.1177/0013164406294776

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sireci, S. G. (2007). On validity theory and test validation. *Educational Researcher, 36*(8), 477–481. doi: 10.3102/0013189x07311609

Sireci, S. G. (2013). Agreeing on validity arguments. *Journal of Educational Measurement, 50*(1), 99–104. doi: 10.1111/jedm.12005

Skrzypek, A. (2009). Phonological short-term memory and L2 collocational development in adult learners. *EUROSLA Yearbook, 9*, 160–184. doi: 10.1075/eurosla.9.09skr

Smith, E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of applied measurement, 6*(2), 147–163.

Sonbul, S., & Schmitt, N. (2013). Explicit and mplicit lexical knowledge: acquisition of collocations under different input conditions. *Language Learning, 63*(1), 121–159. doi: 10.1111/j.1467-9922.2012.00730.x

Sowden, C. (2012). ELF on a mushroom: the overnight growth in English as a Lingua Franca. *ELT journal, 66*(1), 89–96.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Voss, E. (2012). *A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design.* (3539432 Ph.D.), Iowa State University, Ann Arbor. ProQuest Dissertations & Theses Global database.

Walker, C. P. (2011a). A corpus-based study of the linguistic features and processes which influence the way collocations are formed: Some implications for the learning of collocations. *TESOL Quarterly, 45*(2), 291–312. doi: 10.5054/tq.2011.247710

Walker, C. P. (2011b). How a corpus-based study of the factors which influence collocation can help in the teaching of business English. *English for Specific Purposes, 30*(2), 101–112. doi: http://dx.doi.org/10.1016/j.esp.2010.12.003

Waugh, C. K., & Gronlund, N. E. (2013). *Assessment of student achievement*. Upper Saddle River, NJ: Pearson.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition, 27*(01), 33–52. doi: doi:10.1017/S0272263105050023

Webb, S. (2008). Receptive and productive vocabulary sizes of L2 learners. *Studies in Second Language Acquisition, 30*(01), 79–95. doi: doi:10.1017/S0272263108080042

Webb, S., & Kagimoto, E. (2011). Learning collocations: Do the number of collocates, position of the node word, and synonymy affect learning? *Applied Linguistics, 32*(3), 259–276. doi: 10.1093/applin/amq051

Webb, S., Newton, J., & Chang, A. (2013). Incidental learning of collocation. *Language Learning, 63*(1), 91–120. doi: 10.1111/j.1467-9922.2012.00729.x

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum Associates.

Wolfe, E. W., & Smith, E. V. (2007a). Instrument development tools and activities for measure validation using Rasch models: Part I-Instrument development tools. *Journal of applied measurement, 8*(1), 97–123.

Wolfe, E. W., & Smith, E. V. (2007b). Instrument development tools and activities for measure validation using Rasch models: Part II-Validation activities. *Journal of applied measurement, 8*(2), 204–234.

Wolter, B., & Gyllstad, H. (2011). Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge. *Applied Linguistics, 32*(4), 430–449. doi: 10.1093/applin/amr011

Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: Oxford University Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: Mesa Press.

Wright, B. D., & Masters, G. N. (2002). Number of person or item strata. *Rasch Measurement Transactions, 16*(3), 888.

Yamamoto, Y. M. (2011). Bridging the gap between receptive and productive vocabulary size through extensive reading. *Reading Matrix: An International Online Journal, 11*(3), 226–242.

APPENDICES

Appendix A. Test specification of the ACCT

Component	Characteristic	Description
Test purpose	Purpose	The purpose of the test is to provide scores that are meaningfully interpreted as an indicator of academic English collocational competence and used for placement decision in university or other academic institutions at tertiary level.
Construct to be measured	Construct	The construct to be measured is the collocational competence defined based on the interactionalist approach which views performance as a sign of underlying traits and is influenced by the context under which it occurs. The underlying trait or competence in this test is thus presumed to be the academic verb-noun collocational competence demonstrated in the context of academic written English.
Situation	Participants	Learners of English as a foreign language who are studying at a university or an academic institution at tertiary level.
	Content	Published and unpublished texts consisting of seven academic disciplinary areas, Applied Science, Art, Belief & Thought, Commerce & Finance, Natural & Pure Sciences, Social Science, World Affairs
	Setting	University or academic institutions at tertiary level
	Purpose	Primarily representational: Conveying meaning about academic topics and content
	Register	Formal written academic English
Content of the test	Number of the tasks	One multiple-choice task
	Number of the questions	30 questions
	Time allowance	30 minutes
	Test level	Mixed levels of English language proficiency
Text collocation materials	Language Features	High-frequency restricted verb-noun collocations
	Pragmatic Features	Ideational functions: to express or exchange information about ideas and knowledge
	Discourse Features	Knowledge of genre, register, and collocation

Component	Characteristic	Description
Rubric	Instruction	The test is designed to measure your ability to recognise verb-noun collocations used in academic written English. The test consists of 50 multiple-choice items. You have 60 minutes to complete the tasks.
	Direction	Test questions are incomplete sentences. Beneath each sentence, you will see five verbs, marked a, b, c, d, and e. Read each question carefully and choose the one verb that best complete the sentence with an appropriate meaning for the academic context. Circle the letter of the answer you have selected.
	Response Formats	Best-answer multiple-choice: Select the most appropriate verb to complete the appropriate collocation with an appropriate meaning for the academic context.
	Rules for Scoring	Best-answer multiple-choice: Selecting best-answer choice based on the target collocates identified in the collocation identification process from academic written sub-corpora in BNC receives full mark (1). Selecting alternatives responses receive no mark (0).
Administration	Delivering format	The test is a paper-based test format with no answer sheet.
	Test taker	Test-takers answer the questions on the ACCT form. Test-takers have five minutes to read the instruction and direction of the test before doing the test.
	Test user	Teachers or test users administer the test, clarify the instruction and direction of the test and monitor test-takers during the testing period.

Appendix B. Results of evaluated sentences from BNC

No	Sentences	% in Longman 3,000 keywords		Flesch-Kincaid Readability index	
		Word count	%	Grade level	Ease score
1.	The study was approved by the local Ethics Committee and informed written consent was obtained from all subjects.	18	88.9	12	38
2.	It is widely assumed that these scoring systems can be used for comparisons.	13	100	11	37
3.	In either case it is feasible to collect data about a trainee's performance and modify the presented information accordingly.	20	75	17	5
4.	However, it is plausible to imagine that class background has a causal effect on the type of school that the son attends.	22	90.9	12	50
5.	The women actually had rather strong ties, since they spent much of their time doing household tasks communally outside.	19	86.5	12	45
6.	The relationships crossing this terrain take specific forms in specific societies and must be analysed in that context.	18	88.9	14	29
7.	The emotive function uses words to evoke subjective feelings or attitudes, by means of the associations that words carry with them.	21	85.7	15	24
8.	Linguists, psychologists and AI workers have theories about what procedures might select the right sense on the basis of sensible rules and reject the wrong one.	26	84.6	15	34
9.	A similar mechanism may perhaps account for the fact that some group-living animals drive sick or injured individuals out of the group.	23	95.7	14	33
10.	That is why what is validly required by a legitimate authority is one's duty, even where previously it was merely something one had sufficient reason to do.	28	85.7	18	13
11.	It is in virtue of such rules that we can make sense of the idea that we are objectively correct to call the new sensation a pain.	27	88.9	13	51
12.	Cases decided under the 1973 Act may therefore not be a reliable guide to reasonableness under the 1977 Act.	19	78.9	14	30
13.	Their power to admit and expel members has the important consequence of granting and revoking authority to carry on investment business.	21	90.5	16	20
14.	And philosophers talk of ‘sensations’ in this connection because of views they hold about perception.	15	80	12	39
15	During our investigation of these patients, we diagnosed seven new cases of cancer of the prostate.	26	81.3	14	27
16.	This species has been recorded from the Atlantic, Indian and Pacific Oceans.	12	75	11	40
17.	The history of theories of electricity provides an example of the changing fortunes of rival research programmes.	17	82.4	15	15
18.	Prescribing is one possible treatment option; others include counselling, educating patients on self-limiting illnesses, and changes in lifestyle to improve health.	22	81.8	20	-8
19.	Meanwhile an interim award of £1 was made to full-time workers; part-	17	88.2	12	34

No	Sentences	% in Longman 3,000 keywords		Flesch-Kincaid Readability index	
		Word count	%	Grade level	Ease score
	time workers got nothing.				
20.	It was not concerned with the position of local authorities which have the function of enforcing the law in their districts in the public interest.	25	96	14	43
21.	The Act of 1975 was passed in May 1975, but only took effect from 4 May 1976.	17	76.5	5	90
22.	The government was elected in October 1974 with an overall majority of three.	13	69.2	11	37
23.	Those groups have brought pressure to bear on government to provide resources or pursue policies to the benefit of their members.	21	95.2	13	36
24.	In 1986, as part of its wider proposals for the reform of local government finance, the government declared its intention to introduce a new grant system.	26	84.6	15	31
25.	Various groups within the party were formed to try and achieve particular objectives.	13	100	13	24
26.	In the subgroup of patients who are treated with two injections daily, adequate gall bladder concentration might also be achieved after dinner.	22	72.7	15	27
27.	The control group was treated with an oral triple therapy regimen which had previously been evaluated in a pilot study.	20	80	13	34
28.	It is evident that the larger and more popular temples may have played a considerable part in the economy of any province.	22	81.8	15	27
29.	Twenty three children had more severe but intermittent symptoms and nine had chronically severe disease throughout the year.	18	66.7	15	19
30.	All results are expressed as the median value with the range to indicate dispersion.	14	85.7	13	29
31.	This personal and inner satisfaction give way to taking account of the viewpoints of others, and the experience of others.	20	90	13	38
32.	Perhaps it is not in the grammar books because the grammar books do not reflect how people actually use language.	20	95	11	51
33.	It would be possible to test a large sample of readers, who would read nine texts and place them on the network.	22	95.5	10	66
34.	Very little work has been done in accounting for the development of an individual dramatic character in pragmatic or discourse terms.	21	85.7	15	24
35.	There is no general nature in common to those things, and any idea we have is never general or abstract, but always of some particular thing.	26	96.2	15	35
36.	All agree that some of our beliefs are justified by their relation to other beliefs. Belief and Thought	18	94.4	10	52
37.	And there is a particularly close connection in the case we are considering.	13	100	13	23
38.	Under a contract of sale, breach of condition by the seller allows the buyer to reject the goods (if delivered) and terminate the contract.	24	83.3	13	41
39.	These are the terms implied in a contract where the supplier has	25	80	15	27

No	Sentences	% in Longman 3,000 keywords		Flesch-Kincaid Readability index	
		Word count	%	Grade level	Ease score
	agreed to carry out a service (paragraph 8–09 above). Commerce and Finance				
40.	If the parties agree on a procedure and the expert does not, the parties should appoint another expert.	18	100	11	48
41.	It is submitted that a number of cases which in the past applied the literal rule would now be decided in the opposite way. Commerce and Finance	27	92.6	15	38
42.	We conducted a two-year study to assess the effectiveness of the family smoking education and smoking and me projects in influencing smoking behaviour.	24	83.3	16	18
43.	The subject has been reviewed (White et al, 1981) and will be briefly described here (see Figure 5.1).	19	68.4	6	66
44.	We do not have space for a full description of all the experimental techniques used in obtaining the results discussed in this book.	23	95.7	14	36
45.	The following additional cases were cited in argument in the Court of Appeal. Social Science	15	93.3	12	34
46.	After leaving school, most of his friends moved away to university.	11	100	9	50
47.	This means that lecturers and tutors will have to find ways of connecting with their students' outlooks.	17	82.4	8	65
48.	These will cover areas such as equal opportunities, multi-cultural education, cross-curricular themes, competences and dimensions and special needs.	20	80	20	-18
49.	Research would inevitably concentrate on informal relations and social structures through which power is exercised.	15	86.7	17	-0
50.	That unemployment fell as a result of war is an undeniable fact, but it was not the primary reason for the decision to fight the war.	26	96.2	13	51
	Overall	725	86.6	14	28

Appendix C. Research instruments

The Academic Collocational Competence Test

Please give your factual background information (กรุณากรอกข้อมูลภูมิหลังตามความเป็นจริง)

1) Age (อายุ)	_____ years		
2) Gender (เพศ)	☐ Male	☐ Female	
3) Native language (ภาษาแม่)	☐ Thai	☐ Other _____	
4) Field of study (สาขาวิชาที่ศึกษา)	_____		
5) Study level (ระดับที่กำลังศึกษา)	☐ Bachelor	☐ Master	☐ Doctorate
6) English score (คะแนนภาษาอังกฤษ)	☐ CU-TEP_____		☐ IELTS _____
	☐ TOEFL PBT_____		☐ TOEFL iBT _____

Instruction

The test is designed to measure your ability to recognise **verb-noun collocations** used in academic written English. The test consists of **30 multiple-choice items**. You have **30 minutes** to complete the test.

Direction

Test questions are incomplete sentences. Read each question carefully and *choose a verb that best collocates with a **noun*** in boldface to complete the sentence with the most appropriate meaning for academic written English. Circle the letter of the answer you have selected.

Look at the following example

0) He _____ the English placement **test** and was put in the most advanced class.

a. failed (b.) passed c. won d. did e. took

In case you want to change your answer, put a cross (X) on the circle and then circle another letter of the new answer.

Look at the following example

0) He _____ the English placement **test** and was put in the most advanced class.

a. failed (b.) passed c. won d. did (X) took

1) This means that lecturers and tutors will have to _____ **ways** of connecting with their students' outlooks.
   a. pave b. give c. lose d. make e. find
2) The following additional **cases** were _____ in argument in the Court of Appeal.
   a. cited b. diagnosed c. decided d. shown e. settled
3) After _____ **school**, most of his friends moved away to university.
   a. starting b. leaving c. teaching d. entering e. abandoning
4) It was not concerned with the position of local authorities which have the function of _____ the **law** in their districts in the public interest.
   a. enforcing b. exploiting c. spreading d. imposing e. utilising
5) Meanwhile an **award** of £1 was _____ to full-time workers**;** part-time workers got nothing.
   a. done b. built c. made d. formed e. created

6) These will _____ **areas** such as equal opportunities, multi-cultural education, cross-curricular themes, competences and dimensions and special needs.
a. engage b. search c. cover d. incorporate e. integrate

7) The subject has been reviewed (White et al, 1981) and will be briefly described here (_____ **Figure** 5.1).
a. show b. notice c. observe d. see e. look

8) We do not have space for a full description of all the experimental techniques used in _____ the **results** discussed in this book.
a. obtaining b. earning c. expressing d. manifesting e. exhibiting

9) The history of theories of electricity _____ an **example** of the changing fortunes of rival research programmes.
a. generates b. provides c. expresses d. produces e. exhibits

10) Prescribing is one possible treatment option; others include counselling, educating patients on self-limiting illnesses, and changes in lifestyle to _____ **health.**
a. damage b. renovate c. impair d. enhance e. improve

11) We _____ a two-year **study** to assess the effectiveness of the family smoking education and my projects in influencing smoking behaviour.
a. conducted b. committed c. operated d. performed e. produced

12) There is no general nature in common to those things, and any **idea** we _____ is never general or abstract, but always of some particular thing.
a. have b. believe c. make d. create e. build

13) It is by virtue of such rules that we can _____ **sense** of the idea that we are objectively correct to call the new sensation a pain.
a. make b. create c. build d. take e. form

14) All agree that some of our **beliefs** are _____ by their relation to other beliefs.
a. nullified b. identified c. justified d. purified e. falsified

15) And philosophers talk of 'sensations' in this connection because of **views** they _____ about perception.
a. hear b. carry c. hold d. make e. get

16) A similar mechanism may perhaps _____ the **fact** that some group-living animals drive sick or injured individuals out of the group.
a. search for b. argue for c. look for d. account for e. find out

17) It is evident that the larger and more popular temples may have _____ a considerable **part** in the economy of any province.
a. shown b. played c. taken d. done e. made

18) Those groups have brought pressure to bear on government to provide resources or _____ **policies** to the benefit of their members.
a. progress b. produce c. purchase d. pursue e. persuade

19) That unemployment fell as a result of war is an undeniable fact, but it was not the primary reason for the decision to _____ the **war**.
a. operate b. conduct c. produce d. fight e. perform

20) Research would inevitably concentrate on informal relations and social structures through which **power** is _____.
a. executed b. exercised c. produced d. expressed e. harnessed

21) In 1986, as part of its wider proposals for the reform of local government finance, the government declared its intention to _____ a new grant **system.**
a. renovate b. integrate c. introduce d. invent e. install

22) It has been said that these **rules** will be _____ less stringently to a commercial contract than to other types of document.
a. applied b. functioned c. spent d. respected e. violated

23) Their power to admit and expel members has the important consequence of granting and revoking authority to _____ investment **business.**
a. draw on b. carry on c. take on d. keep on e. go on

24) If the parties agree on a procedure and the expert does not, the parties should _____ another **expert.**
a. establish b. constitute c. dismiss d. expel e. appoint

25) Under a contract of sale, breach of condition by the seller allows the buyer to reject the goods and _____ the **contract**.
a. extend b. violate c. terminate d. eradicate e. abandon

26) Here are some **words** which are commonly _____ in essay.
a. used b. made c. spent d. spelled e. taken

27) Very little **work** has been _____ in accounting for the development of an individual dramatic character in pragmatic or discourse terms.
a. made b. devised c. built d. invented e. done

28) Many of these **texts** can be _____ as elaborate commentaries on the nature of writing and reading.
a. written b. posted c. pasted d. read e. typed

29) Twenty three children had more severe but intermittent symptoms and nine _____ severe **disease** throughout the year.
a. had b. held c. attached d. contained e. contacted

30) The control **group** was _____ with an oral triple therapy regimen which had previously been evaluated in a pilot study.
a. dealt b. handled c. fixed d. helped e. treated

Academic Vocabulary Level Test (Version 2)

Direction

This is a vocabulary test. You must choose the right word to go with each meaning. Write the number of that word next to its meaning. Here is an example.

You answer it in the following way.

1 business		
2 clock	___*6*___	part of a house
3 horse	___*3*___	animal with four legs
4 pencil	___*4*___	something used for writing
5 shoe		
6 wall		

Some words are in the test to make it more difficult. You do not have to find a meaning for these words. In the example above, these words are *business, clock,* and *shoe*.

If you have no idea about the meaning of a word, **do not guess.** But if you think you might know the meaning, then you should try to find the answer. You have **30 minutes** to complete the test.

1 area			
2 contract	_____	written agreement	(1)
3 definition	_____	way of doing something	(2)
4 evidence	_____	reason for believing something is or is not true	(3)
5 method			
6 role			
1 adult			
2 exploitation	_____	end	(4)
3 infrastructure	_____	machine used to move people or goods	(5)
4 schedule	_____	list of things to do at certain times	(6)
5 termination			
6 vehicle			
1 debate			
2 exposure	_____	plan	(7)
3 integration	_____	choice	(8)
4 option	_____	joining something into a whole	(9)
5 scheme			
6 stability			
1 alter			
2 coincide	_____	change	(10)
3 deny	_____	say something is not true	(11)
4 devote	_____	describe clearly and exactly	(12)
5 release			
6 specify			
1 access			
2 gender	_____	male or female	(13)
3 implementation	_____	study of the mind	(14)
4 license	_____	entrance or way in	(15)
5 orientation			
6 psychology			

1 correspond		
2 diminish	_____ keep	(16)
3 emerge	_____ match or be in agreement with	(17)
4 highlight	_____ give special attention to something	(18)
5 invoke		
6 retain		
1 accumulation		
2 edition	_____ collecting things over time	(19)
3 guarantee	_____ promise to repair a broken product	(20)
4 media	_____ feeling a strong reason or need to do something	(21)
5 motivation		
6 phenomenon		
1 bond		
2 channel	_____ make smaller	(22)
3 estimate	_____ guess the number or size of something	(23)
4 identify	_____ recognizing and naming a person or thing	(24)
5 mediate		
6 minimize		
1 explicit		
2 final	_____ last	(25)
3 negative	_____ stiff	(26)
4 professional	_____ meaning ‘no’ or ‘not’	(27)
5 rigid		
6 sole		
1 abstract		
2 adjacent	_____ next to	(28)
3 controversial	_____ added to	(29)
4 global	_____ concerning the whole world	(30)
5 neutral		
6 supplementary		

Test Reflection Questionnaire

1) Were you thinking about academic English as you took **the Academic Collocational Competence Test**?
ในขณะที่คุณกำลังทำ**แบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ**คุณได้นึกถึงภาษาอังกฤษเชิงวิชาการหรือไม่
☐ Yes (ใช่) ☐ No (ไม่ใช่) ☐ I don't know (ฉันไม่รู้)

2) Do you think the English in **the Academic Collocational Competence Test** is similar to academic English used in university textbooks?
คุณคิดว่าภาษาอังกฤษใน**แบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ**คล้ายกับภาษาอังกฤษเชิงวิชาการที่ใช้ในตำราเรียนในมหาวิทยาลัยหรือไม่
☐ Yes (ใช่) ☐ No (ไม่ใช่) ☐ I don't know (ฉันไม่รู้)
* If "*Yes or No*", then answer question 3 (ถ้าตอบ "*ใช่ หรือ ไม่ใช่*" ให้ตอบข้อ 3)

3) Please explain how the English in **the Academic Collocational Competence Test** is similar to or different from English used in university textbooks. (you can reply in Thai)
กรุณาอธิบายว่าภาษาอังกฤษใน**แบบสอบสามัตถิยะคำปรากฏร่วมเชิงวิชาการ**คล้ายหรือต่างกับภาษาอังกฤษเชิงวิชาการที่ใช้ในตำราเรียนในมหาวิทยาลัยอย่างไร (ตอบเป็นภาษาไทยได้)
_____
_____
_____
_____

Appendix D. Summary of expert item evaluation results

	Question 1: stem			Question 1: best choice			Question 1: distractors		
Item	Expert 1	Expert 2	Expert 3	Expert 1	Expert 2	Expert 3	Expert 1	Expert 2	Expert 3
1	✓	✓	✓	✓	✓	✓	✓	✓	✓
2	✓	✓	✓	✓	✓	✓	✓	✓	✓
3	✓	✓	✓	✓	✓	✓	✓	✓	✓
4	✓	✓	✓	✓	✓	✓	✓	✓	✓
5	✓	✓	✓	✓	✓	✓	✓	✓	✓
6	X	✓	X	✓	✓	X	✓	✓	✓
7	X	✓	X	✓	✓	✓	X	✓	✓
8	✓	✓	X	✓	✓	X	✓	✓	✓
9	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	X	✓	X	✓	✓	✓	✓	✓	✓
11	X	✓	✓	✓	✓	✓	✓	✓	✓
12	✓	X	✓	✓	✓	X	✓	✓	✓
13	✓	✓	X	✓	✓	X	✓	X	✓
14	✓	✓	✓	✓	✓	✓	✓	✓	✓
15	✓	✓	✓	✓	X	✓	✓	✓	✓
16	X	X	X	✓	✓	X	✓	✓	X
17	✓	✓	✓	✓	✓	✓	✓	✓	✓
18	✓	✓	✓	✓	✓	✓	X	✓	✓
19	✓	✓	✓	✓	X	✓	✓	✓	✓
20	X	✓	✓	✓	✓	✓	✓	✓	✓
21	X	✓	✓	✓	✓	✓	✓	✓	✓
22	✓	✓	✓	✓	✓	✓	✓	✓	✓
23	✓	X	✓	✓	✓	✓	X	✓	✓
24	✓	✓	✓	✓	✓	✓	✓	✓	✓
25	✓	✓	✓	✓	✓	✓	✓	✓	✓
26	✓	✓	✓	✓	✓	✓	✓	✓	✓
27	✓	✓	✓	✓	✓	✓	✓	✓	✓
28	✓	✓	✓	✓	✓	✓	✓	✓	✓
29	✓	✓	✓	✓	X	✓	✓	✓	✓
30	✓	X	✓	✓	X	✓	✓	✓	✓
31	✓	X	X	✓	✓	✓	✓	✓	✓
32	X	X	✓	✓	✓	✓	✓	✓	✓
33	X	✓	✓	X	✓	✓	X	X	X
34	✓	✓	✓	✓	✓	✓	✓	✓	✓
35	✓	✓	✓	✓	✓	✓	✓	✓	✓
36	✓	✓	✓	✓	✓	✓	✓	✓	✓
37	X	✓	✓	✓	X	✓	✓	✓	✓
38	✓	✓	✓	✓	✓	✓	✓	✓	✓
39	✓	✓	✓	✓	X	X	✓	✓	✓
40	✓	✓	✓	✓	✓	✓	✓	✓	✓
41	X	X	✓	✓	✓	✓	✓	✓	X
42	✓	✓	X	✓	✓	✓	✓	✓	✓
43	✓	✓	✓	✓	✓	✓	✓	✓	✓
44	✓	✓	✓	✓	✓	✓	✓	✓	✓
45	✓	X	✓	✓	✓	✓	✓	✓	✓
46	✓	✓	✓	✓	✓	✓	✓	✓	X
47	✓	✓	✓	✓	✓	✓	✓	✓	X
48	✓	✓	✓	✓	✓	✓	✓	✓	X
49	✓	✓	✓	✓	✓	✓	✓	✓	✓
50	✓	✓	✓	✓	✓	✓	✓	✓	✓

Appendix E. Summary of related studies

Variable / Authors	Jaen (2007)	Keshavarz and Salimi (2007)	Webb and Kagomoto (2009)	Webb, Newton, and Chang (2013)	Sonbul and Schmitt (2013)
Title	A corpus-driven design of the test for assessing the ESL collocational competence of university students	Collocational competence and cloze test performance: a study of Iranian EFL learners	The effects of vocabulary learning on collocation and meaning	Incidental learning of collocation	Explicit and implicit lexical knowledge acquisition of collocations under different input conditions
Journal source	International Journal of English Studies	International Journal of Applied Linguistics	TESOL Quarterly	Language Learning	Language Learning
Collocation knowledge	Receptive and productive collocation aspect	Receptive and productive collocation aspect	Receptive and productive collocation aspect	Receptive and productive collocation aspect	Explicit and implicit collocation aspect
Collocation type	Lexical adjective-noun collocation	Grammatical and lexical collocation	Lexical verb-noun collocation	Lexical verb-noun collocation	Lexical adjective-noun collocation
Target language use	General English	General English	General English	General English	Specific medical English
Item input source	Bank of English and BNC	Not reported	Bank of English and BNC	Bank of English	Textbooks
Test taker characteristic	ESL student of English applied linguistics	Iranian EFL university students	Japanese EFL university students	Taiwanese EFL university students	Native-English undergraduate students
Proficiency level	Advanced	Intermediate	Not clear	Not clear	Not clear
Item response format	A multiple-choice format and a gap-filling format	An open-ended close format and a multiple-choice close format	A multiple-choice format, a close-test format and a productive and receptive translation format	A multiple-choice, a gap-filling, and a translation format	An explicit multiple-choice format, an explicit close test format, and an implicit priming test format
Test delivery format	A paper and pencil administration	A paper and pencil administration	A paper and pencil administration	A paper and pencil administration	A paper and pencil administration
Test scoring method	Dichotomous scoring	Dichotomous scoring	Partial credit scoring	Dichotomous and partial credit scoring	Dichotomous scoring
Test quality	Classical test	Classical test	Not reported	Not reported	Not reported

Variable / Authors	Jaen (2007)	Keshavarz and Salimi (2007)	Webb and Kagomoto (2009)	Webb, Newton, and Chang (2013)	Sonbul and Schmitt (2013)
analysis	theory	theory			

Variable / Authors	Voss (2012)	Wolter and Gyllstad (2011)	Sadeghi (2009)	Laufer (2011)	Kim (2009)
Title	A validity argument for score meaning of a computer-based ESL academic collocational ability test based on a corpus-driven approach to test design	Collocational links in the L2 mental lexicon and the influence of L1 intralexical knowledge	Collocational differences between L1 and L2: implications for EFL learners and teachers	The contribution of dictionary use to the production and retention of collocations in a second language	A study of the use of lexical collocations of Korean heritage learners: Identifying the sources of errors
Journal source	Unpublished doctoral dissertation, Iowa State University	Applied Linguistics	TESL CANADA Journal	International Journal of Lexicography	Unpublished master's thesis, University of Southern California
Collocation knowledge	Productive knowledge	Receptive knowledge	Productive knowledge	Productive knowledge	Productive knowledge and comprehension
Collocation type	Verb-noun collocation	Verb-noun collocation	Mixed-collocation types	Verb-noun collocation	Noun-verb collocation
Target language use	Academic written English	General English	General English	general English	General English
Item input source	BNC	BNC	English coursebooks	Selection based on experience	English coursebooks
Test taker characteristic	Chinese ESL university students	Swedish non-native and Native English university students	Iranian EFL university students	Israeli L2 high school learners	Korean EFL students
Proficiency level	Low, moderate and high proficiencies	Not specified	Various levels of proficiency	Pre-intermediate and intermediate proficiency	high-intermediate
Item response format	A gap-filling format	A primed lexical decision task and a receptive multiple-choice test	a multiple-choice format	A gap-filling format	a translation task

Variable / Authors	Voss (2012)	Wolter and Gyllstad (2011)	Sadeghi (2009)	Laufer (2011)	Kim (2009)
Test delivery format	Computer-delivered administration	Computer-delivered administration and a paper-pencil administration	Paper-pencil administration	Paper-pencil administration	Paper-pencil administration
Test scoring method	Dichotomous and partial credit scoring	Dichotomous scoring	Dichotomous scoring	Dichotomous scoring	Dichotomous scoring
Test quality analysis	CTT and Rasch IRT	Not reported	Not reported	Not reported	Not reported

Variable / Authors	Daskalovska (2013)	Skrzypek and Singleton (2013)	Chan and Liou (2005)	Miyakoshi (2009)
Title	Corpus-based versus traditional learning of collocations	Productive knowledge of English collocations in adult Polish learners: The role of short-term memory	Effects of web-based concordancing instruction on EFL students' learning of verb-noun collocations	Investigating ESL learners' lexical collocations: The acquisition of verb + noun collocations by Japanese learners of English
Journal source	Computer Assisted Language Learning	Vigo International Journal of Applied Linguistics	Computer Assisted Language Learning	Unpublished doctoral dissertation, University of Hawaii at Manoa
Collocation knowledge	Productive and receptive knowledge	Productive knowledge	Productive knowledge	Productive and receptive knowledge
Collocation type	Verb-adverb collocation	Lexical and grammatical collocations	Verb-noun collocation	Verb-noun collocation
Target language use	General English	General English	General English	General English
Item input source	Not identified	Gitsaki (1999)	Online lesson materials	Not reported
Test taker characteristic	EFL university students in the Republic of Macedonia	Adult Polish learners of English	Taiwanese EFL students	Japanese EFL students
Proficiency level	Advanced EFL students of English	Elementary and pre-intermediate	Not reported	Intermediate and advanced students
Item response format	Multiple-choice, gap-filling- matching, constructed response formats	Gap-filling format and translation task (Gitsaki (1999)	Gap-filling format	Gap-filling format and multiple choice format
Test delivery format	Paper-pencil administration	Paper-pencil administration	Paper-pencil administration	Paper-pencil administration

Variable / Authors	Daskalovska (2013)	Skrzypek and Singleton (2013)	Chan and Liou (2005)	Miyakoshi (2009)
Test scoring method	Dichotomous and partial credit scoring	Dichotomous scoring	Dichotomous scoring	Dichotomous scoring
Test quality analysis	Not reported	CTT	Not reported	Not reported

Appendix F. Person measure estimation

Examinee ID	Proficiency level	ACCT score	Bayesian ability	Rasch ability	S.E. estimate	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd	PTM correlation
1	H	24	1.550	1.61	0.48	1.2	0.78	1.06	0.28	0.15
2	H	27	2.173	2.50	0.63	0.91	-0.05	0.66	-0.26	0.36
3	H	23	1.366	1.39	0.46	0.93	-0.2	0.93	-0.06	0.39
4	H	22	1.203	1.19	0.44	0.93	-0.24	0.89	-0.23	0.42
5	H	22	1.184	1.19	0.44	0.95	-0.16	1.43	1.23	0.33
6	H	24	1.548	1.61	0.48	0.96	-0.05	0.82	-0.29	0.37
7	H	26	1.939	2.15	0.56	1.05	0.26	1.03	0.26	0.22
8	H	21	1.026	1.00	0.43	0.91	-0.41	0.8	-0.61	0.47
9	H	19	0.720	0.65	0.41	0.91	-0.52	0.87	-0.51	0.47
10	H	22	1.189	1.19	0.44	1.2	0.95	1.73	1.87	0.08
11	H	28	2.459	2.97	0.75	0.8	-0.17	0.33	-0.63	0.46
12	H	24	1.532	1.61	0.48	1	0.07	0.95	0.04	0.32
13	H	28	2.440	2.97	0.75	0.87	-0.03	0.45	-0.4	0.38
14	H	20	0.869	0.82	0.42	0.72	-1.67	0.67	-1.31	0.64
15	H	14	-0.008	-0.15	0.4	0.98	-0.11	0.99	0.01	0.41
16	H	21	1.012	1.00	0.43	1.28	1.4	1.35	1.13	0.06
17	H	25	1.720	1.86	0.52	0.77	-0.69	0.55	-0.85	0.55
18	H	18	0.562	0.48	0.4	1.09	0.59	1.29	1.29	0.26
19	H	23	1.365	1.39	0.46	0.84	-0.63	0.79	-0.45	0.49
20	H	24	1.507	1.61	0.48	0.96	-0.05	0.89	-0.09	0.36
21	H	25	1.718	1.86	0.52	1.04	0.23	0.86	-0.1	0.29
22	H	18	0.550	0.48	0.4	0.91	-0.58	0.84	-0.73	0.49
23	H	25	1.707	1.86	0.52	0.98	0.04	0.96	0.09	0.32
24	H	29	2.754	3.72	1.03	1.07	0.38	1.37	0.71	0.01
25	H	25	1.689	1.86	0.52	0.94	-0.1	0.71	-0.44	0.4
26	H	21	1.003	1.00	0.43	0.86	-0.69	0.8	-0.61	0.5
27	H	20	0.860	0.82	0.42	1.21	1.18	1.19	0.74	0.16
28	H	26	1.960	2.15	0.56	0.92	-0.1	0.77	-0.2	0.36
29	H	24	1.543	1.61	0.48	1.05	0.26	0.86	-0.17	0.31
30	H	27	2.160	2.50	0.63	0.99	0.14	0.72	-0.16	0.29
31	H	22	1.194	1.19	0.44	1.05	0.31	1.03	0.19	0.3
32	H	21	0.994	1.00	0.43	1.27	1.36	1.24	0.85	0.1
33	H	20	0.872	0.82	0.42	1.09	0.53	1.38	1.36	0.24
34	H	27	2.194	2.50	0.63	1.03	0.21	0.79	-0.05	0.25
35	M	19	0.703	0.65	0.41	0.97	-0.15	1.23	0.96	0.36
36	M	23	1.361	1.39	0.46	0.95	-0.16	0.8	-0.42	0.42
37	H	22	1.166	1.19	0.44	0.89	-0.49	1.35	1.04	0.4
38	H	28	2.406	2.97	0.75	1.02	0.24	1.45	0.73	0.12
39	M	21	1.028	1.00	0.43	1.01	0.14	1.52	1.59	0.27
40	H	19	0.706	0.65	0.41	1.07	0.47	1.04	0.26	0.31
41	M	16	0.258	0.16	0.4	1.18	1.25	1.29	1.48	0.19
42	M	21	1.038	1.00	0.43	1.25	1.28	1.18	0.66	0.13
43	H	17	0.413	0.32	0.4	0.97	-0.16	0.92	-0.36	0.42
44	H	25	1.731	1.86	0.52	1.05	0.27	0.96	0.1	0.26

Examinee ID	Proficiency level	ACCT score	Bayesian ability	Rasch ability	S.E. estimate	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd	PTM correlation
45	H	14	-0.007	-0.15	0.4	1.02	0.22	1.04	0.29	0.36
46	M	16	0.260	0.16	0.4	0.79	-1.63	0.72	-1.57	0.61
47	H	21	1.037	1.00	0.43	1.17	0.89	1.56	1.69	0.14
48	H	27	2.193	2.50	0.63	1.06	0.29	1.05	0.32	0.18
49	M	24	1.535	1.61	0.48	1.06	0.32	1.17	0.5	0.22
50	H	20	0.848	0.82	0.42	1.06	0.38	1.03	0.19	0.31
51	H	19	0.712	0.65	0.41	1.14	0.85	1.41	1.6	0.19
52	M	19	0.728	0.65	0.41	0.89	-0.67	0.84	-0.62	0.49
53	H	26	1.934	2.15	0.56	1.03	0.2	1.09	0.35	0.23
54	M	16	0.269	0.16	0.4	1.03	0.25	0.98	-0.05	0.37
55	M	21	1.022	1.00	0.43	1.14	0.75	1.52	1.58	0.17
56	M	22	1.175	1.19	0.44	0.94	-0.22	0.85	-0.36	0.42
57	H	21	1.006	1.00	0.43	1.15	0.8	1.59	1.76	0.14
58	M	16	0.268	0.16	0.4	0.83	-1.3	0.78	-1.23	0.57
59	M	21	1.015	1.00	0.43	0.81	-1	0.71	-0.95	0.56
60	H	22	1.170	1.19	0.44	0.85	-0.67	0.75	-0.7	0.51
61	M	17	0.397	0.32	0.4	0.79	-1.48	0.73	-1.4	0.6
62	M	13	-0.158	-0.31	0.4	0.89	-0.75	0.8	-1.01	0.51
63	M	12	-0.326	-0.47	0.4	1.01	0.11	0.99	0.02	0.38
64	H	20	0.845	0.82	0.42	1.12	0.73	1.09	0.43	0.25
65	M	20	0.840	0.82	0.42	0.79	-1.21	0.74	-0.97	0.58
66	M	16	0.260	0.16	0.4	1.21	1.46	1.22	1.17	0.18
67	M	18	0.573	0.48	0.4	0.86	-0.91	0.79	-0.95	0.53
68	M	23	1.342	1.39	0.46	0.92	-0.27	0.86	-0.26	0.41
69	M	20	0.861	0.82	0.42	1.09	0.57	1.04	0.24	0.28
70	M	19	0.722	0.65	0.41	0.97	-0.13	0.89	-0.39	0.42
71	M	23	1.313	1.39	0.46	0.81	-0.78	0.76	-0.55	0.51
72	M	18	0.552	0.48	0.4	1.02	0.2	0.97	-0.05	0.37
73	M	12	-0.318	-0.47	0.4	1.19	1.31	1.32	1.45	0.16
74	M	19	0.709	0.65	0.41	0.86	-0.86	0.8	-0.8	0.52
75	M	20	0.843	0.82	0.42	0.74	-1.56	0.68	-1.28	0.63
76	M	12	-0.310	-0.47	0.4	0.83	-1.16	0.92	-0.3	0.53
77	M	11	-0.468	-0.64	0.41	0.86	-0.92	0.76	-1.05	0.54
78	M	10	-0.614	-0.81	0.42	0.97	-0.14	0.93	-0.2	0.41
79	M	25	1.715	1.86	0.52	0.83	-0.49	0.56	-0.84	0.51
80	M	11	-0.454	-0.64	0.41	0.77	-1.58	0.67	-1.56	0.63
81	M	14	-0.002	-0.15	0.4	1.01	0.13	0.96	-0.14	0.39
82	M	10	-0.606	-0.81	0.42	0.84	-0.9	0.73	-1.09	0.55
83	M	16	0.247	0.16	0.4	1.21	1.46	1.31	1.55	0.16
84	M	20	0.839	0.82	0.42	1.15	0.88	1.09	0.42	0.23
85	M	13	-0.177	-0.31	0.4	0.91	-0.63	0.88	-0.58	0.48
86	L	13	-0.192	-0.31	0.4	0.87	-0.98	0.8	-1.06	0.53
87	M	18	0.562	0.48	0.4	1.03	0.24	1.29	1.29	0.31
88	M	14	-0.023	-0.15	0.4	0.81	-1.45	0.74	-1.46	0.59
89	M	18	0.571	0.48	0.4	0.76	-1.66	0.71	-1.4	0.62
90	L	9	-0.751	-0.99	0.43	0.75	-1.4	0.63	-1.38	0.63

Examinee ID	Proficiency level	ACCT score	Bayesian ability	Rasch ability	S.E. estimate	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd	PTM correlation
-91	L	8	-0.941	-1.18	0.44	0.8	-0.95	0.65	-1.13	0.58
-92	M	8	-0.934	-1.18	0.44	0.82	-0.8	0.67	-1.03	0.56
93	L	11	-0.457	-0.64	0.41	0.96	-0.24	0.84	-0.63	0.45
94	M	14	-0.001	-0.15	0.4	0.98	-0.14	0.95	-0.23	0.42
95	M	12	-0.304	-0.47	0.4	1.01	0.12	0.97	-0.06	0.38
96	M	11	-0.450	-0.64	0.41	0.92	-0.45	0.93	-0.22	0.45
97	L	8	-0.951	-1.18	0.44	1.09	0.46	1.16	0.58	0.26
98	L	9	-0.783	-0.99	0.43	1.06	0.36	0.91	-0.21	0.34
99	L	16	0.267	0.16	0.4	1.05	0.37	1.01	0.13	0.35
100	L	7	-1.124	-1.38	0.46	1.3	1.19	1.66	1.59	-0.01
101	L	10	-0.609	-0.81	0.42	1.07	0.46	1.05	0.26	0.31
102	L	9	-0.766	-0.99	0.43	1.07	0.42	1.13	0.52	0.28
103	L	13	-0.162	-0.31	0.4	0.71	-2.32	0.63	-2.1	0.69
104	L	9	-0.763	-0.99	0.43	1.03	0.19	1.22	0.81	0.3
105	L	8	-0.954	-1.18	0.44	0.78	-1.02	0.63	-1.19	0.6
106	M	9	-0.778	-0.99	0.43	0.94	-0.28	0.91	-0.23	0.43
107	M	11	-0.469	-0.64	0.41	0.81	-1.21	0.72	-1.27	0.58
108	L	6	-1.314	-1.61	0.49	0.99	0.07	1.42	1.01	0.26
109	M	10	-0.626	-0.81	0.42	0.78	-1.34	0.68	-1.29	0.61
110	L	12	-0.317	-0.47	0.4	0.97	-0.14	1.1	0.52	0.38
111	M	6	-1.293	-1.61	0.49	0.78	-0.78	0.63	-0.85	0.56
112	M	9	-0.770	-0.99	0.43	0.83	-0.88	0.71	-1	0.55
113	M	10	-0.615	-0.81	0.42	0.86	-0.81	0.77	-0.89	0.53
114	M	13	-0.156	-0.31	0.4	1.14	1.01	1.15	0.8	0.24
115	L	11	-0.474	-0.64	0.41	0.81	-1.23	0.71	-1.33	0.59
116	L	5	-1.518	-1.86	0.52	1.05	0.26	0.94	0.04	0.28
117	L	6	-1.283	-1.61	0.49	0.78	-0.77	0.63	-0.83	0.56
118	L	9	-0.767	-0.99	0.43	1.02	0.15	0.88	-0.33	0.38
119	M	9	-0.748	-0.99	0.43	1.03	0.21	1.05	0.28	0.33
120	L	6	-1.305	-1.61	0.49	0.78	-0.76	0.72	-0.57	0.54
121	L	8	-0.932	-1.18	0.44	1.09	0.47	1.04	0.23	0.28
122	M	4	-1.690	-2.15	0.57	1.06	0.29	2.27	1.77	0.1
123	L	8	-0.955	-1.18	0.44	0.96	-0.12	0.87	-0.29	0.41
124	L	5	-1.490	-1.86	0.52	1.15	0.55	1.42	0.9	0.12
125	L	7	-1.112	-1.38	0.46	1.14	0.61	1.04	0.22	0.23
126	L	11	-0.466	-0.64	0.41	0.9	-0.64	0.85	-0.61	0.49
127	L	4	-1.709	-2.15	0.57	0.8	-0.43	0.68	-0.4	0.48
128	L	9	-0.768	-0.99	0.43	1.11	0.64	1.23	0.82	0.23
129	L	6	-1.291	-1.61	0.49	1.01	0.15	0.91	-0.07	0.34
130	M	14	-0.021	-0.15	0.4	1.03	0.24	1.17	0.95	0.33
131	L	9	-0.783	-0.99	0.43	0.9	-0.49	0.76	-0.82	0.5
132	L	6	-1.302	-1.61	0.49	1.26	0.95	1.08	0.32	0.13
133	L	6	-1.285	-1.61	0.49	1.07	0.33	0.92	-0.05	0.29
134	L	8	-0.948	-1.18	0.44	0.85	-0.66	0.7	-0.92	0.53
135	M	14	-0.025	-0.15	0.4	0.92	-0.55	0.95	-0.22	0.46
136	M	6	-1.318	-1.61	0.49	0.82	-0.59	0.7	-0.64	0.51

Examinee ID	Proficiency level	ACCT score	Bayesian ability	Rasch ability	S.E. estimate	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd	PTM correlation
137	L	5	-1.472	-1.86	0.52	1.08	0.35	1.37	0.82	0.2
138	L	11	-0.467	-0.64	0.41	0.88	-0.72	0.77	-1.01	0.52
139	L	6	-1.275	-1.61	0.49	1.23	0.85	1.01	0.16	0.17
140	L	8	-0.930	-1.18	0.44	0.94	-0.2	1.1	0.42	0.38
141	L	7	-1.117	-1.38	0.46	1.2	0.85	1.4	1.07	0.13
142	L	9	-0.771	-0.99	0.43	1.14	0.75	1.37	1.25	0.18
143	L	10	-0.618	-0.81	0.42	1.03	0.25	1.06	0.31	0.33
144	L	12	-0.319	-0.47	0.4	0.77	-1.66	0.68	-1.63	0.63
145	L	8	-0.933	-1.18	0.44	1.07	0.4	1.19	0.65	0.27
146	L	4	-1.717	-2.15	0.57	1.25	0.75	1.9	1.4	-0.01
147	L	14	-0.012	-0.15	0.4	1.01	0.11	1	0.06	0.38
148	L	9	-0.773	-0.99	0.43	1.14	0.76	1.3	1.04	0.2
149	L	11	-0.442	-0.64	0.41	0.78	-1.45	0.68	-1.49	0.61
150	L	7	-1.090	-1.38	0.46	1.22	0.93	1.82	1.88	0.04
151	L	3	-1.960	-2.51	0.64	0.98	0.12	1.75	1.08	0.18
152	L	4	-1.677	-2.15	0.57	0.98	0.08	0.76	-0.25	0.35
153	L	6	-1.273	-1.61	0.49	0.85	-0.46	0.77	-0.44	0.48
154	L	12	-0.311	-0.47	0.4	1.04	0.34	1	0.08	0.35
155	M	6	-1.298	-1.61	0.49	1.12	0.51	1.54	1.22	0.15
156	L	4	-1.717	-2.15	0.57	0.87	-0.23	0.55	-0.7	0.48
157	L	5	-1.505	-1.86	0.52	0.84	-0.43	0.7	-0.49	0.48
158	L	8	-0.937	-1.18	0.44	1.09	0.47	1.18	0.61	0.26
159	L	12	-0.306	-0.47	0.4	1.13	0.88	1.11	0.58	0.26
160	L	6	-1.295	-1.61	0.49	1.08	0.38	1	0.14	0.27
161	L	12	-0.320	-0.47	0.4	0.93	-0.48	1.01	0.12	0.44
162	M	15	0.124	0.01	0.4	0.92	-0.53	0.93	-0.32	0.46
163	M	11	-0.452	-0.64	0.41	1.09	0.63	1.2	0.89	0.26
164	L	7	-1.122	-1.38	0.46	0.91	-0.3	1.35	0.97	0.36
165	L	8	-0.956	-1.18	0.44	0.77	-1.07	0.64	-1.14	0.6
166	L	8	-0.932	-1.18	0.44	1.29	1.31	1.71	1.89	0.01
167	L	7	-1.138	-1.38	0.46	1	0.08	1.44	1.16	0.25
168	L	11	-0.472	-0.64	0.41	0.88	-0.75	0.83	-0.7	0.51
169	L	10	-0.614	-0.81	0.42	1.16	0.92	1.14	0.61	0.22
170	L	4	-1.704	-2.15	0.57	0.97	0.05	0.72	-0.31	0.36
171	L	12	-0.306	-0.47	0.4	0.73	-1.97	0.65	-1.85	0.66
172	L	11	-0.475	-0.64	0.41	1.11	0.73	1.35	1.45	0.22
173	L	9	-0.784	-0.99	0.43	1.48	2.24	1.86	2.48	-0.17
174	L	10	-0.589	-0.81	0.42	1.29	1.59	1.43	1.57	0.06
175	L	8	-0.939	-1.18	0.44	1.18	0.85	1.14	0.5	0.19
176	L	7	-1.126	-1.38	0.46	0.9	-0.36	0.76	-0.57	0.47
177	L	11	-0.452	-0.64	0.41	1.15	0.94	1.29	1.23	0.2
178	L	6	-1.297	-1.61	0.49	0.78	-0.77	0.63	-0.83	0.56
179	L	7	-1.118	-1.38	0.46	1.37	1.45	2.1	2.36	-0.11
180	L	5	-1.485	-1.86	0.52	1.19	0.66	1.66	1.26	0.05
181	L	8	-0.937	-1.18	0.44	0.93	-0.25	1.01	0.13	0.4
182	L	7	-1.107	-1.38	0.46	1.15	0.66	1.26	0.76	0.18

Examinee ID	Proficiency level	ACCT score	Bayesian ability	Rasch ability	S.E. estimate	Infit Mnsq	Infit Zstd	Outfit Mnsq	Outfit Zstd	PTM correlation
183	L	9	-0.783	-0.99	0.43	0.93	-0.33	0.9	-0.25	0.44
184	L	6	-1.300	-1.61	0.49	1.51	1.68	2.24	2.28	-0.29
185	L	11	-0.454	-0.64	0.41	0.94	-0.36	0.96	-0.09	0.44
186	L	11	-0.464	-0.64	0.41	1.07	0.49	1.07	0.35	0.31
187	L	13	-0.166	-0.31	0.4	1	0.07	1.08	0.47	0.36
188	L	12	-0.306	-0.47	0.4	0.82	-1.3	0.83	-0.78	0.56
189	L	9	-0.789	-0.99	0.43	1.15	0.81	1.24	0.88	0.19
190	L	12	-0.310	-0.47	0.4	1.18	1.25	1.26	1.22	0.18
191	L	8	-0.922	-1.18	0.44	1.3	1.33	1.64	1.75	0
192	H	22	1.192	1.19	0.44	1.04	0.26	1.17	0.57	0.27
193	H	20	0.869	0.82	0.42	1.12	0.73	1.07	0.34	0.26

Appendix G. Multiple-choice distractor functioning analysis

Item number	Option code	Score value	Response data count	Response data %	Average ability	S.E. mean	Outfit Mnsq	PTM correlation
1	b	0	28	15	-1.1	0.15	0.6	-0.32
	c	0	3	2	-1	0.22	0.5	-0.09
	d	0	43	22	-0.53	0.17	1.6	-0.17
	a	0	13	7	0.07	0.32	2.1	0.04
	e	1	106	55	0.31	0.12	1.1	0.37
2	c	0	35	18	-1.03	0.13	0.4	-0.34
	d	0	46	24	-0.54	0.14	0.7	-0.19
	b	0	27	14	-0.49	0.18	0.7	-0.12
	e	0	48	25	0.54	0.17	2.3	0.3
	a	1	37	19	0.68	0.23	1.6	0.31
3	e	0	5	3	-1.68	0.16	0.3	-0.2
	a	0	11	6	-1.13	0.26	0.9	-0.2
	d	0	11	6	-1.13	0.18	0.7	-0.2
	c	0	6	3	-0.92	0.31	0.9	-0.11
	b	1	160	83	0.1	0.1	1	0.38
4	c	0	24	12	-1.05	0.14	0.5	-0.28
	b	0	29	15	-1.04	0.14	0.6	-0.31
	d	0	24	12	-1.02	0.18	0.7	-0.27
	e	0	16	8	-0.75	0.22	0.8	-0.15
	a	1	100	52	0.69	0.11	0.7	0.66
5	a	0	31	16	-0.82	0.15	0.8	-0.24
	b	0	7	4	-0.68	0.37	0.9	-0.09
	e	0	36	19	-0.55	0.18	1.2	-0.16
	d	0	23	12	-0.51	0.23	1.3	-0.11
	c	1	96	50	0.4	0.13	1.2	0.41
6	b	0	21	11	-1.26	0.11	0.4	-0.32
	a	0	18	9	-0.88	0.22	0.8	-0.19
	d	0	13	7	-0.88	0.16	0.7	-0.16
	e	0	25	13	-0.85	0.17	0.8	-0.22
	c	1	116	60	0.45	0.11	0.9	0.55
7	e	0	3	2	-1.32	0.14	0.3	-0.12
	c	0	4	2	-1.15	0.38	0.5	-0.12
	b	0	12	6	-1.12	0.14	0.4	-0.2
	a	0	82	42	-0.97	0.08	0.6	-0.58
	d	1	92	48	0.85	0.11	0.6	0.73
8	b	0	8	4	-1.1	0.13	0.4	-0.16
	c	0	47	24	-0.7	0.12	0.8	-0.26
	d	0	11	6	-0.62	0.32	1.2	-0.1
	e	0	24	12	-0.56	0.2	1.1	-0.13
	a	1	103	53	0.38	0.13	1.2	0.42
9	c	0	8	4	-1.17	0.3	0.6	-0.17
	d	0	17	9	-1.14	0.19	0.7	-0.25
	a	0	31	16	-0.96	0.17	1	-0.29
	e	0	20	10	-0.12	0.25	2.2	0

Item number	Option code	Score value	Response data count	Response data %	Average ability	S.E. mean	Outfit Mnsq	PTM correlation
	b	1	117	61	0.32	0.11	1	0.43
10	a	0	14	7	-1.09	0.16	0.5	-0.21
	b	0	13	7	-0.96	0.19	0.7	-0.18
	c	0	16	8	-0.85	0.23	0.9	-0.17
	d	0	37	19	-0.48	0.19	1.9	-0.14
	e	1	113	59	0.32	0.12	1.1	0.41
11	e	0	7	4	-1.11	0.39	0.6	-0.15
	c	0	26	13	-1.11	0.11	0.4	-0.31
	b	0	34	18	-1.07	0.08	0.5	-0.35
	d	0	30	16	-1.04	0.12	0.5	-0.31
	a	1	96	50	0.85	0.1	0.6	0.76
12	e	0	5	3	-0.9	0.21	0.4	-0.1
	d	0	53	27	-0.7	0.14	0.9	-0.28
	b	0	48	25	-0.66	0.13	0.9	-0.25
	c	0	16	8	-0.6	0.21	0.8	-0.11
	a	1	71	37	0.84	0.14	0.9	0.58
13	d	0	9	5	-1.69	0.17	0.4	-0.27
	e	0	5	3	-0.79	0.51	1.4	-0.09
	c	0	5	3	-0.48	0.58	2	-0.05
	b	0	18	9	-0.08	0.23	2.3	0.01
	a	1	156	81	0	0.1	1.1	0.19
14	d	0	11	6	-1.31	0.18	0.4	-0.23
	b	0	57	30	-0.92	0.11	0.8	-0.41
	e	0	14	7	-0.43	0.29	1.4	-0.07
	c	1	111	58	0.45	0.12	0.9	0.52
15	a	0	16	8	-0.98	0.21	0.5	-0.2
	d	0	23	12	-0.66	0.21	0.8	-0.16
	e	0	60	31	-0.63	0.13	1.1	-0.27
	b	0	33	17	-0.22	0.2	1.4	-0.04
	c	1	61	32	0.86	0.15	0.9	0.53
16	a	0	27	14	-0.81	0.21	0.9	-0.22
	c	0	46	24	-0.8	0.12	0.6	-0.3
	e	0	43	22	-0.52	0.14	0.8	-0.17
	b	0	20	10	0.19	0.28	2	0.08
	d	1	57	30	0.95	0.16	0.9	0.55
17	d	0	12	6	-1.32	0.1	0.2	-0.24
	a	0	47	24	-0.82	0.14	0.8	-0.31
	e	0	22	11	-0.67	0.18	0.7	-0.16
	c	0	56	29	-0.16	0.15	1.4	-0.02
	b	1	56	29	0.98	0.14	0.8	0.56
18	c	0	17	9	-1.07	0.12	0.3	-0.23
	a	0	41	21	-0.65	0.17	0.8	-0.22
	e	0	34	18	-0.61	0.13	0.6	-0.18
	b	0	45	23	-0.13	0.19	1.7	0
	d	1	56	29	0.86	0.16	1	0.49
19	e	0	22	11	-0.54	0.22	1.1	-0.12

Item number	Option code	Score value	Response data count	Response data %	Average ability	S.E. mean	Outfit Mnsq	PTM correlation
	c	0	13	7	-0.42	0.29	1.3	-0.06
	a	0	31	16	-0.04	0.21	1.9	0.03
	b	0	38	20	0.1	0.21	3.4	0.09
	d	1	89	46	-0.09	* .14	1.9	0.02
20	e	0	19	10	-0.98	0.17	0.4	-0.22
	d	0	52	27	-0.74	0.14	0.8	-0.3
	c	0	22	11	-0.31	0.18	0.8	-0.05
	a	0	49	25	-0.22	0.14	1.1	-0.05
	b	1	51	26	1.01	0.19	1.2	0.53
21	b	0	12	6	-0.69	0.28	0.7	-0.12
	a	0	37	19	-0.61	0.14	0.6	-0.19
	e	0	78	40	-0.48	0.12	0.9	-0.23
	d	0	17	9	-0.09	0.27	1.2	0.01
	c	1	49	25	0.96	0.19	1	0.49
22	c	0	11	6	-1.05	0.14	0.5	-0.18
	b	0	29	15	-0.92	0.17	0.8	-0.26
	d	0	35	18	-0.92	0.13	0.7	-0.3
	e	0	15	8	-0.15	0.31	1.9	-0.01
	a	1	103	53	0.48	0.12	1	0.51
23	d	0	43	22	-1.02	0.1	0.5	-0.38
	e	0	13	7	-0.7	0.23	0.8	-0.12
	a	0	20	10	0.08	0.25	2.1	0.05
	c	0	32	17	0.23	0.26	3	0.12
	b	1	85	44	0.25	0.14	1.3	0.26
24	d	0	25	13	-1.16	0.13	0.4	-0.32
	c	0	12	6	-0.89	0.22	0.6	-0.16
	b	0	31	16	-0.88	0.14	0.7	-0.26
	a	0	36	19	-0.78	0.11	0.6	-0.25
	e	1	89	46	0.81	0.12	0.8	0.68
25	d	0	8	4	-1.3	0.12	0.3	-0.19
	a	0	34	18	-0.97	0.12	0.6	-0.31
	e	0	29	15	-0.7	0.16	0.8	-0.19
	b	0	25	13	-0.62	0.22	1.2	-0.15
	c	1	97	50	0.58	0.13	0.9	0.55
26	e	0	7	4	-1.38	0.15	0.4	-0.19
	c	0	14	7	-1.3	0.13	0.4	-0.26
	d	0	31	16	-1.16	0.12	0.6	-0.36
	b	0	12	6	-1.15	0.16	0.5	-0.21
	a	1	129	67	0.42	0.1	0.8	0.61
27	d	0	23	12	-1.12	0.12	0.5	-0.29
	c	0	19	10	-0.96	0.14	0.6	-0.22
	a	0	26	13	-0.76	0.21	1	-0.2
	b	0	22	11	-0.57	0.23	1.3	-0.13
	e	1	103	53	0.52	0.12	0.9	0.54
28	c	0	15	8	-0.9	0.22	0.5	-0.18
	e	0	25	13	-0.72	0.22	0.8	-0.18

Item number	Option code	Score value	Response data count	Response data %	Average ability	S.E. mean	Outfit Mnsq	PTM correlation
	a	0	64	33	-0.23	0.14	1.2	-0.06
	b	0	44	23	-0.06	0.18	1.6	0.02
	d	1	45	23	0.58	0.21	1.7	0.31
29	c	0	27	14	-1	0.14	0.5	-0.28
	d	0	32	17	-0.93	0.16	0.7	-0.29
	b	0	19	10	-0.42	0.24	1.2	-0.08
	e	0	40	21	0.04	0.19	1.9	0.06
	a	1	75	39	0.54	0.15	1.1	0.41
30	c	0	36	19	-1.19	0.11	0.5	-0.4
	d	0	12	6	-0.84	0.21	0.8	-0.15
	b	0	20	10	-0.58	0.17	1	-0.12
	a	0	15	8	-0.38	0.28	1.5	-0.06
	e	1	110	57	0.43	0.12	1	0.5

* Average ability does not ascend with category score

Appendix H. Test evaluation form

The evaluation form of the collocation test item

Purpose of the test

The collocation test will be used as a placement test or a supplement test of existing placement tests in academic English courses at university or other institutions of higher education in the EFL context. It is aimed to measure a receptive dimension of a general academic verb-noun collocational competence, which is part of vocabulary depth and academic writing ability. The test scores will be interpreted based on a norm-referenced evaluation.

Components of a multiple-choice item

A single multiple-choice item consists of three aspects. The first aspect is a step, which is a problem in the form of an incomplete sentence. The second aspect is one best correct choice, which is an intended answer to the problem. The third aspect is four incorrect choices, which are distractors to the best correct choice.

☞ Look at the following example

A stem	*He _____ the English placement **test** and was put in the most advanced class.*
a distractor	*a. failed*
a correct answer	*b. passed*
a distractor	*c. won*
a distractor	*d. did*
a distractor	*e. took*

Instruction

In the evaluation form, please evaluate each test item by ticking "yes" or "no" for each of the following questions. Please use the evaluation form in conjunction with the test form.

Question Number 1: *Does the stem/problem present a single, clear sentence and a sufficient context for a verb-noun collocation?*

Question Number 2: *Is the correct answer clearly the best among plausible incorrect alternatives?*

Question Number 3: *Are the incorrect alternatives overall plausible enough to distract uninformed examinees away from the correct answer?*

☞ Look at the following example

No	Items	Question Number						Other Comments
		1		2		3		
		yes	no	yes	no	yes	no	
0	He _____ the English placement **test** and was put in the most advanced class. a. failed b. passed (a correct answer) c. won d. did e. took	✓		✓		✓		

Item	Question Number						Other Comments
	1		2		3		
	yes	no	yes	no	yes	no	
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
11							
12							
13							
14							
15							
16							
17							
18							
19							
20							
21							
22							
23							
24							
25							
26							
27							
28							
29							
30							
31							
32							
33							
34							
35							
36							
37							
38							
39							
40							
41							
42							
43							
44							
45							
46							
47							
48							
49							
50							

VITA

Apichat Khamboonruang was born in Mahasarakham province, Thailand. He obtained his current M.A. degree in English as an International Language with emphasis on language assessment and evaluation in 2013 from Chulalongkorn University, where he previously earned an M.Ed. degree in educational research and a B.Ed. degree in music education with specialisation in traditional Thai music in 2009 and 2006 respectively. His main research areas of interest include phraseology, corpus linguistics, second language assessment, meta-analysis, applied linguistics, and applied psychometric methods in the behavioural and social sciences.