University Library

A gateway to Melbourne's research publications

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:
Ngo, Phuong Viet Ha

Title:
Understanding Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests: A Case of English as a Foreign Language (EFL) Teachers in Vietnamese Higher Education Settings

Date:
2020

Persistent Link:
https://hdl.handle.net/11343/246526

Terms and Conditions:
Terms and Conditions: Copyright in works deposited in Minerva Access is retained by the copyright owner. The work may not be altered without permission from the copyright owner. Readers may only download, print and save electronic copies of whole works for their own personal non-commercial use. Any use that exceeds these limits requires permission from the copyright owner. Attribution is essential when quoting or paraphrasing from these works.

Understanding Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests: A Case of English as a Foreign Language (EFL) Teachers in Vietnamese Higher Education Settings

Phuong Viet Ha Ngo

ORCID identifier: 0000-0002-0442-0892

Submitted in total fulfilment of the requirements of the degree of Doctor of Philosophy

September 2020

Melbourne Graduate School of Education

The University of Melbourne

Abstract

This study investigated Vietnamese English as a Foreign Language (EFL) teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. The current study sought to construct and empirically calibrate a developmental framework that operationalises Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. The framework defines this competence as a continuum of competencies in terms of major tasks performed by Vietnamese EFL teacher item writers and delineates typical observable behaviours indicating their competencies at varying levels along a developmental progression. It also sought to examine to what extent teacher engagement in professional learning, their educational and professional background characteristics predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. For each of the three constructs explored, a scale was developed or adapted for the teachers to self-assess their level of competence in multiple-choice test item writing for English reading and listening skill tests and to self-report their engagement in self-regulated professional learning and access to formal induction and training. One hundred and fifty EFL teachers from various tertiary institutions in Vietnam responded to self-reported measures of their competence in multiple-choice test item writing for English reading and listening skill tests and their engagement in professional learning through an online survey. The results of both the classical test and item response theory analyses demonstrated that each of the scales had satisfactory measurement properties. The teachers' self-reported responses were further explored using multiple regression analysis procedures. Four factors were determined to be the most significant predictors of Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests (i.e., teachers' engagement in self-regulated professional learning, teachers' access to institutional induction

and training, their pre-service training in TESOL and their English language proficiency). Teachers' self-regulated and deliberate professional learning and institutional induction and training were confirmed to be the most powerful mechanisms for predicting Vietnamese EFL teachers' judgements of their competence in multiple-choice test item writing. Pre-service training in TESOL degree retained a moderate influence on the variance of the teacher competence while no evidence of a relationship between length of work-related experience and the competence was detected. It was also revealed that English language proficiency was a significantly positive factor associated with Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests. The findings had direct implications for theoretical models of multiple-choice test item writing for English reading and listening skill tests, policy-related issues for developing teachers' capability in multiple-choice test item writing, training and practice of EFL teacher item writers in Vietnamese higher education settings.

Acknowledgements

My doctoral research was supervised by Associate Professor Shelley Gillis and Doctor Cuc Nguyen at the Melbourne Graduate School of Education, The University of Melbourne. They provided feedback and advice which helped me to engage with and finish the project. As my principal supervisor, Shelley Gillis played a formative role in developing my thinking, and in helping me shape this dissertation.

I would also like to thank Professor Hamish Coates, Professor Margaret Wu, Associate Professor Graham Hepworth and Doctor Sue Finch for their thought-provoking suggestions and generous consultation throughout the project.

My warm thanks go to colleagues and PhD student fellows at the Assessment Research Centre and the Centre for Vocational Education and Policy for their engaging conversations, valued collegiality and fruitful comments.

The qualitative and empirical parts of the project depended on the generous input from approximately 200 English teachers and subject matter experts at foreign language universities in Vietnam and in Australia, who participated in the focus groups, workshops, expert reviews and the survey. I am thankful to them for their time and commitment.

I am grateful to my parents, my parents-in-law and my husband for their love, encouragement and support during my candidature. I am truly blessed to have my daughters, Chi and Quynh Anh and my son, Minh, who provide me with a daily dose of joy, a rainbow of laughter and a great burst of positive energy, without which the completion of the thesis would not have been possible.

Table of Contents

Abstract i

Acknowledgements iii

List of Tables x

List of Figures xiii

Declaration xvi

CHAPTER 1: INTRODUCTION 1

1.1 Background to the Study 1

1.2 Rationale for the Study 11

1.3 Aims of the Study and the Research Questions 17

1.4 Structure of the Study 20

CHAPTER 2: THE NATURE OF COMPETENCE 23

2.1 Concepts of Competence 23

2.1.1 The Generic Approach 24

2.1.2 The Behaviourist Approach 26

2.1.3 The Integrated Approach 27

2.1.3.1 The triangular model. 29

2.1.3.2 The cube model and the pyramid model. 30

2.1.3.3 The developmental model of competence. 33

2.2 Development of Competence 38

2.2.1 Conscious Competence Theory 39

2.2.2 Zone of Proximal Development 41

2.2.3 Staged Skill Acquisition Model 42

2.3 Methodological Approaches to Deriving Competencies and to Competency Framework Development 48

2.3.1 The Functional Analysis Approach 49

2.3.2 The Actuarial Approach 53

2.3.3 The Consensus Approach....54
2.3.4 The Developmental Multiple-Methods Approach....58
2.4 Chapter Summary....65
CHAPTER 3: TEST ITEM WRITING COMPETENCE....67
3.1 Item Writing Formats....67
3.2 Existing Models of Multiple-Choice Test Item Writing....73
3.2.1 Cognitive, Knowledge-Based Models of Item Writing....73
3.2.2. Performance Task-Based Model of Item Writing....81
3.3 Test Item Writing Competencies....88
3.3.1 Analysing Test Specifications....89
3.3.2 Generating Materials for Tests....90
3.3.3 Writing Test Items....95
3.3.3.1 Approaching item writing....96
3.3.3.2 Writing item stems and distractors....97
3.3.3.3 Providing answer keys....104
3.3.3.4 Deciding on item difficulty....104
3.3.4 Reviewing and Revising Test Items....106
3.3.4.1 Editorial review....108
3.3.4.2 Item content review....109
3.3.4.2.1 Qualitative item review....110
3.3.4.2.2 Statistical item analysis....111
3.4 Chapter Summary....114
CHAPTER 4: FACTORS AFFECTING TEACHER EXPERTISE DEVELOPMENT....116
4. 1 Professional Learning: A Mechanism for Teacher Competence Growth....116
4.1.1 Defining Professional Learning for EFL Teacher Item Writers....122
4.1.2 Measuring English Teacher Item Writers' Professional Learning....130

4.1.2.1 Measuring teacher engagement in self-regulated professional learning: The categorical model. 130
4.1.2.2 Measuring institutional induction and training for teacher item writers: The RITCME framework 134
4.2 Educational Background Characteristics Affecting Teacher Competence 141
in Test Item Writing 141
4.3 Professional Background Characteristics Affecting Teacher Competence in Multiple-Choice Test Item Writing 143
4.4 Research Questions Revisited 145
4.5 Chapter Summary 148
CHAPTER 5: RESEARCH METHODOLOGY 149
5.1 Scales and Questionnaire Development 149
5.1.1 The Development of the Progression Scale for Vietnamese EFL Teacher Competence in Multiple-Choice Test Item Writing (The IWCOMP Scale) 150
5.1.1.1 Scale Development Stages 150
5.1.1.1.1 Drafting. 152
5.1.1.1.2 Focus groups. 157
5.1.1.1.3 Workshops with Subject Matter Experts (SMEs). 161
5.1.1.1.4 Expert review. 164
5.1.1.2 The IWCOMP Scale 166
5.1.2 The Adaptation of the Professional Learning Scales (the PROLEARN and INTRAIN scales) 169
5.1.2.1 The PROLEARN Scale 171
5.1.2.2 The INTRAIN Scale 173
5.1.3 Professional and Educational Background Characteristics 178
5.2 Data Collection Procedures 180
5.2.1 The Participants and the Sampling Framework 180

5.2.2 Response Rate 183
5.2.3 Questionnaire Administration 183
5.3 Data Analysis 184
5.3.1 Item Response Modelling 184
5.3.1.1 Fit of the Model. 190
5.3.1.2 Reliability and its Contributions to Validity Arguments. 192
5.3.1.3 Summary Goodness-of-Fit. 194
5.3.2 Multiple Regression Analysis 194
5.3.2.1 Tests of assumptions for individual variables. 196
5.3.2.2 Estimating the regression model. 197
5.3.2.3 Tests of assumptions for the regression variate. 197
5.3.2.4 Assessing the overall model fit. 198
5.3.2.5 Interpreting the regression variate. 199
5.3.2.6 Validating the results. 199
5.4 Chapter Summary 202
CHAPTER 6: SCALES CALIBRATION 203
6.1 Calibration of Variables 203
6.1.1 Calibration of the IWCOMP Scale 203
6.1.2 Calibration of the PROLEARN Scale 226
6.1.3 Calibration of the INTRAIN Scale 234
6.2 Summary of Statistics 241
CHAPTER 7: MULTIPLE REGRESSION ANALYSES 243
7.1 Survey Sample Characteristics 243
7.2 Univariate Descriptive Statistics for Assumption Checks 246
7.2.1 Data Screening for Normality, Linearity and Homoscedasticity of Variables 246
7.2.2 Normality, Linearity and Homoscedasticity of Residuals 252
7.2.3 Independence of Errors 255

7.3 Bivariate Results 255
7.4 Multivariate Results 261
CHAPTER 8: CONCLUSIONS AND IMPLICATIONS 266
8.1 An Overview of the Study 266
8.1.1 Revisiting the Rationale and Aims of the Study 266
8.1.2 A Review of Methodology. 270
8.1.2.1 Scale Development Procedures 270
8.1.2.2 Data Collection Procedures 272
8.1.2.3 Data Analysis Procedures 273
8.2 Key Findings and Discussion 274
8.2.1 The Developmental Progression for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests 274
8.2.2 Factors Predicting Vietnamese Teachers' Judgements of Their Competence in Multiple- Choice Test Item Writing for English Reading and Listening Skill Tests. 282
8.2.2.1 The pre-eminence of professional learning. 282
8.2.2.2 The power of English language proficiency. 285
8.2.2.3 The pertinence of pre-service training. 285
8.2.2.4 The paradox of experience. 286
8.3 Contributions to Practice and Research 289
8.3.1 Contributions to the Field of Multiple-Choice Test Item Writing for English Reading and Writing Skill Tests Research 289
8.3.2 Contribution to the Practice and Training of Item Writers. 291
8.4 Study Limitations and Recommendations for Future Research 296
References 300
Appendix A. Plain Language Statements 317
Appendix B. Consent Forms 323
Appendix C. Ethics Approval 326
Appendix D. The Questionnaire 327

Appendix E. The Initial Draft Framework for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests 334

Appendix F. Amendments Made to the Framework following Focus Groups, Workshops with SMEs and Expert Review 341

Appendix F.1 Amendments to Performance Indicators Following Focus Groups 342

Appendix F.2 Amendments Made to the Quality Criteria Following Workshops.....343

Appendix F.3 Amendments Made to the Wording of the Quality Criteria Following Expert Review 352

Appendix F.4 Amendments Made to the Ordering of the Quality Criteria Following Expert Review 356

Appendix G. The Final Draft Framework for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests 361

List of Tables

Table 1.1 *English Language Proficiency Learning Outcomes for Vietnamese Students* 6

Table 1.2 *A Summary of the* VSTEP® Listening *and Reading Test Specifications* 8

Table 2.1 *Stages of Skill Acquisition from Novice to Expert* 43

Table 3.1 *The Cognitive Item Writing Processes*.....77

Table 3.2 *Cognitive Processes Employed by Item Writers* 79

Table 3.3 *Strategies for Adapting Texts/Scripts Applied by IELTS Reading and Listening Test Item Writers* 94

Table 3.4 *Item Writing Guidelines*.....99

Table 3.5 *Generic Item Shells*.....101

Table 4.1 *The List of Activities for Teacher Professional Development at the Workplace* ...133

Table 4.2 *Induction and Training Activities for English Reading and Listening Test Item Writers at Cambridge Assessment English* 137

Table 4.3 *The Observation Checklist for High-Quality Professional Development Training (HQPD Checklist)* 140

Table 5.1 *Methods Employed in the Development of the Draft Framework for Vietnamese EFL Teacher Competence in Multiple-Choice Test Item Writing* 152

Table 5.2 *Draft Competencies and Performance Indicators for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests* 154

Table 5.3 *Draft Quality Criteria Levels for Competency 1.1: 'Analysing Test Specifications'* 156

Table 5.4 *Focus Group Participants' Background Information* 159

Table 5.5 *Items within the IWCOMP Scale* 167

Table 5.6 *Categories of Professional Learning and Associated Items in the TPD @ Work Survey Instrument and in the PROLEARN Scale* 172

Table 5.7 *Items within the PROLEARN Scale* 173

Table 5.8 *The Core Features of Professional Training and Associated Items in the Original Instruments and in the INTRAIN Scale* 175

Table 5.9 *Items within the INTRAIN Scale* 178

Table 5.10 *Professional and Educational Background Information Sought about the Participants* 179

Table 5.11 *Summary of Survey Participants from Major Universities by Location* 181

Table 5.12 *The Nine Predictor Variables* 195

Table 6.1 *Details of Item 9's Analysis* 205

Table 6.2 *Details of Item 9's Analysis after Recode* 205

Table 6.3 *Calibration Estimates for the Test Item Writing Competence Scale* 207

Table 6.4 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 4* 217

Table 6.5 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 3* 218

Table 6.6 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 2* 219

Table 6.7 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 1* 220

Table 6.8 *The Developmental Progression for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests* 221

Table 6.9 *Calibration Estimates for the PROLEARN Scale* 228

Table 6.10 *Item Criteria, Delta Parameters and Derived Levels of Vietnamese EFL Teacher Engagement in Self-Regulated Professional Learning* 232

Table 6.11 *Interpretability of the Variable Map for the PROLEARN Scale* 233

Table 6.12 *Calibration Estimates for the INTRAIN Scale* 235

Table 6.13 *Item Criteria, Delta Parameters and Derived Description for Vietnamese EFL Teacher Access to Institutional Induction and Training*.....239

Table 6.14 *Summary Estimates of the Classical and Rasch Analyses for Each Scale.* 241

Table 7.1 *Professional Background Characteristics of the Sample.* 245

Table 7.2 *Professional Background Characteristics of the Sample.* 246

Table 7.3 *Mean, Standard Deviation, Skewness and Kurtosis Estimates.*.....247

Table 7.4 *Descriptive Statistics for the Nine Predictors and the Dependent Variable* 257

Table 7.5 *Correlation Matrix for the Variables involved in the Model.*.....258

Table 7.6 *Regression Coefficients with all 9 Independent Variables.* 262

Table 7.7 *The Overall Model Fit Results from Standard Multiple Regression Analysis*.....263

Table 7.8 *The Fit Indices of 10 Regression Models*.....264

List of Figures

Figure 2.1 An integrated model of competence 30

Figure 2.2 The cube model. 31

Figure 2.3 The pyramid model. 32

Figure 2.4 Representation of Conscious Competence Theory. 40

Figure 3.1 Model of item writing expertise based on cognitive processes 74

Figure 3.2 Cognitive process placement for item writers, indicating movement forward and backward in the problem space. 78

Figure 3.3 The conceptual framework for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. 86

Figure 3.4 The core competencies and corresponding performance indicators required for multiple-choice test item writers. 114

Figure 4.1 Model for prediction of Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. 147

Figure 5.1 Hierarchical organisation of the components of the framework for Vietnamese EFL teacher competence in multiple-choice test item writing. 151

Figure 5.2 The process of fitting the regression model. 201

Figure 6.1 Item Characteristics Curve of Item 6 (Adapts texts for texts-fit-items and items-fit-texts). 211

Figure 6.2 Item Characteristics Curve of Item 12 (Check item content fit). 212

Figure 6.3 Item Characteristic Curve of Item 5 (Make textual changes to the selected materials for tests). 212

Figure 6.4 Interpretability of the variable map for the IWCOMP scale. 215

Figure 6.5 The variable map for the PROLEARN scale 231

Figure 6.6 The variable map for the INTRAIN scale. 238

Figure 7.1 Scores distribution of the Multiple-Choice Test Item Writing Competence scale. 248

Figure 7.2 Scores distribution of the Professional Learning scale. 248

Figure 7.3 Scores distribution of the Induction and Training scale. 248

Figure 7.4 Normal Q-Q plot for the Multiple-Choice Test Item Writing Competence scale. 249

Figure 7.5 Normal Q-Q plot for the Professional Learning scale. 249

Figure 7.6 Normal Q-Q plot for the Induction and Training scale. 250

Figure 7.7 Linearity through bivariate scatterplot for IWCOMP and PROLEARN. 251

Figure 7.8 Linearity through bivariate scatterplot for IWCOMP and INTRAIN. 251

Figure 7.9 Linearity through bivariate scatterplot for INTRAIN and PROLEARN. 252

Figure 7.10 Normal P-P Plot of regression standardised residual. 253

Figure 7.11 Scatterplot of the regression standardised residual. 254

Figure 8.1 A single-dimension 'rocket' chart of Vietnamese EFL teacher competence in multiple-choice test item writing. 279

Figure 8.2 Level statements of five multiple-choice test item writing competencies demonstrated by Vietnamese EFL teachers. 281

List of Acronyms and Abbreviations

AERA	American Educational Research Association
ALTE	Association of Language Testers in Europe
APA	American Psychological Association
CET	College English Test
EFL	English as a Foreign Language
ESL	English as a Second Language
ETS	English Testing Service
IELTS	International English Language Testing System
GEPT	General English Proficiency Test
MOET	Ministry of Education and Training of Vietnam
NCME	National Council on Measurement in Education
NEA	National Education Association
SME	Subject Matter Expert
STEP EIKEN	The Eiken Test in Practical English Proficiency
TEFL	Teaching English as a Foreign Language
TEP	Test of English Proficiency
TESOL	Teaching English to Speakers of Other Languages
TOEFL iBT	Test of English as a Foreign Language – Internet -based Test
ULIS	University of Languages and International Studies
VNU	Vietnam National University
VSTEP®	Vietnamese Standardised Test of English Proficiency
ZPD	Zone of Proximal Development

Declaration

This is to certify that:

1. the thesis comprises my original research towards the PhD
2. due acknowledgement has been made in the text to all other material used; and
3. the thesis is less than 100,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Phuong Viet Ha Ngo

September 24th 2020

Preface

The author received the financial support for the research of this dissertation from the University of Melbourne (Melbourne International Fee Remission Scholarship and Melbourne International Research Scholarship 2015).

The author would like to acknowledge Ms. Rosemary Viete for her proofreading of the thesis. Her editorial intervention was undertaken in accordance with the Australian Standards for Editing Practice. Ms. Rosemary Viete proofread the thesis in its final draft, providing grammatical and semantic feedback, but preserving the author's meaning, style, organisation and voice.

CHAPTER 1: INTRODUCTION

1.1 Background to the Study

How the status of English in Vietnam should be framed has been debated in both the public media and the scholarly literature. One of the major newspapers in Vietnam released an article titled 'TTO – it is time to see English not as a foreign language' and the abstract below stating 'English should not be seen as a foreign language but a second language' (the author's translation) (Tuoi Tre News, 2005). This article attracted conflicting views from Vietnamese readers. Some advocated that Vietnam should institutionalise English as the second national language (i.e., English as a Second Language, ESL), alongside Vietnamese as other East Asian countries do, namely Singapore, Hong Kong (China), Malaysia, the Philippines and Brunei Darussalam. Others insisted on preserving Vietnamese as the only official national language and English as a Foreign Language (EFL). In one recent publication in the English language literature, Doan, Pham, Pham, and Tran (2018) argued that neither of the conceptualisations of English as an EFL or an ESL fully reflected the functions of English in this nation. Doan et al. (2018) suggested expanding the conceptualisation of English to encompass the national use of English for both international and intra-national communication and the national use of English as an instruction medium. While such debates are not of this study's interest, they indicate the enthusiasm for English in Vietnam and the radical change in its value and usefulness associated with the language (Blommaert, 2010). English was originally entitled the major foreign language in a government document in 1968 (Government of Vietnam, 1968), being specified in the document as English as a language of the Western countries. In 1994, the new Directive (Government of Vietnam, 1994) was released and characterised English as the language for international collaboration, and most recently, as a medium for the country's economic development (Government of Vietnam, 2008). At school levels, from 1982 to 2002, English was introduced nationwide as a compulsory subject at senior secondary school level (Grades10-12)

and as an elective subject at junior secondary level (Grades 6-9). After the successful piloting of teaching English at primary schools in large cities, the Vietnam Ministry of Education and Training (MOET) officially introduced English as an elective subject from Grade 3 to Grade 5 for two periods per week (each period was 40 minutes long) in primary schools that had adequate teaching resources and learning demand from pupils and parents (MOET, 1996, 2005). Since 2010, English has become mandatory with four periods per week in primary schools across the country. English language education has been introduced earlier (i.e., changing from Grade 6 to Grade 3) to students and is now offered to school students for 10 years of studying, as opposed to seven years as was the case prior to 1994. In the present, English is taught throughout the national curricula from Grade 3 to Grade 12 alongside other foreign languages such as Chinese, French, Russian, German and Japanese, though English has proved its dominant role as the preferred foreign language in Vietnam with 94% of the third graders choosing English (Hoang, 2011). Similarly, at the higher education level, among the foreign languages studied, namely English, Chinese, French, and more recently Japanese and Korean, English has been the most popular choice since the 1986 Open Door policy which initiated the explosive demand for English which played the role as the most crucial foreign language for Vietnamese people (Pham, 2014). The status of English was officially recognised in Vietnam's foreign language policy in which English is one of the six compulsory examination subjects for the national secondary education certificate (Hoang, 2011). The nationwide demand for English studying and English proficiency testing and assessment for academic, communicative, diplomatic, and career purposes in Vietnam has never shown sign of slowing down since 1986 (Truong & Wang, 2019).

The start of the twenty-first century witnessed the promulgation of numerous legal documents by the Government of Vietnam to promote the status of English and make English learning a priority (Pham, 2014). The latest innovation in foreign language education policy to

reinforce the status of English in particular by the Government of Vietnam was their enactment of Decision 1400/QD-TTg on the approval of the project entitled "Teaching and Learning Foreign Languages in the National Education System, Period 2008 – 2020" (*hereafter referred to as Project 2020*)" with a budget of US$443 million. Given the low per-capita GDP of US$ 1,164 in 2008 and US$ 2,111 in 2015, and many unresolved problems inherent in the nation's education system, this large-scale foreign language education reform policy has been acknowledged as the most ambitious government-initiated endeavour in the educational history of Vietnam (V. C. Le, 2015). It is common knowledge that the phrase *foreign language* automatically means "English" for many Vietnamese people and the dominant position of English as a foreign language in the society of Vietnam and in its educational curricula since Doi Moi (the Economic Reform) is unquestionable (Denham, 1992; Do, 1999, 2006; N. Nguyen, 2012; Shapiro, 1995). As claimed by the first Executive Director of Project 2020, among all foreign languages, the emphasis is placed on English as the government has recognised English as the priority foreign language, one which plays a critical role in the socio-economic development of Vietnam (N. H. Nguyen, 2013). The implementation of Project 2020 has demonstrated the government's strong commitment to improving the English proficiency of its national workforce. Project 2020 aims at making a substantial improvement in foreign language proficiency in the Vietnamese labour force through revolutionising the teaching, learning and assessment of foreign languages. The main objective of Project 2020 is to "renovate thoroughly the tasks of teaching and learning foreign languages within the national educational system in the period 2008-2020" (Government of Vietnam, 2008, p. 1). It envisaged that:

> …by the year 2020 most Vietnamese youth whoever graduate from vocational schools, colleges, and universities gain the capacity to use a foreign language independently. This will enable them to be more confident in communication, further their chance to study and work in an integrated and multi-cultural environment with a variety of languages.

> This goal also makes language as a comparative advantage for Vietnamese people, serving the cause of industrialization and modernization for the country.
>
> (Government of Vietnam, 2008, p. 1)

The amendments and supplements to the foreign language teaching and learning project for the 2017-2025 period expressed in Decision 2080/CP-TTg was approved by Deputy Prime Minister Vũ Đức Đam in December 2017. Accordingly, Project 2020 has now been extended to 2025 to achieve one of its main goals, which has been revised as follows: "By the year 2025, 100 per cent foreign language majored and 80 percent non-foreign language majored university graduates meet the required proficiency levels as specified in CEFR –VN" (Government of Vietnam, 2017, p. 1).

In order to achieve this strategic goal, Project 2020 has taken a series of radical actions that include:

1. Promulgating the six-level English Proficiency Standards for Vietnamese adapted from the Common European Framework of Reference (CEFR).
2. Setting benchmarking levels of English proficiency as Level 1 (A1) for primary students, Level 2 (A2) for lower secondary and Level 3 (B1) for upper secondary, college, and non-English major tertiary students; Level 5 (C1) for university graduates majoring in English language education.
3. Developing an English Teacher Competency Framework.
4. Reviewing, innovating and developing pre-service English language teacher education curricula, assessment and pedagogy in light of standard-based and outcome-based education with required levels of English proficiency upon graduation.
5. Conducting re-verification of in-service English teachers' language proficiency.
6. Re-educating or providing intensive training programs for in-service English teachers whose English proficiency has not met the required standards.

7. Rewriting English textbooks (from Grade 3 to Grade 12).
8. Institutionalising teaching English as a foreign language from Grade 3
9. Establishing National Language Testing and Assessment Centres in six regional foreign language centres (RFLCs) across Vietnam.
10. Developing the Vietnamese Standardised Test of English Proficiency (VSTEP®) from Level 3 to Level 5.
11. Harnessing Information Computer Technology for better and more efficient English language programs.

One of the innovations made by the Project 2020 on behalf of the Vietnamese government was their enactment of the Circular: 01/2014/TT-BGDĐT, which demands that English language proficiency be a mandatory higher education learning outcome for all college and university students across the country. The Circular approved the establishment of the six-level English language proficiency standards for Vietnamese users (CEFR-VN) and the prescribed levels of English proficiency required for students at each level of education in order to graduate. Table 1.1 summarises the English language proficiency levels students are required to achieve upon their graduation as specified in the Circular 1/2014.

Table 1.1 *English Language Proficiency Learning Outcomes for Vietnamese Students* (Vietnam Ministry of Education and Training (MOET), 2014)

STUDENTS	CEFR	CEFR-VN	International equivalents	
			Cambridge Assessment English Tests	IELTS
Primary school graduates	A1	Level 1		
Lower secondary and vocational school graduates	A2	Level 2	Key (KET)	
Upper secondary graduates; Non-English majored college and university graduates	B1	Level 3	Preliminary (PET)	4.0 - 5.0
English – majored college graduates	B2	Level 4	First Certificate of English (FCE)	5.5 - 6.5
English – majored university graduates	C1	Level 5	Advanced (CAE)	7.0 - 8.0
None	C2	Level 6		9.0

This English proficiency requirement and its standard setting entailed the creation and operation of the Vietnamese Standardised Tests of English Proficiency (VSTEP®). VSTEP® has been the latest undertaking of the Vietnamese government represented by Project 2020 to introduce a single 'made-in-Vietnam' English language proficiency test for use across all schools and universities in Vietnam. This is similar to other international initiatives such as the introduction of the Test of English Proficiency (TEPS) developed by Seoul National University for Korean students, the Eiken Test in Practical English Proficiency (STEP EIKEN) for Japanese students, College English Test (CET) for China, and the General English Proficiency Test (GEPT) for Taiwanese students to name a few. It was envisioned that VSTEP® would be recognised by all educational institutions in the country and it would enable a valid conversion

into IELTS and TOELF iBT scores that would be internationally recognised by most universities all over the world.

VSTEP® is the first standardised paper-based English proficiency test in Vietnam designed to assess four English language macro skills. The test specifications were developed by EFL teachers and language testing and assessment specialists from the University of Languages and International Studies, Vietnam National University, Ha Noi (ULIS-VNU). VSTEP® was officially launched nationwide by MOET in March 2015. VSTEP® aims to assess Vietnamese adult language learners' English proficiency for both educational and professional purposes. Specifically, VSTEP® is primarily used to validate university graduates' English language proficiency levels as presented in Table 1.1. VSTEP® certificates are among the prerequisites for candidates applying for post-graduate studies or job opportunities at government-owned corporations or organisations. Since March 2015, VSTEP® has become an official English proficiency test whose results, alongside those of IELTS, TOEFL iBT and other international English tests, are recognised by MOET Vietnam and as a result by all universities, government-owned organisations and corporations in Vietnam. VSTEP® has been administered extensively across Vietnam in response to the unprecedented demands to sit English tests from an annual cohort of 1.7 million college and university students in fulfilment of their university degrees and hundreds of thousands of job seekers, prospective employees and candidates for promotion and further study opportunities (Carr, Nguyen, Nguyen, Nguyen, & Nguyen, 2016).

Currently, the specifications for VSTEP® level 3-5 have been developed in accordance with three levels of English proficiency that VSTEP® aims to assess, namely level 3 (equivalent to CEFR –B1), level 4 (CEFR –B2) and level 5 (CEFR –C1). The VSTEP® specifications were approved in Decision 730/QD-BGDDT dated 11/3/2015 by MOET. The specifications give the detailed requirements for each of the skill tests regarding the theme, complexity and difficulty of input materials and the test item blueprints. Since this study proposed to investigate Vietnamese

teacher competence in multiple-choice test item writing for English reading and listening skill tests, the VSTEP® listening and reading test specifications were reviewed and summarised in Table 1.2.

Table 1.2 *A Summary of the* VSTEP® Listening *and Reading Test Specifications*

Skills	Part	Structure	Theme	Text/ Script length	No. items	Item difficulty levels	Timing
Listening	1	Two announcements, or instructions	Various	30-60 words per script	8 multiple-choice items	Lower level 3 to lower level 4 ($B1^-$ to $B2^-$)	10 -11 minutes
	2	Three conversations between two English native and/or ESL speakers	Occupations Social issues Educational topics in an Asian, ASEAN or Vietnamese context	2 minutes long	12 multiple-choice items	Lower level 3 to lower level 5 ($B1^-$ to $C1^-$)	10 -12 minutes
	3	Three short lectures	Social issues Educational topics in an Asian, ASEAN or Vietnamese context	2-3 minutes long	15 multiple-choice items	Upper level 3 to middle level 5 ($B1^+$ to mid-C1)	12 -15 minutes
Reading	1	One general text at Level 3	Everyday life	450 words	10 multiple-choice items	Level 3 to level 4 (B1 – B2)	10 minutes
	2	One academic text at Level 4	Social or natural science topics	450 words	10 multiple-choice items	Level 3 to level 5 (B1 – C1)	13 minutes
	3	One academic text at Level 4	A topic in social sciences, natural sciences, or another discipline	450 words	10 multiple-choice items	Level 3 to level 5 (B1 – C1)	13 minutes
	4	One academic text at Level 5	a specialised subject or literature	500 words	10 multiple-choice items	Level 3 to level 5 (B1 – C1)	14 minutes

The VSTEP® specifications also include the descriptors for the listening and reading skills at level 3-5 extracted from the six-level English Proficiency Standards for Vietnamese

people, the requirements regarding the format and content for each answer option and some example item stem shells. They also provide a short example script/text and five multiple-choice items extracted from a sample VSTEP® listening/reading test.

At the time of this study, no document or guideline that detailed the recruitment, induction, training, coordination, monitoring and evaluation processes used with VSTEP® item writers had been published. It was observed that VSTEP® examination boards in Vietnam followed the standard procedures commonly used for ordering and securing tests for end-of-course and university exit examinations. Notably, it appeared that the VSTEP® examination boards in Vietnam began to be aware of and showed interest in adapting the test development procedures conducted by the international examinations centre, Cambridge Assessment English. However, at the time of this investigation, the adaptation was still in its trial and error stage, so the recruitment, induction, training, coordination, monitoring and evaluation processes used with VSTEP® item writers in Vietnam were not yet complete. Regarding recruitment of item writers, over the last four years, groups of elite EFL teachers from the leading foreign language universities across Vietnam had been entrusted with the task of writing items for VSTEP® (Dunlea et al., 2016; 2018). At ULIS-VNU where VSTEP® was created, the selection of teachers to write VSTEP® tests was based on their academic qualification (usually at least a Master's degree in TESOL or Applied Linguistics), experience in teaching and designing English language classroom tests (at least three to five years), and reputation for high-achieving performance in the workplace as perceived by the teacher leaders. It was indicated that the teacher item writers' levels of expertise had not been directly assessed based on empirical research evidence, but rather based on social judgement by supervisors and colleagues with the primary criteria being completed education, accumulated knowledge, length of experience, and social reputation in the profession (Chi, 2006; Chi, Glaser, & Farr, 1998 cited in Ericsson, 2009). There was no formal, face-to-face induction for the item writers who were invited to write items

for VSTEP® tests except that a copy of the VSTEP® test specifications was provided to the item writers. Training of VSTEP® item writers appeared to be prioritised by the VSTEP® examination boards in Vietnam. ULIS-VNU was among the first institutions in Vietnam to organise training workshops for the EFL teachers who engaged in VSTEP® test item writing. The training was conducted once a year by experts in language test development and offered to the VSTEP® item writers from within ULIS-VNU as well as from other universities to equip them with theories and best practices in test item development. The coordination process for VSTEP® test writing involved limited communication between the VSTEP® item writers and the examination board. The item writers worked independently to write items for the tests, and written feedback on their submitted items may be sent to them for revisions. How VSTEP® test item performance was monitored and evaluated could be an area that VSTEP® examination boards in Vietnam considered to invest their time and effort into in the future. It was informed that analyses of test items of used VSTEP® reading and listening tests were conducted to generate data on the test item performance. The data was not provided to the item writers, and it was unknown if the data was used to inform the examination board's future recruitment decision.

With the recent implementation of VSTEP®, whose results are intended to be comparable with those of the internationally recognised IELTS and TOEFL iBT, the common practice of recycling test items from published test books and online resources has become obsolete. The practice of test item writing has been introduced to EFL teachers in Vietnam for the very first time. The development and operation of VSTEP® has not only marked the first attempt by the government to create the first local standardised English proficiency test but also to build teachers' capability and expertise in EFL test item writing (Dunlea et al., 2016; Dunlea et al., 2018; T. N. Q. Nguyen & Do, 2015). There has been a clear and pressing need for EFL teachers to acquire and develop competence in writing English test items in accordance with the standards to accurately measure learners' English proficiency and to inform curricula and

instruction. This extensive national demand for recognising and developing teacher competence in multiple-choice test item writing for large-scale English reading and listening skill tests in the local context of Vietnamese higher education served as the inspiration for this current study, which investigates what constitutes Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests, and what teacher-related factors predict the development of this competence. Adopting the integrated, developmental approach to defining competence advocated by Griffin (2014, 2018) and Masters (2003), it was argued that Vietnamese EFL teacher competence in multiple-choice test item writing could be manifested by sets of underlying attributes and visibly demonstrated through the completion of a series of related tasks. These tasks are expressed in observable and measurable behaviours/actions evidencing precisely what these item writers do to demonstrate their competence. The item writers' self-assessment of their performances on these tasks could be used to make inferences about their level of competence on a developmental progression. These inferences could subsequently inform educational and professional learning activities and experiences to facilitate the teachers' development of expertise in multiple-choice test item writing.

1.2 Rationale for the Study

The expertise of item writers has been an area of active inquiry given their critical role in terms of human resources in the process of test development as pointed out by language test experts such as Davidson and Lynch (2002). Shin (2012) and Kim et al. (2010) emphasise the general lack of a thorough documentation of item writers' expertise. It is claimed that the expertise of item writers has been mystified and research on one of the most "under-identified job descriptions" for item writers has been called for (Shin, 2013, p. 590). Research studies attempting to 'demystify' teachers' item writing expertise have been scarce. Among a few rare pieces of research focusing on this area are those conducted by Fulkerson, Mittelholtz, and

Nichols (2009), Fulkerson, Nichols, and Mittelholtz (2010), Fulkerson, Nichols, and Snow (2011) and Salisbury (2005, 2016).

A series of case studies led by Fulkerson (2009, 2010, 2011) explored the cognitive processes used by a small group of experienced and inexperienced item writers. These item writer participants were required to complete a test item writing assignment given by the researchers and were requested to verbalise cognitive information they generated during the item writing process, known as think-aloud protocols. Their verbal statements were audio-recorded, analysed and coded into categories of cognitive processes and knowledge structures that item writers made use of to write test items. The findings from these studies were generalised into a cognitive model of item writing expertise. In their first project in 2009, Fulkerson, Mittelholtz, and Nichols conducted a study of the cognitive process of three experienced item writers, and it was revealed that expert item writers typically engaged in a three-phrase problem solving process. In 2010, Fulkerson, Nichols, and Mittelholtz extended this cognitive model of item writing expertise to develop a better understanding of the processes employed by not only expert item writers but also novice ones. The results from this second study continued to confirm that regardless of their experience, item writers went through a three- phase problem solving process as they wrote items for tests. In their most recent expert study of item writers, Fulkerson et al. (2011) expanded the cognitive model of item writing expertise developed earlier by Fulkerson and colleagues (2009, 2010) by adding core knowledge structures for item writing into their previously proposed model of the cognitive processes of item writers' (Fulkerson et al., 2009, 2010). The analyses of transcripts of the participants' think-aloud sessions conducted in the study suggested that, regardless of their experience, all the item writers made use of nine cognitive processes and five knowledge structures in their item writing process. The cognitive, knowledge-based model of item writing developed from the latest study led by Fulkerson (2011) contributed an understanding of the three-phased item writing process all item writers moved

through with some noticeable differences between expert and novice item writers. It also offered a glimpse into the knowledge structures and cognitive processes applied by item writers. However, viewing test item writing from a cognitive perspective, rather than a competency-based perspective, the model restricted our understanding of item writing expertise to a list of cognitive processes and knowledge structures. Such a list seems of limited usefulness informing and supporting the practice and expertise development of item writers. Through this model of item writing, little has been revealed about what item writers do, create, write and say at each stage of their item writing process. Thus, the cognitive, knowledge-based model of item writing expertise provided a very limited understanding of the core competencies item writers should exhibit.

The other model of item writing found in the literature is the task-based model of item writing, which emerged from Salisbury's (2005) study exploring the sequential steps that expert and non-expert English for speakers of other languages (ESOL) listening test item writers undertake to write the test items using Fulkerson and colleagues' methodological approach.

The study findings demonstrated that there were three broad phases that all item writers went though, further confirming the three-phased problem-solving process both inexperienced and experienced item writers engaged in as revealed by Fulkerson et al. (2009, 2010, 2011). While Fulkerson et al. (2009, 2010, 2011) categorised the process into cognitive problem-solving stages, Salisbury (2005) conceptualised these stages in terms of actions carried out by item writers. In the first phase, the *Exploratory* phase, item writers search and identify potentially exploitable input materials for tests from various written or oral sources such as magazines, newspaper, specialist journal articles and radio programs. In the second phase, the *Concerted* phase, they work intensively through five 'episodes' (Salisbury, 2005, p.120) to produce test items for initial submission. These episodes are *Review test specifications, Read text, Devise context, Devise script, and Devise items*. The final phase, the *Refining* phase,

happens after either self-, peer- or editor-review. It is in this final phase that the item writers responded to a critical review by the writers themselves and/ or by others and improve the test items.

Unlike Fulkerson and colleagues who conceptualised test item writing expertise as a set of cognitive processes and knowledge structures, Salisbury (2005) adopted the behaviourist approach to conceptualising test item writing, characterising it as a set of five key performance tasks executed by test item writers regardless of their experience. These major tasks are analysing test specifications, generating input materials for test, writing test items, reviewing and revising test items. Salisbury's (2005) performance task-based model is the only one found in the current literature that views test item writing from a competency-based perspective and contributes to demystifying competencies required of item writers and expressing these competencies in five observable behaviours. The expert studies of item writers led by Fulkerson (2009, 2010, 2011) and Salisbury (2005) have contributed a solid understanding of item writing expertise which can be modelled as a three-phase process delineating five major competencies in terms of core, observable tasks performed by item writers regardless of their levels of expertise (Salisbury, 2005). However, it is not yet known how these item writing competencies are developed.

A large part of the literature on test item writing competencies seems to be scattered across several case studies and an indefinite number of manuals, guidelines and handbooks for test development. A few separate studies on how expert and non-expert item writers demonstrate the evidence of their competence in performing the five core test item writing tasks as identified by Salisbury (2005). For example, Kim et al. (2010) and Lynch and Davidson (1994) studied how item writers utilise test specifications; Green and Hawkey (2011) and Sydorenko (2011) explored how item writers search, select and adapt input materials for tests and make judgements of test item difficulty. However, these studies focused on distinguishing between novice and

expert practice, limiting the competence to dichotomous levels and was thus not able to capture the progression of these competencies as item writers develop from novice to expert. Similar studies on the other competencies are extremely scarce. Instead, all the five test item writing competencies are discussed mainly in the forms of checklists, tips and 'do's and don'ts' in comprehensive handbooks for test development and test item construction such as those by Haladyna (2004), Downing and Haladyna (2006), Haladyna and Rodriguez (2013), and Lane, Raymond, and Haladyna (2015). For instance, as reported by Rodriguez (2016), multiple-choice test item writing guidelines have been present in approximately 100 textbooks and book chapters since 1951 and reviewed multiple times by such authors as Haladyna and Downing (1989), Haladyna, Downing, and Rodriguez (2002), and Moreno, Martínez, and Muñiz (2015). However, these products for test development are expressed more in terms of checklists, guidelines, principles and procedures to be followed as opposed to developmental tools that promote the acquisition and development of test writing expertise and inform the professional development and training of item writers. It is argued that while current research and guidelines contribute a solid understanding of how these item writing competencies are demonstrated, little is known about how these competencies are acquired, nurtured, and developed. There is a lack of empirical research offering a holistic insight into test item writing competence, and an absence of empirical research into the developmental pathways of these competencies. There is a need to operationalise these competencies on a continuum in light of skill acquisition and expertise development theories so that item writers' acquisition of and development in these competencies could be better understood. Such a developmental continuum of competence could be used for a range of educative, self-monitoring and developmental purposes.

In summary, it remains unknown how item writing competencies are developed, and how expertise can be improved. Such limited understandings of item writing competencies are inadequate to inform the current self-directed professional development, training, and practice of

item writers. There is no empirical evidence on which to base how teachers develop expertise in test item writing, so their professional development needs and goals can be specified and their competence growth could be scaffolded in the light of competence development discussed in works such as that by Gillis and Macpherson (1995), and Griffin, Gillis, and Calvitto (2007). Thus, there is a great need for a more theoretically and methodologically sound model of item writing competency development which can contribute to the current scholarly literature by providing a more thorough understanding of how item writing competencies can be identified and developed.

Furthermore, to date no research has adequately and empirically measured and examined factors influencing teacher competence in multiple-choice test item writing for English reading and listening skill tests. While factors such as academic qualifications and English teaching experience have been identified as core recruitment criteria for item writers, and professional development has seemed to play a crucial role in testing organisations' agenda, previous research has not investigated these in detail as factors contributing to test item writing competence. Moreover, no research has accounted for the correlation between these factors and the development of teachers' in multiple-choice test item writing for English reading and listening skill tests. This inadequacy and scarcity have imposed constraints on understanding the extent to which the factors have individually and in concert contributed to English test item writing competence. Thus, it is unknown what educational and professional interventions can be targeted and implemented to support item writers' development of expertise. However, the substantive research in expertise development and teachers' professional competence growth in the workplace has suggested a number of potential contributory factors. Of interest to this research are those factors that are acquirable or malleable. Compelling evidence from research has shown that engagement in self-regulated professional learning that stimulate reflection, self-evaluation, and collaboration is conducive to professional competence growth (e.g., Boshuizen,

Bromme, & Gruber, 2004; Dunn & Shriner,1999; Ericsson, Krampe, & Tesch-Römer, 1993, 1996; Sonnentag, 1998; van de Wiel, Szegedi, & Weggeman, 2004). Besides, considerable research has suggested that pre-service formal training and in-service professional development focusing on language assessment improved teachers' levels of language testing and assessment literacy respectively (see Chen, 2005; Griffin, Care, Roberton, Crigan, Awwal, & Pavlovic, 2013; Koh, 2011; Levy-Vered & Alhija, 2015; Mahapatra, 2016; McGee & Colby, 2014; Mertler & Campbell, 2005; Montee, Back, Donovan, & Thompson, 2013; Riestenberg, Di Silvio, Donovan, & Malone, 2010; Sato, Wei, & Darling-Hammond, 2008; Walters, 2010; Yastibas, 2018; Zhang & Burry-Stock, 2003). In summary, both informal and formal professional learning activities at individual and institutional levels have been consistently found in the scholarly literature to be powerful mechanisms for teacher competence growth and expertise development (Desimone, 2009). Additionally, other educational and professional background characteristics have been recognised as factors possibly contributing to teacher competence growth in few case studies of test item writers and numerous studies on teachers' expertise. These background factors are possibly teachers' length of work-related experience (Charness, Krampe, & Maye, 1996; Ericsson et al., 1993; Sonnentag, 1995), academic qualifications, degree majors (Alderson, Clapham, & Wall, 1995), specific training in testing and assessment and their English language proficiency levels (Green & Hawkey, 2007; Kim et al., 2010). These findings have offered directions for further research into factors that predict teacher development of expertise in multiple-choice test item writing as they have the potential to enhance our understanding of what interventions could be targeted to improve teachers' multiple-choice item writing expertise.

1.3 Aims of the Study and the Research Questions

Test item writers or item writers are the titles used by examination boards such as English Testing Service (ETS), Association of Language Testers in Europe (ALTE) and

Cambridge Assessment English to target those producing tests and/or test items from given test specifications; therefore, this term is used throughout the study. According to ALTE (2005), item writers are those who are involved in the whole process of test item development using multiple cognitive processes and knowledge domains to carry out a range of tasks from interpreting the test specifications, searching for materials to devising test items. The term distinguishes *item writers* from test developers, test designers and test writers. *Test developers* are those involved in the creation of test specifications to build validity arguments for the tests (ALTE, 2005). *Test writers or test designers* refer to the people who are engaged in the whole test development cycle from specification design to final test construction. Thus, *item writers* are those who execute the test specifications and translate it into physical tests meeting all the requirements of the test item blueprint in the test specifications. Within this current study, item writers are EFL teachers in Vietnamese universities who are commissioned to write items for large-scale standardised English tests.

The study aspired to respond to the need for empirical research into test item writing competence to provide more thorough understanding of how test item writing competencies can be demonstrated and developed, and of factors that can predict the development of these competencies in the researcher's local context. Firstly, the study aimed to develop and validate a developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. Drawing together past research and gathering data on the practice of EFL teacher item writers in Vietnamese universities, the first purpose of the current study was to construct and empirically calibrate a framework that represents a developmental conceptualisation of Vietnamese EFL teacher competence in multiple-choice test item writing. Within this framework, test item writing competence was defined as a continuum of competencies in terms of major tasks performed by EFL item writers delineating typical observable behaviours at varying levels along a developmental progression.

This study aimed to analyse, specify and validate competencies that are unique to EFL multiple-choice test item writing in the context of Vietnamese higher education. It aimed to shed light on the practice of item writers, providing insights into their knowledge and cognitive processing skills in the form of observable behaviours demanded of them to fulfil their major item writing tasks. The goal was, thus, to develop a theoretically based, contextually and empirically validated, and operationalisable model that captures the core constituents of the Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests and depicts developmental trajectories of test item writing competencies. The study worked in a number of ways to unravel the competencies required of EFL test item writers. Possible competencies were exposed through extensive reviews of the scholarly literature, focus groups, and workshops with Vietnamese English teacher item writers as well as expert consultation. Following these exploratory investigations, a psychometrically calibrated instrument was used to gather data to refine and validate the measurement of these competencies. In doing so, it laid the foundation for subsequent, more thorough interpretation of the competence, which constitutes a theoretical and practical contribution to the understanding of teacher competence in multiple-choice test item writing for English reading and listening skill tests.

Secondly, the research sought to explore the factors that contribute to the development of Vietnamese teachers' perceived competence in multiple-choice test item writing for English reading and listening skill tests. Three sets of teacher-related factors, namely teachers' engagement in professional learning, and their professional and educational background characteristics, were hypothesised to make an explanatory contribution to the understanding of the teachers' competence.

The primary and subsidiary research questions were:

1. What is the typical developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests?

2. What teacher-related factors predict Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.1 To what extent does professional learning *(i.e., self-regulated professional learning and access to institutional induction and training)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.2 To what extent do teachers' professional background characteristics *(i.e., length of experience in teaching English language skills, teaching English exam preparation courses, constructing English tests from published resources, and writing multiple-choice test items)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.3 To what extent do teachers' educational background characteristics **(***i.e., their highest degree's major, uptake of coursework in language testing and assessment and their English language proficiency)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

1.4 Structure of the Study

There are eight chapters in the thesis. Chapter One establishes the context for the study by discussing the pressing needs for EFL teachers' acquisition and development of competence in multiple-choice test item writing for English reading and listening skill tests in Vietnamese higher education settings due to the Government's mandate of the latest language education renovation. It presents the rationale for the study by addressing the local demand to which the study aspires to respond and the gaps in the literature it wishes to fill. The chapter also outlines the aims, the research questions and structure of the study.

Chapter Two explores the nature of competence from both conceptual and methodological stances. It discusses approaches to conceptualising competence, presents

developmental expertise models for describing levels of competence, and investigates approaches to deriving competency frameworks. The main purpose of this chapter is to gain an understanding of the conceptualisation, derivation and development of competence that informs the critical analyses of the existing models of test item writing competence in the next chapter. It also aims to provide the theoretical and methodological approach to defining Vietnamese EFL teacher competence in multiple-choice test item writing for this study.

Chapter Three contains reviews of the body of research and scholarship pertinent to the study. It reviews and critiques the existing models of item writing on theoretical, practical and methodological grounds in light of their applicability to the developmental approach to competence presented in the previous chapter. It also presents a synthesis of research in best practices in multiple-choice test item writing. This synthesis is organised around the five major competencies required of EFL test item writers as identified in earlier research. The chapter establishes the conceptual framework for operationalising Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests within this current research study, based on the combined insights of Vygotsky (1978), Glaser (1963, 1981) and Rasch (1960, 1980), as brought together by Griffin and colleagues (2004, 2007, 2009, 2014, 2018).

Chapter Four draws upon previous research in expertise development and teachers' professional competence growth in the workplace to give an account of the factors that potentially influence teacher competence in multiple-choice test item writing. Identifying that professional learning is the main mechanism that drives teachers' professional competence growth, models and measures of professional learning in the field of teacher education and development are reviewed. Existing measures of professional learning are contextualised and adapted for use in defining and measuring EFL test item writers' engagement in professional learning. The chapter also explores educational and professional background characteristics that

have been identified as potentially influencing teacher competence in multiple-choice test item writing within the context of large-scale English testing and assessment. At the end of the chapter, the research questions underpinning the investigation are described.

Chapter Five discusses details of the research participants including sample composition and method of recruitment, and the development of the data collection instruments. It also elaborates on the methods and procedures for collecting and analysing the data for the investigation into Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests in Vietnamese higher education settings.

Chapter Six provides an overview of the calibration of the three construct-measuring scales used in this study to measure English teachers' perceived competence in multiple-choice test item writing and the factors that were hypothesised to affect the competence. The CONQUEST software version 4.0 (Adams, Wu, & Wilson, 2015) was employed to calibrate each of the scales. The analyses aimed to describe the measurement properties of each scale and levels of each of the constructs, to examine the performance of individual items on each scale, and to generate estimates of English teacher competence in item writing, their engagement in self-regulated professional learning, and the accessibility of institutional induction and training for them. To start with, the calibration of each of the three latent variables is reported, followed by the discussion of the properties of each of the calibrated scales and how these inform scale item revision.

Chapter Seven reports the results obtained from the multiple regression analyses of the data obtained in the study.

The thesis concludes in Chapter Eight with the discussion of the key findings and the contributions of the study to the field of test item writing research, including the practice and training of test item writers. Based on the findings and limitations of this study, the chapter also identifies implications for future inquiry.

CHAPTER 2: THE NATURE OF COMPETENCE

This chapter explores the nature of competence from both conceptual and methodological stances. It discusses approaches to conceptualising competence, presents developmental models for describing levels of competence, and investigates methodological approaches to deriving competency frameworks. The main purpose of the chapter is to gain a conceptual and methodological understanding of the conceptualisation, derivation and development of competence that informs the critical analyses of the existing models of test item writing competence in the next chapter. It also aims to provide a theoretical and methodological approach to modelling teacher competence in multiple-choice test item writing under this study.

2.1 Concepts of Competence

The notion of competence has been widely debated in the scholarly literature, including discussion in a number of comprehensive reviews and concept analyses (Axley, 2008; Cowan, Norman, & Coopamah, 2007; Fernandez et al., 2012; Frank et al., 2010). Competence is regarded as an ambiguous notion (Gardner, Hase, Gardner, Dunn, & Carryer, 2007), or as a 'wicked word' – an evil word that is very difficult to define - as maintained by Guerrero and De los Rios (2012, p. 1291). The concept of competence, originally introduced by White (1959) and spreading worldwide in the 1980s, has evolved enormously (Guerrero & De los Rios, 2012). As claimed by Cowan et al. (2007), the literature about defining competence is replete with "controversy, ambiguity, confusion and contraction" (p. 23). Within the literature, terms such as *competency, capability, performance, qualification, capacity, ability, attributes, adequacy, expertise, proficiency and potential* are offered as either alternatives or explanations as to how competence might be understood (Eraut, 1998; Griffin et al., 2007; Griffin, Nguyen, & Gillis, 2004; Hager & Gillis, 1995; Hager & Gonczi, 1993; Heywood, Gonczi, & Hager, 1992). The interchangeable use of all these terms may have contributed to variety of interpretations of the concept of competence. In this study, the term 'competence model' was used to refer to the

approach to which competence is conceptualised, and to describe different conceptual models of competence presented in the literature review. The term ‘competency framework’ was used to describe the framework for Vietnamese EFL teacher competence in multiple choice test item writing for English reading and listening skill tests developed and empirically validated within the present study. Given the diverse conceptualisations of competence, it is not surprising that there is no universally agreed approach to defining competence (Watson, Stimpson, Topping, & Porock, 2012). Hager and Gillis (1995) identified three broad approaches to defining competence: the generic approach, behaviourist approach, and integrated approach. These approaches to defining competence potentially have pragmatic benefits, including the ability to develop better methods of instructions, training and coaching experts and the ability to inform research on the learning and acquisition of expertise.

2.1.1 The Generic Approach

The *generic* approach conceptualises competence as possession of skills, knowledge and attitudes that allow a person to perform the job functions successfully and effectively in the workplace. Some examples of attributes thought to be salient in the generic approach to competence are problem solving, critical thinking, teamwork, decision making and communication skills (Eraut, 1998; Heywood et al., 1992). This approach viewed competencies as prescribed attributes, the presence of which allows inference of the underlying competence across the contexts in which it may be applied (Hager, 1995, p. 143). This attribute-based model of competence assumes that specifiable general attributes are transferable into other contexts, facilitating competent performance of an array of specific tasks in professional workplaces (Hager & Gillis, 1995). Thus, someone who, for example, has a generic attribute, such as problem-solving skills, will be capable of applying them in a wide range of tasks. Adopting this approach, it was revealed by Klemp (1988) that highly successful performers in a range of jobs typically possessed general attributes such as systematic thinking, diagnostic thinking,

conceptualisation, influence skills, sensitivity to others, confidence, initiative and persistence. Similarly, Hearn, Close, Smith, and Soutley (1996) used this approach to identify a set of generic attributes characteristic of a range of professions such as marketing, teaching and human resource management and endorsed its usefulness. However, Reeves, Fox, and Hodges (2009) argue that this approach may not be helpful for descriptions of attributes of competence that are unique to those professions (Hager, 2017; Reeves et al., 2009). This is supported by the evidence from the novice/expert literature suggesting that what characterises high levels of competence is not any generic attribute, but the knowledge and skills that are significantly specific to the professions (e.g.,Tennant, 1991; Yates & Chandler, 1991). The idea that any generic attribute acquired can be translated into practice in most professions was argued by Hager (2004, p.429) to be "naïve". As the generic approach creates difficulties in describing unique and distinctive bodies of knowledge and practice that define competence for particular professions, it is considered by Hager (2004) to be of limited value in terms of the practicalities in informing professional development, education and training. Added to that, others have argued that attributes underpinning competence such as capabilities, abilities and skills are so diverse that it is barely possible to specify them precisely (Hager, 2004; Lum, 2004; Magee, 1998). Another criticism regarding the generic approach is that even if the capabilities, abilities and skills constituting competence are identifiable, it is difficult to assess such attributes as critical thinking, problem solving, and empathy through inference (Cowan et al., 2007; Hager & Beckett, 1995) and out of the work context (Hager, 1995; 2017). The generic approach to competence, thus, is not favoured by those seeking concrete evidence of competence in performance (Eraut, 1998). This approach was not aligned with the aim of this study which was to describe Vietnamese EFL teacher competence in multiple-choice test item writing in concrete and observable behaviours to inform self-development, training and practice of Vietnamese EFL test item writers.

2.1.2 The Behaviourist Approach

The second approach to conceptualising competence is the task-based or behaviourist approach. This approach defines competence as the myriad of activities or separate tasks that a person performs in order to demonstrate their competence in an occupation (Gonczi, Hager, & Oliver, 1990; Hager, 1995, 2004, 2017; Hager & Gillis, 1995). This approach is concerned with the outcomes in the demonstrations of competent performance, rather than potential attributes that contribute to competent performance as indicated in the generic approach. In this approach, tasks become the competencies and evidence of competence is typically generated from direct observations of individuals' performance of "an exhaustive and compulsory list of tasks" (Griffin & Gillis, 1999, p. 33) in the actual professional practice. As reported by Lester (2015), this approach to competence was used in over 87% (35 out of 40) of the total professional standards he examined and is considered to be appealing to those seeking evidence to "tick off" (Hager, 2017, p. 204) a list of discrete tasks that a person knows and can do. However, this approach has been criticised as "conservative and atheoretical" (Hager, 1995, p. 142) and "too atomistic" (Hager, 2017, p. 204) given that it reduces a profession to a checklist of precisely defined observable tasks. It is argued that a 'sum of the tasks' does not always equate to competence as it provides "a trivial and superficial representation" (Hager, 1995, p.143) of the profession and ignores contributing attributes, the importance of professional judgement, and the complexity of competent performance in the real world (Gonczi et al., 1990; Grant, 1999; Sutton & Arbon, 1994). Hager (2004) and Lum (2004) emphasise that to define all discrete behaviours to obtain a complete description of a profession is futile and unrealistic. Huddle and Heudebert (2007) support the argument by reasoning that competence of professional practice can rarely be expressed precisely once it has have been broken down into pieces of task performance. Fish and de Cossart (2006) and Gonczi and Hager (2010) further added that describing professional

practice as a discrete list of tasks possibly hinders the development of professional expertise and results in the failure to fully capture the complexity of professional practice.

Both the behaviourist and generic approaches have their own strengths and drawbacks, and the choice of the approach is often based on practical grounds. The generic models of competence are potentially useful for informing training and education, curriculum design and aiding people's movement into new areas of practice or new ways of working while the behaviourist version is more relevant to assessment and signing someone off as fit to practice (Lester, 2014, p. 40). As argued by Hager and Gonczi (1993) while the logic is that attributes are essential for competence, they are not adequate by themselves. Similarly, performance tasks are necessary but not adequate for competence. Hager (2004) argued that neither of these approaches adequately captures the abstract and holistic nature of real-world professional practice which often involves simultaneously myriad performance tasks and a diversity of underpinning attributes that are typically prerequisites for competent performances. Hence, these approaches, by themselves, tend to view the constituents of competence as atomic and static entities, ignoring the dynamics of competence acquisition and downplaying the role of work contexts. As further argued by Hager (2004, p. 426), attributes of competence can be acquired or reinforced and progressively nurtured by learning and adapting to constantly changing work contexts and requirements. Consequently, Hager (2004) postulates that as the contextuality of work continually evolves, it is required that practitioners constantly nurture and develop new capabilities through ongoing workplace learning and collaborative practices. It could be concluded that neither of the abovementioned approaches offers a holistic understanding of competence and the underpinnings of competence growth. That necessitates the third approach.

2.1.3 The Integrated Approach

The integrated conception of competence embodies the integrated schema of attributes and performance, committed to describing professional practice with a significantly increased

degree of holism (Hager, 2017). This integrated approach conceptualises competence as the demonstration of skills, knowledge, and attitudes in the execution of well-selected realistic professional tasks that are of a relevant level of generality (Hager, Gonczi, & Athanasou, 1994). As succinctly summarised by Hager (2017), competence is "contextualised capability involving an integration of assorted practitioner attributes" (p. 206). These attributes range from cognitive skills, technical skills to interpersonal skills and affective attributes. It is implied that competency standards developed within this approach are those in which core professional tasks are mingled with the attributes required for successful performance, making both attributes and performance on key professional tasks essential constituents of competence. One distinguishing feature of this approach is that it overcomes the problems of atomising professional practice into myriad specific and discrete tasks. Instead, the approach focuses on choosing key tasks that are critical to the professional practice. The attributes requisite for the proficient performance of the tasks will be subsequently discussed. As a result, the issue of separating generic attributes from their contexts is overcome. Essentially, the concept of competence is relational as it connects attributes with key tasks.

Over the last three decades, the employment of the integrated approach to competence has ensured that the holistic richness of a range of professional practices that could not be achieved by either of the other approaches could be successfully captured (see Gonczi, 1994, 2013; Hager, 2004; Hager & Gonczi, 1998; Stone, Boud, & Hager, 2011). It is also reported that this approach to competence facilitates professional development and assessment frameworks design (Hager, 2017).

Of the three approaches to conceptualising professional competence, Gonczi et al. (1990) strongly recommended the integrated approach to the analysis of conceptualisation of professional competence be adopted so that both professional attributes (knowledge, skills, and attitudes) and professional tasks that constitute professional competence are captured. Given the

scarcity of research closely pertinent to multiple-choice test item writing competence, the current study drew on concepts and approaches from competence research in various disciplines. The early 21st century has observed a widespread employment of this holistic approach to the development of competence models, competency standards, and competency assessment instruments in Australia, America and the Asia Pacific in a wide range of disciplinary areas such as psychology (see Nash & Larkin, 2012; Rodolfa et al., 2005); nursing and healthcare services (see Homer et al., 2007; Nicholson, Griffin, Gillis, Wu, & Dunning, 2013; Palermo et al., 2016); teaching (see Griffin, Nguyen, et al., 2004; Zhu & Wang, 2014); management and leadership (see Department of Education Victoria, 2007); hospitality and tourism (see Suhairom, Musta'amal, Mohd Amin, & Johari, 2014); orthotics/prosthetics (see Ash, O'Connor, Anderson, Ridgewell, & Clarke, 2015), to name a few. Though these studies were not closely pertinent to the current study on EFL teacher competence in multiple-choice test item writing, they contributed different configurations for holistically understanding the professional competence of interest, operationalising and extending the original notion of integrated competence. Thus, these studies provided helpful conceptual approaches that could be adapted or adopted to define multiple-choice test item writing competence and methodological approaches to developing the framework for Vietnamese EFL teacher competence in multiple-choice test item writing to inform self-development, training and practice of Vietnamese EFL test item writers. The following section presents four models of integrated competence that emerged from the abovementioned research.

2.1.3.1 The triangular model.

The integrated approach to competence is elaborated and illustrated by Potolea and Toma (2010) and Potolea (2013). Potolea (2013) employed this approach to define research competencies for higher research programs in the social-humanistic field. In Potolea's (2013) research, it is proposed that competence comprises elements such as professional roles/tasks,

performance standards, knowledge, skills, personality traits and the context within which the competence functions (see Figure 2.1). This integrated model holds that competence is the ability to select, organise, and use complex combinations of internal resources or attributes (knowledge, skills, and attitudes) to solve professional tasks, relying on the task to be solved and the features of the context. This can be considered as an elaborated version featuring the integrated approach to and the context-dependent nature of professional competence.

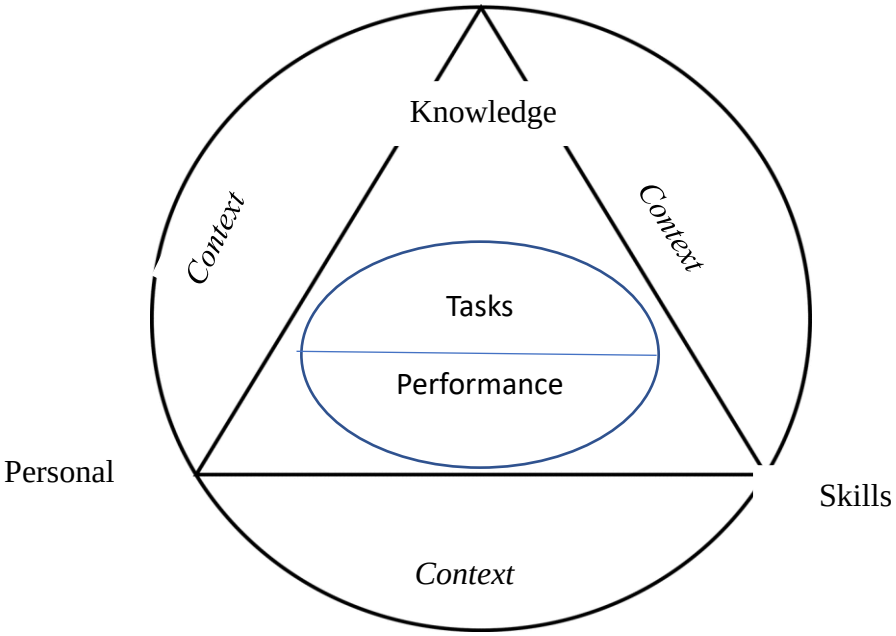


Figure 2.1 An integrated model of competence
Reprinted from "Doctoral studies and research competences" by D. Potolea, 2013, *Procedia – Social and Behavioral Sciences,* 76, p. 439. Copyright 2013 by Elsevier, Ltd.

2.1.3.2 The cube model and the pyramid model.

Two other variations of competence models from the integrated perspective, namely, namely the cube model and the pyramid model, are found in the literature of professional competencies. These models are useful structures representing additional efforts to define professional competencies in a holistic way (Nash & Larkin, 2012). The cube model of competence put forward by Rodolfa et al. (2005) for use in describing competency development of professional psychologists. The cube model conceptualises competence as three dimensions (see Figure 2.2). The first dimension represents the domain of skills, knowledge, attitudes, and

values that are described as the foundational competencies. The second dimension is the domain of functional competencies broadly describing a range of tasks professionals carry out in practice. The third dimension depicts the development of foundational and functional competencies of a professional across the stages of career development. While the cube model features the holistic approach to defining competence, as found in the other integrated approaches, the distinction is that the cubical approach explicitly recognises the continuous progression of competence by associating it with career stages.

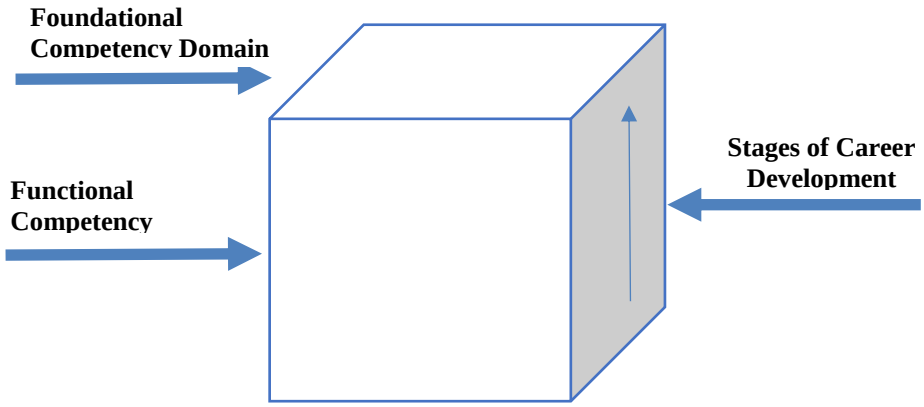


Figure 2.2 The cube model.

Adapted from "A cube model for competency development: implications for psychology educators and regulators" by E. Rodolfa, R. Bent, E. Eisman, P. Nelson, L. Rehm and P. Ritchie, 2005, *Professional Psychology: Research and Practice*, 36, p. 350. Copyright 2005 by the American Psychological Association.

However, Nash and Larkin (2012) maintained that the diverse pathways of competency development in the area of specialist psychology are not reflected in the cube model. Thus, they proposed a newer model called the pyramid model, which builds on the cube model and illustrates both competency structures and competency acquisition processes for specialist psychologists. The pyramid model depicts four "potential pathways for developing competence in specialty psychology" (Nash & Larkin, 2012, p. 42), two of which capture competency development commencing at the doctoral level and two at the postdoctoral level (see Figure 2.3). These pathways represent varying levels of experience, exposure and immersion in specialty

areas at each training level and are recommended study pathways for those aspiring to become specialists.

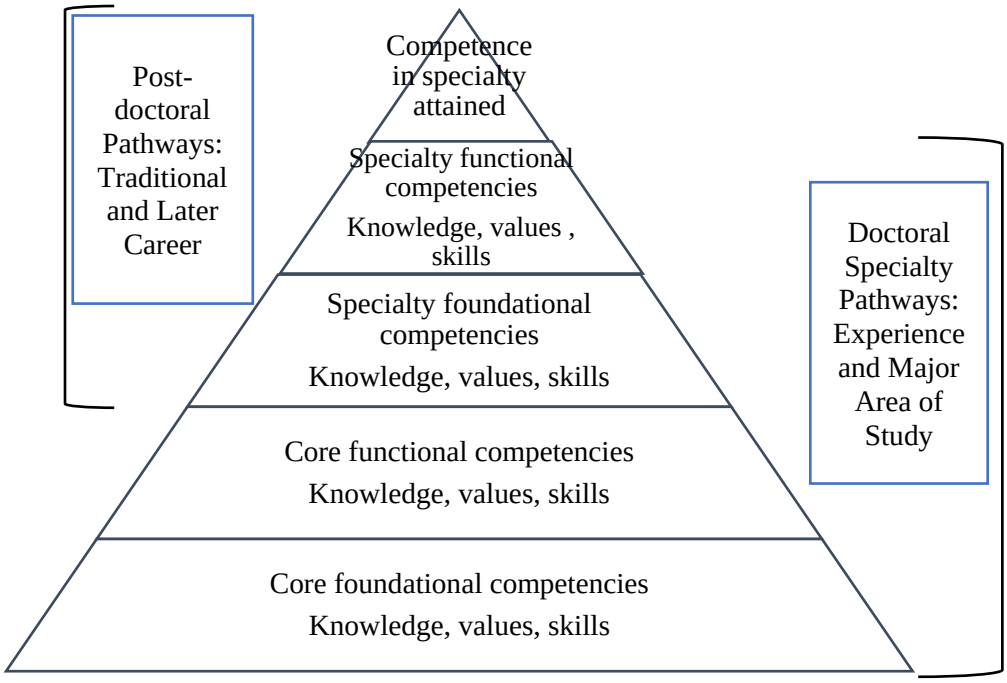


Figure 2.3 The pyramid model.

Reprinted from "Geometric models of competency development in specialty areas of professional psychology" by J. Nash and K. Larkin, 2012, *Training and Education in Professional Psychology*, 6, p. 42. Copyright 2012 by the American Psychological Association.

The integrated approach to competence provides a more comprehensive insight into the nature of competence than the behavioural and generic approaches. The cubical model by Rodolfa et al. (2005) and the pyramid model by Nash and Larkin (2012) could be regarded as the operational models of the integrated approach to competence, expanded to better indicate the acquisition pathways of competencies. These models suggest that a competence model should have a developmental approach as recognised in the cube model and the pyramid model. However, the competence model may not merely be focusing on career progression or formal training pathways and be used for a regulatory purpose (Mahony & Hextall, 2000; Power, 1997). Instead, it may be attached with a range of other developmental and educational purposes, such as to operationalise a set of aspirational competencies for professionals to develop and accomplish over a lifelong career or for professionals to reflect on their practice or specialise in

an area of practice. Masters (2003) applauds the power of the developmental uses of competence models because these models enable individuals to locate where they are on a map of increasing competence, identify current strengths and deficiency and clarify how to progress in an area of development. Ingvarson and Kleinhenz (2003) further emphasise that such models offer "the basis on which the profession can lay out its own agenda and expectations for professional development and accountability" (p. 10). An attempt could be made to operationalise the progression of competence acquisition in light of a developmental model, given that monitoring individual growth along the continuum of competence is the central idea in the era of competency-based assessment and education (Masters, 2003). The development trajectories of competence, once identified, could allow individuals to monitor their development of expertise and to facilitate professional development and performance improvement (Ingvarson,1996; Masters, 2003). This is a feature of the developmental model of competence (Griffin, 2001, 2007, 2018; Griffin et al., 2004; Masters, 2003; Masters et al., 1990) to be discussed in the next section.

2.1.3.3 The developmental model of competence.

Applications of a developmental approach to conceptualising competence by researchers such as Griffin and colleagues (Griffin, 1997; Griffin & Gillis, 2014; Griffin et al., 2007; Griffin, Gillis & Taylor, 2002; Griffin et al., 2004; Masters, 2003; Masters et al., 1990) hold much potential for an alternative approach to competence models that incorporate the notion of developmental progression. Such developmental models of competence respond to the need for a frame of reference for monitoring individual competence growth and identifying interventions and experiences to further raise professional competence (Larter, 1991; Masters, 2003). The developmental approach to defining competence potentially resonated with the aim of the current study which was to construct a developmental progression for Vietnamese EFL teacher competence in multiple-choice test item writing and inform the teachers of their next stages in

their professional development. Building on the integrated approach to defining competence, Griffin, Francis, and Robertson (2018) offer a definition of competence that encompasses both the constituents of competence and its development. The definition is:

> Competence is a latent trait that defines a person's capacity to perform tasks and demonstrate skills. It is a latent or 'hidden' personal attribute that can only be inferred from observing how much and how well a person manifestly performs a set of skills related to the trait in a range of contexts.
>
> Griffin et al., (2018, p. 100)

It is clear that in these researchers' view, competence is manifested by sets of underlying attributes (i.e., skills, knowledge and attitudes) that can be visibly displayed through the accomplishment of a series of related tasks or through the possession of core competencies. These tasks/competencies are expressed in observable and measurable behaviours/actions evidencing precisely what a person does, says, creates, and writes. The person's performances on these tasks (i.e., how well a person performs) could be used to make inferences about his/her level of competence on a continuum of development. These inferences subsequently inform educational and professional intervention activities and experiences to facilitate further growth and to identify the amount of competence he/she possesses (Griffin, 2014, 2018; Masters, 2003). The concept of a continuum of development is a central idea underlying assessment designs, curriculum frameworks, and competency standards, including the curriculum profiles for Australian schools, *First Steps* Literacy and Mathematics (Western Australian Ministry of Education, 1991), and similar initiatives in America and the Toronto Benchmark in Canada. This developmental notion of competence has also been proposed by researchers as a more appropriate or desirable approach to conceptualising professional competence for teachers. This continuum perspective has been adopted in a number of competency standards for teachers, for example the Teacher Professional Competency Standards in Estonia, Hungary, Netherland,

Vietnam and Australia (see Educational Services Australia, 2011; Griffin, Nguyen, et al., 2004; Pedaste, Leijen, & Poom-Valickis, 2019; Snoek, Dengerink, & de Wit, 2019; Symeonidis, 2019). Such a developmental model of competence is supposed to identify and describe meaningful dimensions of teachers' competence, which is useful for monitoring the professional competence development of teachers and for setting performance standards (Abbott-Chapman, Radford, & Hughes, 1993; Louden, 1994; Masters, 2003). As Ingvarson and Chadbourne (1994) argue, the developmental model of competence is useful for setting a clear direction about what teachers should "get better at" or what it means to become a 'better' teacher (p. 12). Louden (1994) further elaborated by pointing out that "progression along a continuum of competence is a much more appropriate conception of occupational competence in a profession such as teaching, where much of the competence continues to develop on the job, long after initial entry to the profession" (p. 105).

Thus, competence is conceptualised as a developmental trait as recognised in the cubical model of competence by Potoleo (2013) and the pyramid model by Nash and Larkin (2012). However, while the developmental pathways of competence in the cube and pyramid model are defined by career progressions and formal training pathways, those of the developmental model are more linked to individual growth and development of expertise over time. According to Griffin and colleagues (1997, 204, 2007, 2014, 2018), this developmental approach to competence is grounded on the following theoretical and psychometric underpinnings:

- Glaser's (1981) criterion-referencing as referring to distinct levels of competence,
- Vygotsky (1978)'s Zone of Proximal Development (ZPD) as a zone of readiness to progress to a higher level of competence, where professional learning should be optimally focused, and

- The Rasch mathematical model (Masters, 1982; Rasch, 1960, 1980; Wright & Masters, 1982) that enables teachers to be located against the proficiency indicators of their current level of competence

The combined contribution of the insights from these influential thinkers informed the developmental approach to competence. According to Glaser (1963, 1981), a person's competence could be inferred from the tasks they perform. He emphasised that criterion-referencing was the process of monitoring a person's growth through "stages of increasing competence" which are defined by tasks of increasingly complex knowledge, skills and abilities (Griffin, 2014; Griffin, Rickards, & Francis, 2018). Criterion referencing of competence, in Glaser's view requires that distinct levels of competence are identified, and these levels of competence are differentiated among themselves in an increasing order of sophistication. A crucial feature of Glaser's criterion-referenced approach is that each level of competence is described in terms of the tasks performed and the skills displayed. As a whole, the level of competence describes the observable behaviours that allow for inferences about one's capability in relation to the underlying domain of competence. Glaser (1981) further developed his idea of criterion-referenced interpretations to the idea of assessing performance quality at various levels of competence along a developmental continuum, which allows individuals to know what they have achieved, what they can do now and to predict what to do to progress to the next level of competence.

Based on the notion of Rasch (1960) regarding the relation between task difficulty and competence, Glaser (1980) consolidated the interpretative framework of competence by distinguishing one competence level from another on a continuum through a mathematical modelling procedure. The mathematician theorised that a person's competence equates with the task demand and developed a mathematical model to match a person's competence to the difficulty level of the tasks on the same metric on the grounds that when they are equal, the

probability of a person's success on the task or a cluster of tasks is 50/50. As a person's level of competence on a continuum can be mathematically located in relation to the tasks performed, a person's growth through stages of increasing competence as reflected in Glaser's criterion-referenced framework, from simple to complex, easy to difficult or novice to expert can be interpreted and monitored.

Vygotsky (1978) defined the Zone of Proximal Development (ZPD) as "the distance between the actual developmental level as determined by independent problem-solving and the level of potential development as determined through problem-solving under adult guidance, or in collaboration with more capable peers" (p. 86). ZPD is the level of development at which a person is most likely to learn if assistance and support is provided and is the zone between what a person is able to do and that he/she is incapable of. The concept of "readiness to learn" embodied in the ZPD emphasises progress towards higher competence levels. As the learner's competence is increasing, the boundaries of ZPD are changing. What one can accomplish today with guidance, they will be capable of accomplishing tomorrow autonomously. This zone can be identified only if a performance is interpreted in a developmental framework (Ingvarson, 1996; Ingvarson & Kleinhenz, 2003; Masters, 2003; Griffin & Robertson, 2014). When Vygotsky's ZPD is combined with Glaser's concept of knowledge acquisition as a continuum of increasing competence and Rasch's mathematical connection between the competence of a person and the difficulty of a task, it is possible to describe and locate their current level of competence and set directions for moving forward to reach the next level (Griffin &Robertson, 2014). Competence as a continuum entails the arrangement of observable tasks, behaviours or action manifesting the underlying ability in terms of increasing demand or difficulty so that the emerging developmental progression of the competence can be observed. Such a developmental progression is not found in the other approaches to inferencing competence. Furthermore, the continuum denotes that there does not have to be a single cut off point on. Rather, there can be

multiple levels of competence or different levels of expertise along the continuum a person can attain at different stages, illustrating a learning trajectory in which novices progress successively into experts, opening up opportunities and providing inspiration and motivation for teachers to monitor their competence growth (Abbott-Chapman, Radford, & Hughes, 1993; Griffin, 2014; Griffin et al., 2007; Griffin, Nguyen, et al., 2004; Louden, 1994; Louden & Wildy, 1998; Snoek et al., 2019; Symeonidis, 2019).

This approach to defining competence enables the inclusion of knowledge, skills, attributes and observable performance tasks and the description of how one develops their ability to complete specific tasks. This helps to unveil the nature of the competency, the typical order of its development, complexity and levels of sophistication. Most importantly, within this perspective, competence is not static, restricting itself to a list of attributes and tasks, but competence is integrative and evolutionary and can be increased through learning, experience and mentoring/coaching from more competent colleagues (Griffin et al., 2018, p. 102). In other words, the developmental model of competence contributes a more dynamic understanding of the profession, creating a potential for mapping professional learning trajectories and development paths (Epstein & Hundert, 2002; Snoek et al., 2019); thus, to some extent, competence is teachable, learnable and malleable by exposure to learning, training and practice. Given the developmental nature of competence, it is essential to explore the literature on how developmental progressions of competence could be generalised and described.

2.2 Development of Competence

Research studies in the field of skill acquisition and expertise development support the understanding that developing a higher level of proficiency or expertise is a developmental, continuous and transitional process (Alexander, 2003; Bransford, Brown, Cocking, Donovan, & Pellegrino, 2004; Ge & Hardré, 2010; Lajoie, 2003; Moon, Kim, & You, 2013; Sternberg, 1999). This finding further supports the developmental model of competence discussed in the previous

section. The researchers of these studies contend that expertise development is a process of constantly learning, relearning, consolidating, and reflecting on a set of skills and knowledge needed to master a higher level of proficiency. They maintain that expertise is not an ultimate goal nor an end product nor a defined stage but a continuum where the novice is at one end and the expert is on the other. Individuals move along the continuum of the novice to expert transition as they gain more practice and experience in their field of practice. Various skill acquisition and expertise development models and theories generated by leading researchers could be employed to describe increasing levels of competence. According to Attri's (2019) review, there are at least twenty- three major models and theories that explain expertise development, reflecting different views on how expertise is developed. Attri's (2019) systematic review of the literature suggested that the most prominent models for describing expertise development appear to be the stage-based models, which propose that there is a staged progression in the novice-to-expert transition. The notion of stages supports the view that developing from a novice to a higher level of proficiency and expertise is a process and competence is developmental. Some of the classic and more recent stage-based models for describing levels of competence are the Conscious Competence Theory, Zone of Proximal Development (Vygotsky, 1978), Staged Skill Acquisition Model (Dreyfus & Dreyfus, 1986), Proficiency Scale (Hoffman, 1998), Levels of Human Competence (Jacobs, 1997), and Novice to Expert Transition (Rosenberg, 2013).

2.2.1 Conscious Competence Theory

The earliest stage-based model for describing expertise development may probably be the Conscious Competence Theory (Attri, 2019). The origin of conscious competence theory is credited to Lewis Robinson (1974), however, its theoretical grounding remains unknown. The model demonstrates development of expertise as the progression from Unconscious Incompetence through Conscious Incompetence and Conscious Competence to Unconscious

Competence (see Figure 2.4). At the unconscious incompetence stage, a person is not aware that he/she does not have a certain skill or may not know the relevance and usefulness of the skill. As the person moves to the conscious incompetence stage, he/she begins to recognise his lack of the skill and works to learn and practice the skill. Then, he/she is in the conscious competence stage where he/she can perform the learned skill with intensive concentration and attention. Eventually, the person reaches the unconscious competence stage in which the skill becomes automatic. This model is also subject to debates on whether the last stage should be unconscious competence or conscious competence, when the individual can justify what he/she is doing (Cheetham & Chivers, 2005). It is questionable whether an individual could stay unconsciously competent all the time without continued practice for improved competence or without continuing to learn new skills and knowledge (Attri, 2019). Chapman (n.d) also argued that regression back to previous stages might be necessary for acquiring more advanced or new skills, which is not represented in the model. Another issue is that the model merely provides a conceptual explanation of how an individual learns but does not offer an indication of how to measure or quantify the individual's attainment at each stage (Attri, 2019).

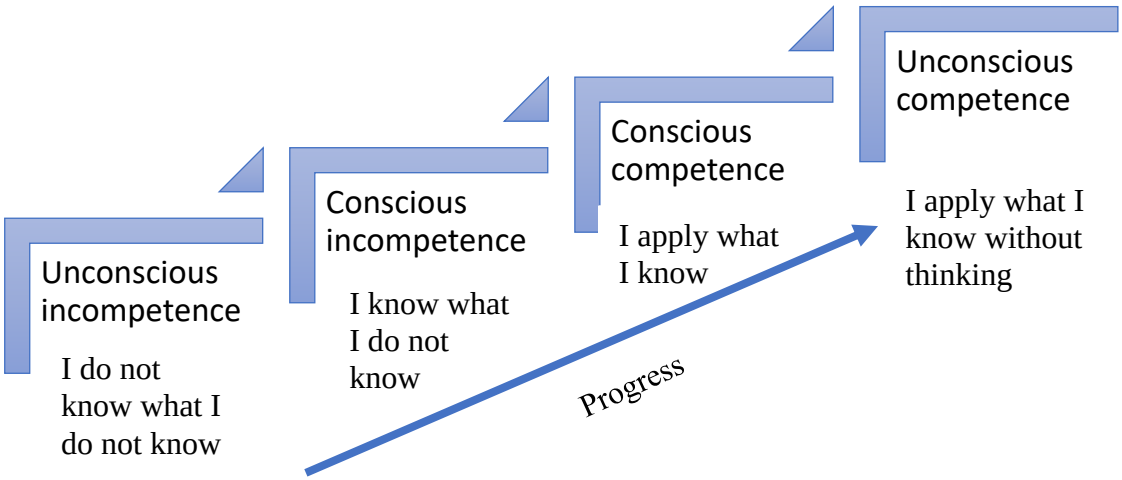


Figure 2.4 Representation of Conscious Competence Theory.

Adapted from "The models of skill acquisition and expertise development: A quick reference of summaries", by R. Attri, 2019, Singapore: Speed to Proficiency Research. Copyright 2019 by Speed to Proficiency Research

2.2.2 Zone of Proximal Development

Based on the premise that social interaction leads to cognitive development, Vygotsky (1978) developed a theoretical framework of the development of cognition explaining how learning occurs. Vygotsky (1978) argued that cognitive development depends on the "Zone of Proximal Development" (ZPD), which he defined as "the distance between the actual developmental level as determined by independent problem-solving and the level of potential development as determined through problem-solving under adult guidance, or in collaboration with more capable peers" (p. 86). ZPD is the level of development at which a person is most likely to learn if assistance and support is provided and is the zone between what a person is able to do and that which he/she is incapable of. The concept of "readiness to learn" embodied in the ZPD emphasises progress towards higher competence levels. As the learner's competence is increasing, the boundaries of ZPD are changing. What one can accomplish today with guidance; they will be able to accomplish tomorrow autonomously. From the expertise development perspective, Bransford et al. (2004) emphasised that "ZPD embodies a concept of readiness to learn that emphasizes upper levels of competence" (p. 81). Vygotsky's (1978) framework implies that proficiency grows when one learns in both formal and informal contexts and with capable peers, mentors and colleagues. The social aspect of learning and developing proficiency in Vygotsky's (1978) framework is also found in other models such as Bandura Social Learning Theory, Lave and Wegner's Situated Learning Theory, and Collin's Cognitive Apprenticeship Model. Vygotsky's (1978) ZPD framework suggested that learning occurs in four stages (Dunphy & Williamson, 2004):

- Assisted phase: In this phase, an individual can perform a task if they are provided with accurate and personalised assistance from competent peers and experts
- Self-assistance phase: An individual can perform the task without any assistance; however, performance is not yet fully automatic.

- Automated performance phase: The task is internalised and performance is fully developed and automated. No assistance is required. This point marks the end of the ZPD.
- De-automatisation of performance: A practitioner may regress back through ZPD again in case he/she can no longer do what he/she formerly/previously could do. At this phase, the goal is enhancement, improvement and maintenance of performance.

This model addressed some criticism targeted at the Conscious Competence Theory as it indicates that the developmental process may be recursive in order to enhance, maintain or restore a skill or capability if the skill or capability may deteriorate. One limitation found in the ZPD framework is that it does not specify in detail how one transits from independent performance to expertise (Dunphy & Williamson, 2004).

2.2.3 Staged Skill Acquisition Model

The most recognised work in identifying expertise development stages was the one put forward by Dreyfus and Dreyfus (1980, 1986, 2008, 2009) and supplemented by Dreyfus (2004, 2005). The Dreyfus and Dreyfus model of skill acquisition is grounded on the conception that skill acquisition is an evolving process and only experience in concrete situations results in higher levels of competence.This model is pertinent to the nature of the integrated, developmental notion of competence which views competence as the ability to apply knowledge, skills and attributes into executing performance tasks in real-world contexts, and acknowledges that competence is evolutionary, contextual, and impermanent, requiring a commitment to ongoing development throughout one's career lifespan. The Dreyfus and Dreyfus model of skill acquisition allows for describing developmental pathways of competencies in a holistic way. Dreyfus and Dreyfus (1986) proposed the five-stage theoretical model of skill acquisition which reflects a "progression from analytic behaviour of a detached subject, consciously decomposing his environment into recognizable elements, and following abstract rules to involved skilled

behaviour based on an accumulation of concrete experiences and the unconscious recognition of new situations as similar to whole remembered ones" (p.35). According to S. E. Dreyfus (2004), the development of competence occurs over five specific stages: Novice, Advanced beginner, Competent, Proficient and Expert. The model can be summarised as follows (see Table 2.1).

Table 2.1 *Stages of Skill Acquisition from Novice to Expert*

Level	Stage	Characteristics
1	Novice	Rigid adherence to taught rules or plans. No store of contingency plans for when circumstances vary from the expected. Little situational perception.
2	Advanced Beginner	Minor adjustment to rules or plans can be carried out in some situations under supervision.
3	Competent	Perceives actions at least partially in terms of long-term goals. Conscious, deliberate planning and use of standardised and routinised procedures. Makes reasoned decisions about new situations without being sure of appropriateness of decision.
4	Proficient	Sees situations holistically and identifies goals or salient aspects intuitively. Perceives deviations from the normal pattern and possesses perspectives on a situation. Decision making is less labored.
5	Expert	No longer relies on rules, guidelines or maxims, but operate from a deep tacit understanding of the total situation. Analytical approaches are only used in novel situations or when a problem occurs; Visions of what is possible.

Note. Reprinted from *Mind over machine: Human intuition and expertise in the era of the computer* (p. 50), by H. L Dreyfus and S. E Dreyfus, 1986, New York: The Free Press. Copyright 1986 by the Free Press.

The novice stage is featured by clearly recognizing relevant, individual facts and characteristics and relying on provided or learned rules, guidelines and clearly defined procedures for decision making (Dreyfus & Dreyfus, 1986). Novices lack a coherent comprehension of the whole task, so they will approach tasks mechanically and spend time

memorizing facts and are only capable of following rules and plans. They perceived their own performance based on their adherence to instructions.

The Advanced Beginner primarily uses “recipe” and begins to recognise situational elements based on some prior practical experiences and relies on rules and plans to solve the problem. At this stage, practical experience rather than verbal description is the most efficient way of learning. Advanced beginners are keen learners who are attuned to feedback, learn from mistakes, and from observing expert colleagues in actions (Dreyfus & Dreyfus, 1986).

The Competent has now gained a mental store of concrete examples in the area of expertise so they are capable of managing contingency plans (Benner, 2001) and making reasoned decisions with new problems though they may be uncertain if the decision is appropriate. They consciously and deliberately plan and use standardised and routinised procedures. They perceive contexts as critical; thus, they develop a sense of whether using rules is pertinent to a given context or not (Flyvbjerg, 2001).

The Proficient will find him/herself “thinking analytically about what to do while intuitively organising and understanding his task” (Dreyfus & Dreyfus, 1986, p. 29). They recognise and make intuitive judgements of situations simultaneously and then consciously analyse options to take actions based on previously successful results (Flyvbjerg, 2001). They have developed deep understanding of rules, theories and alternatives, and they are able to vary their uses of maxims for guidance in accordance with the situations (Benner, 2004).

The Expert can perform fluidly and efficiently. This may be because that they can “retrieve important aspects of their knowledge with little attention effort” (Persky & Robinson, 2017, p. 76). They respond to context based on previous experiences without consciously thinking about it; as Dreyfus and Dreyfus (1986, p. 31) put it, “when things are proceeding normally, experts do not solve problems and do not make decisions; they do what normally works”. The expert can perceive subtleties in the situation and rely on “creative search and cue

sensitivity". The expert makes decisions based on intuition and experience rather than calculation (Flyvbjerg, 2001) as they are independent of rules, guidelines or principles. When new situations arise or problems occur, they use analytical approaches to solve them.

In their later work, Dreyfus and Dreyfus (2001, 2008) include the sixth and seventh stage called 'mastery' and 'practical wisdom' beyond the expertise level as a response to Flyvbjerg's (1990, 1991) criticism that the Dreyfus and Dreyfus model did not take into account practical wisdom and progressive innovation. As summed up by Khan and Ramachandran (2012), a practitioner at these stages "sets new standards of performance, mostly deals with complex situations intuitively, has a unique vision of what may be possibly related to the given task, and is able to train other experts at national or international level" (p. 5).

Opponents of the Dreyfus and Dreyfus model may have questioned the validity of the model. For example, Day (2002) raised the question of the practicality of applying the 5-staged model in professional settings. Day (2002) noted that in real work contexts, practitioners can perform a range of tasks and they may not fit in one stage for all the tasks that they carry out. Another concern is its lack of objective measurement of attainment at each stage and lack of operational definitions of intuition (Peňa, 2010). Despite these concerns, the Dreyfus and Dreyfus model has proven to be a useful model for describing levels of expertise in various professions, namely in software engineering (Page-Jones, 1998), in nursing (Benner, 2001, 2004; Ramsburg, 2010), healthcare (Khan & Ramachandran, 2012), and correctional services (Scobey, 2006). Among these, the best-known study is that of Benner (2004), who applied the Dreyfus and Dreyfus model in nine studies over a period of 21 years. Benner (2004) found that the Dreyfus and Dreyfus model was "predictive and descriptive of distinct stages of skill acquisition in nursing practice" (p. 198). The Dreyfus and Dreyfus model has been effectively employed by high-profile educational researchers (Berliner, 2004; Goodwyn, 2011; Hallam, 2010). In particular, Goodwyn (2011, 2017) and Enow and Goodwyn (2018) have used this model to

analyse how English teachers develop their competence in learning design. The Dreyfus and Dreyfus model has also been effectively applied for over three decades and provided a robust means to establish reasonable expectations of a Novice and a clear pathway to progress to the Competent and to evolving into the Expert (Enow & Goodwyn, 2018). The Dreyfus and Dreyfus model provides a realistic frame of reference to examine the development continuum of competence grounded on the understanding that expertise develops through engaging in increasingly complex problem solving, and by progressing from novice, competent, and finally expert. The theory of professional expertise underlying the Dreyfus and Dreyfus model of skill acquisition is grounded on the assumptions that competencies required to perform a job well can continually evolve during one's career and skills are acquired through instruction and experience shifting from controlled processing towards automatic processing. This model acknowledges the complex nature of profession and the required time to develop one's expertise, which is pertinent to the nature of integrated, developmental notion of competence.

On a similar vein to that of the Dreyfus and Dreyfus model, Hoffman (1998) proposed a Proficiency Scale to define novice and experts in terms of progression of knowledge and capabilities based on European Middle Ages' craft guilds. Hoffman (1998) proposed that a novice develops into a master through seven stages: Naïve, Novice, Initiate, Apprentice, Journeyman, Expert, and Master. Apprentice and Journeyman correspond to Advanced Beginner and are either competent or proficient in the Dreyfus and Dreyfus model respectively. Hoffman's (1998) Proficiency Scale was recognised as a means of describing the progression of increasing knowledge and skills as one moves from novice to master. However, it is criticised for its lack of scientific validation (Farrington – Darby & Wilson, 2006). Similarly, Jacobs (1997) proposed that human competence could be scaled as novice, specialists, experienced specialists, expert, and master based on the outcomes one is able to produce. According to Jacobs' (1997) Human Competence Scale, novices are the ones who produce no outcomes or less valuable outcomes

while experts are those who attain the most valuable outcomes in organisations. Conceptually, this scale is closely aligned with Hoffman (1998), but it approaches the stages of expertise development from a human resource development standpoint; thus, it might be more appealing to organisations (Attri, 2019).

One of the latest stage-based models of expertise development is the Novice to Expert Transition developed by Rosenberg (2012) which explained expertise development as four stages: novice, competent, experienced and master/expert. Each stage is characterised by how an individual performs the job at each stage of progression. According to this model, a novice knows little or nothing about the job and can hardly perform to an acceptable standard. Competent workers can perform the tasks to basic standards and require coaching and practice to perform better. Experienced workers can search for information from resources or coworkers to perform tasks based on specific situations. Master/experts are able to invent better or new ways of doing a job, create new knowledge, and can teach others. This model is considered as a good guide for training program designers because it shows an integration between learning and job performance (Attri, 2019).

Of all the stage-based models for describing expertise discussed above, Hoffman's (1998) Proficiency Scale, Jacobs's (1997) Levels of Human Competence (Jacobs, 1997), and Rosenberg's (2013) Novice to Expert Transition seem to be suitable for describing levels of competence from the generic or behaviourist perspective, which elaborate competencies in terms of knowledge, skills, attitudes or tasks. Conscious Competence Theory and Vygotsy's (1978) ZPD framework provide a conceptual understanding of learning and skill acquisition which is useful for educational program and instructional design. The Dreyfus and Dreyfus model is most appropriate to holistically describe professional learning pathways for practitioners, which is pertinent to the current research aiming at understanding Vietnamese teacher competence in

multiple-choice test item writing for English reading and listening skill tests on a developmental progression.

2.3 Methodological Approaches to Deriving Competencies and to Competency Framework Development

Different approaches to conceptualising competence lead to different methodologies being used to unravel the competencies required for successful performance in a profession (Gonczi et al., 1990). The major distinction between these methodologies is those that concern the individuals' attributes that contribute to one's ability to demonstrate their competencies, and those that focus on the competent activities or functions to be performed. These methodologies reflect the two major traditional approaches to competence: the generic versus the behaviourist approach reviewed earlier in Section 2.1.1 and 2.1.2 (Eraut, 1989; Lester, 2014). Some authors such as Lester (2014) and Potolea (2013) categorised methods for deriving competencies in accordance with these two broad approaches to the notions of competence. Accordingly, within the generic approach, the purpose is normally to generate a list of broad statements of knowledge, skills and dispositions associated with competent job performers that are considered difficult to translate into practice (Gonczi, 1994; Reeves et al., 2009). These attributes are mainly identified through document analyses such as analyses of scholarly literature, occupational analysis, course syllabi and learning needs analysis, behavioural event interview, critical incident analysis and repertory grid technique (Lester, 2014). Meanwhile, within the behaviourist perspectives, the aim is to identify lists of descriptions of job functions and detailed activities within them that professionals are required to undertake. The lists of these competencies are usually unravelled through functional analysis. Similarly, Gonczi et al. (1990) outlined twelve major methods for deriving competencies without labelling these methods as being associated with either the generic or behaviourist views of competence. However, Gonczi et al. (1990) advised that the choice of technique be guided by the profession's needs, purpose of use and

other practical considerations. Gonczi et al. (1990) also recommended that a combination of the methods be used to analyse both tasks and attributes required for professional work so as to capture the full complexity of the profession which reflects the integrated/holistic approach to competence that Gonczi and colleagues advocated. Emerging from the literature are four major approaches to unravelling the competencies required for successful job performance: the functional analysis approach, the actuarial approach, the consensus approach, and the multi-methods approach.

2.3.1 The Functional Analysis Approach

The functional analysis approach is claimed to promote the behaviourist/task-based view of competence and is a deductive approach to splitting professional roles into increasingly detailed depictions of job functions (Lester, 2014). The functional analysis process commences with establishing a Key Purpose statement that captures the unique contribution of the job and is further developed to identify the core functions within the professional role to form a functional 'map'. It is followed by disaggregating the core functions into key tasks and competencies required to fulfil the task. The process of disaggregation is carried out by expert groups and guided by a simple question "What needs to happen for this (key purpose, tasks) to be achieved?" (Gonczi et al., p. 43) and it is repeated until all behavioural indices of the competencies are obtained (Carroll & Boutall, 2011; Gonczi et al., 1990). The functional analysis approach usually uses a consultative process that engages practitioners, leaders and relevant stakeholders.

Functional analysis had been in use as an approach to competency identification, modelling and assessment in the United Kingdom (Lester, 2014), Canada and America, though named differently (Rothwell & Lindholm, 1999), since the 1990s. Rothwell and Kazanas (1999) identified five approaches to competence modelling applied by organisations in the USA, namely: the process-driven approach, the outputs-driven approach, the invented approach, the

trends-driven approach and the work responsibilities-driven approach. The main objective of all the approaches is to derive a list of competencies that characterise exemplary job performance. The competencies could be documented based on extensive investigations of the work duties, responsibilities, and roles through document analysis or focus groups, and behavioural event interviews which are extensive face-to-face interviews which involve eliciting critical incidents from interviewees and recording their thoughts, feelings and actions during the incident (McClelland, 1978). All these approaches use a panel of experts who are nominated exemplary job performers and fully successful performers, except for the invented approach, which used either decision makers or external subject matter experts when it was believed that no exemplary performers existed in the organisation. The trends-driven approach is different from the others in that it places emphasis on what people must know and do in order to react to arising changes in the work environment, not on what people do in real-time. The work-responsibilities approach includes the presence of a facilitator, often the researcher-in-charge who leads discussions among a representative group of expert practitioners and relevant stakeholders, such as managers, and researchers. These participants are asked to respond to a preliminary broad analysis of the profession or to list the functions, responsibilities and behaviours they perform until they can no longer think of any.

The functional analysis approach used in America is similar to the modified functional analysis Griffin and colleagues used to derive competencies for Vietnamese primary teachers and Australian school principals (see Griffin et al., 2004; Department of Education Victoria, 2007). When a more refined analysis is made based on group discussion and suggestions, the representative group work further to rate the tasks according to their essentialness or importance. The outcomes are then mapped and taken as a basis for further amendment and refinement. The derived list of competencies based on the functional analysis and its modified approaches are normally reviewed and verified by the representative groups of experts or exemplary job

performers or another group of exemplary job performers through surveys. Alternatively, the facilitator may invite representative groups to write responses to the proposed analysis and conduct a series of interviews to define the analysis. The same techniques are also applied to define knowledge, skills, and attitudes required to successfully carry out the tasks and to set standards expected at various levels of performance.

Similarly, the Canadian Department of Manpower and the General Learning Cooperation of New York are credited with being the original developers of DACUM, an acronym for Developing A CurriculUM in the 1960s. This technique was quickly adopted and adapted in the education sector as an effective approach to curriculum development. DACUM used focus group discussion with experts to solicit areas of practice, specific tasks and required competencies. In the DACUM procedure, around 12 expert workers are carefully selected with one facilitator whose duties are to elicit and record all the data provided by the experts. Initially, the facilitator collects ideas from the group about the major duties of the profession. Subsequently, the constituent tasks associated with the listed duties are analysed. Then the analysis goes further to determine the general knowledge and skills, personal traits and interpersonal skills (competencies), tools and equipment enabling effective task performance (Fretwell, Lewis, & Deij, 2001; Lester, 2015). In addition, the group of experts can be requested to rank the tasks/competencies on several dimensions (e.g., frequency of task performance, level of importance) and set minimum threshold standards for the tasks. DACUM is generally useful as it provides an efficient, systematic and thorough approach to the analysis of professional tasks at moderate cost while the process-driven functional approaches reviewed above could be time-consuming and potentially expensive (Rothwell & Kazanas, 1999).

One of the criticisms about the functional analysis approach is with regard to its use of experts. First, the validity of functional analysis is reliant on the capacities of the experts participating in the analysis process since they play a vital role in the process of translating their

knowledge, experience and judgement into descriptions of tasks (Gonczi et al., 1990; Lester, 2014). Second, functional analysis has been criticised for not being able to reflect the complexity of real-life professional work as it is a deductive process led by a small group of experts rather than based on research into what professionals actually do (Eraut, 2004; Grugulis, 2000). Substantial modification to the original functional analysis has been made by research practitioners to meet this criticism. For example, a newer version of functional analysis is based on extensive research into what practitioners conduct in their real practice, and is believed to yield relevant and usable outcomes (Lester, 2014).

Another criticism is mainly related to the behaviourist/task-based approach to competence to which functional analysis is tied. It has been criticised as 'overly reductionist' with an excessive focus on tasks and how the tasks should be carried out (Battel-Kirk, Barry, Taub, & Lysoby, 2009, p. 17). As this approach works deductively from an occupational purpose, functional analysis tends to produce exhaustive descriptions of a profession's broad functions and excessive (sometimes, repetitive) detailed activities within them. As a result, competency standards developed from a functional analysis approach may contain a mix of critical and trivial competencies with excessive and sometimes unused details (Lester, 2014). It is argued that such standards are too 'prescriptive, static and closed' (Lester, 2014; Whitehead, 2013; A. Wolf, 2001), failing to address the evolving and context-driven nature of competence. Besides, by focusing on reflecting in a most realistic way the outputs of competent activity in a profession, the functional approach overlooks or ignores who the competent practitioners are and what contributes to their competence development (Cheetham & Chivers, 2005; Gonczi et al., 1990). Such standards might not be helpful to guide practitioners in their practice, support their ongoing professional development and to road map their expertise development. Given these limitations, it is suggested that functional analysis should be used with some modifications and/or in combination with other approaches (Lester 2014).

2.3.2 The Actuarial Approach

The actuarial approach typically involves observations of highly successful individuals or teams of expert practitioners in their workplace. The observations are further consolidated with structured/unstructured or semi-structured interviews and/or written reflections or descriptions by the practitioners. The critical incident technique may be employed to seek reasoning and explanations. In this procedure, the observed professionals are asked to recall incidents from their practice and give detailed information regarding events leading up to the situation and factors critical in determining the outcome. The critical incident technique has the potential to provide researchers with insights into the tacit knowledge that experts use in successful performance and data about factors that characterise outstanding professional competence (Gonczi et al., 1990; Griffin, Nguyen, et al., 2004). Information collected from these procedures form draft statements of competence, and then standards are written and validated by experts as with the functional analysis approach. This approach reflects the integrated - also termed the 'hybrid' - model of competence (Lester, 2014, p. 41). It often results in competency standards that include related attributes listed alongside the competence statements and a set of desirable behaviours. Although the actuarial approach tends to generate fewer and more generic descriptions of competence than the functional approach, ultimately, it is still criticised for its reliance on subjective judgements and experiences of a group of experts. Therefore, competency standards developed through the actuarial approach may be subject to regular revision (Griffin et al., 2004). It could be argued that both functional analysis and actuarial approaches, though employing different methods to describe competence, end up yielding either a list of discrete tasks or a list of knowledge, skills and attributes required for competent performance or a list of both tasks and attributes if they are used in combination. These 'checklists' have limited usefulness for aiding the development and assessment of the competence because they provide

no insights into what the current competence level of a practitioner is and how their competence evolves as practitioners move through different stages of their professional development.

2.3.3 The Consensus Approach

The consensus approach was initiated by an inter-disciplinary panel of experts from a number of health schools in Germany, Austria and Switzerland. There are several consensus methods, namely consensus development conferences or panels, staticised groups, social judgement analysis, structured discussions, Glaser's State of the Art approach and informal consensus forming (Kiessling et al., 2010, p. 262). The Delphi Method and the Nominal Group Techniques are probably the most popular consensus methods (Lloyd-Jones, Fowell, & Bligh, 1999; Stewart, 2001). It is also recommended that multiple methods and a systematic review of the literature should be applied (Kiessling et al., 2010). The major aim of the consensus method is to reach an expert consensus for defining competencies in a structured environment where experts can express their opinions anonymously, overcoming the disadvantages of consensus forming in face-to-face focus groups or workshops, such as individual dominance and iteration.

The Delphi Method and the Nominal Group Techniques (NGT) have been well researched and proposed to be feasible to apply in various professional contexts. NGT and Delphi have been long used to develop professional competencies almost world-wide from Europe, Canada and America to New Zealand and Australia (Kiessling et al., 2010; Lindberg, Lundström-Landegren, Johansson, Lidén, & Holm, 2012; Moaveni et al., 2011; O'Connell & Gardner, 2012). NGT is employed in conducting workshops with experts to derive targeted competencies which are subsequently validated using the Delphi method. It includes three main team working phases: brainstorming, discussion, and ranking. At the team meeting, participants contribute ideas to the facilitator who keeps record of these ideas on a flip chart. After that a group discussion is carried out to clarify and evaluate each idea. Then, ideas are once again ranked and discussed. It is an iterative process between small groups and plenary discussions,

usually guided by the literature research conducted prior to the workshop by the facilitator. After the workshop, all competencies are refined. The competencies are subsequently validated using the Delphi methods.

Originally devised and first used by the RAND Corporation as part of a military defence project in the late 1940s and documented for uses for non-military purposes in the early 1950s (Landeta, 2006), the Delphi technique was named after the famous oracle at Delphi (Hasson, Keeney, & McKenna, 2000). In essence, the Delphi method is "a structured, iterative multiple – step process" (Linstone & Turroff, 2002, p. 3) used to elicit "the most reliable consensus of opinion of a group of experts by means of a series of intensive questionnaires with controlled opinion feedback" (Landeta, 2006, p. 468). The Delphi method has been rapidly spread and adopted for use in a wide range of research themes, and a lot of variants of the classical Delphi method have come into existence, namely the 'modified Delphi ' (McKenna, 1994), the 'policy Delphi ' (Crisp, Pelletier, Duffield, Adams, & Nagy, 1997), and the 'real-time Delphi ' (Beretta, 1996). In the traditional Delphi, a prepared questionnaire is often presented to a panel of experts in a specific field to seek their opinion on a technical topic for several rounds until consensus of opinion is reached. The 'modified Delphi' is different in the way the items for the questionnaire are created. Instead of asking the participants to react to the prepared list of items, they are requested to create the items based on a list of research priorities. In contrast with the traditional Delphi, the prime objective of the 'policy Delphi' is not to generate a consensus but to solicit the strongest opposing views on the possible solutions of a policy-related issue providing a platform for arguments for and against an issue or policy (Turoff, 2002). The 'policy Delphi' is different from the traditional Delphi not only with regard to the type of topic covered and the goal to be achieved but also the use of participants. While the traditional Delphi tends to use a homogeneous group of participants, the 'policy' Delphi chooses a wide variety of stakeholders who are likely to be affected by the targeted issue or policy. The 'real-time Delphi' is more

concerned with the method for collecting and synthesising expert opinions. With the use of the Internet and computer technology, the questionnaire can be done online and the results are updated and aggregated as responses are recorded in real time. The participants can revisit the questionnaire as many times as they want. This is considered to increase the efficiency of the consultative process.

Intensive search of the literature and databases conducted by Gupta and Clarke (1996) between 1975 and 1994 and by Landeta (2006) between 1995 and 2004 showed significant and massive applications of the Delphi techniques in medical, technological and economic research and its extent of use in research as a theme or research method did not decline. In a Delphi procedure, a panel of experts anonymously completed a structured questionnaire on different rounds. Qualitative data used to inform a first- round questionnaire of the Delphi is often collected through prior focus groups or interviews and a pilot test with a number of experts should precede the first- round questionnaire implementation. Statistical analyses of the data from each round is communicated back to the panellists in forms of controlled feedback and assist the formation of the questionnaire in the coming round. This process repeats until agreement or stability in each panel member's responses is achieved or until the number of returns for each round declines (Hasson et al., 2000).

The Delphi technique has several advantages that make it useful and relevant to the proposed study. First, it is an efficient method to obtain information from educational experts and to reach consensus since this technique allows the group work process to be developed even though experts do not coincide in space and time and controlled feedback is used to eliminate irrelevant exchanges of information between experts in case they meet physically. Secondly, successive questionnaires allow opinions to be considered in a non-adversarial manner and after each round the panellists are confronted with their own ratings in comparison with the mean score and standard deviation on each item. This helps to identify items that experts may have

missed, and there is opportunity for them to reconsider their viewpoints (Couper 1984; McKenna, 1994, as cited in Hasson et al., 2000). Furthermore, the influence of dominant views inherent within classic direct interaction groups is avoided, for members can have their opinions and ideas crystallised and individually consider what competencies are important and relevant.

In Delphi studies to validate competencies, a commonly seen design includes two or three rounds. One aim of these rounds is for a panel of experts to judge the relevance and rate the importance of the competencies on a Likert scale. Another aim is to categorise the competencies into domains to provide additional information and edit the wording of the competency statements (see Bok, Jaarsma, Teunissen, Van der Vleuten, & Van Beukelen, 2011; Staggers, Gassert, & Curran, 2002; Tigelaar, Dolmans, Wolfhagen, & Van der Vleuten, 2004). The major difference is some researchers have made use of existing frameworks as an input for the validation as in Tigelaar et al. (2004) whereas some have conducted a literature review and focus groups research and prepared a potential framework prior to Delphi steps as can be seen in Staggers et al. (2002) and Bok et al. (2011).

The consensus approach often involves the opinion of a wide variety of experts who participated as participants in workshops or as informants in the Delphi study. Within a broad research scope, this approach could be used to reach expert consensus at a national level employing a combination of a Steering Committee, a wider Reference Group and public consultation (Kirk, Tonkin, & Skirton, 2014, p. 418). Thus, it might overcome the criticism of being dependent on a small group of experts as with the functional analysis approach, though it still relies on subjective/ value judgement. The major strength of the approach is that it helps to preserve confidentiality, avoid some restrictions of informal consensus in a face-to-face competency development process, such as dominance by a few strong opinions of experts or authorities, difficulties in reaching an agreement and susceptibility to overloading (Kiessling et al., 2010). However, the Delphi method has been seen as possibly having a lack of scientific

rigour or questioned whether or not it is just an alternative way of organising team communication. What considers to be critical is the recruitment of experts, the study leaders, and the design of questionnaires and the defined agreement level (Kiessling et al., 2010).

2.3.4 The Developmental Multiple-Methods Approach

Given that each of the three approaches discussed above has its own merits and potential drawbacks, it would seem to be appropriate to embrace a multiple-methods approach to developing a competency framework in order to mitigate the shortcomings of the methods if used separately and to best reflect the complexities of a profession. A review of the international literature on the development of competency frameworks and standards across professions reveals that a combination of functional, actuarial and consensus methods has frequently been used, such as in the teaching profession (Zhu & Wang, 2014; Griffin et al., 2004), in nursing and healthcare services (see Ash et al., 2015; Battel-Kirk et al., 2009; Griffin, Nguyen, et al., 2004; Homer et al., 2007; Nash & Larkin, 2012; Nicholson et al., 2013; Palermo et al., 2016; Suhairom et al., 2014; Zhu & Wang, 2014), and in psychology (Nash & Larkin, 2012). The methodological approaches from these studies could be adopted or adapted to apply in the present study which aimed to explore Vietnamese EFL teacher competence in multiple-choice test item writing for reading and listening skill tests.

While there are some slight adaptations/modifications in the methods used to develop the competency frameworks, commonalities can be identified. First the development stages are often informed by an extensive literature review, related legal documents, guideline materials and existing competencies, followed by an information-gathering activity focussing on what actually happens in practice. This may be done by observing or interviewing (an) exemplary practitioner(s) or a representative group of practitioners and/or relevant stakeholders to specify the actions, content and context in their professional practice or undertaking a survey of a sample of practising professionals to document what they actually do in practice. The development

stages generally employ both functional analysis and actuarial methods, with an integration of research-based evidence from the literature, ensuring that competencies are strongly grounded on both theory and practice. The establishment of the competency frameworks is often guided by a developmental model of expertise development, namely Dreyfus and Dreyfus five-stage model of skill acquisition, depending on the intended uses of the framework and the conceptual model of competence adopted. The development stages result in a preliminary or draft competency framework that is subsequently validated using consensus methods such as the Delphi method or Nominal Group Techniques. The validation can also be conducted through or in combination with psychometric testing procedures such as Item Response Theory (IRT) modelling namely the Rasch (1960, 1980) models, the Rating Scale Model (Andrich, 1978) and the Partial Credit Model (Masters, 1982). These IRT models model the probability of a person's response to an item through a mathematical function. Through these procedures, it is possible to generate the relative difficulty of items and the relative person ability, and to position them on the same linear scale of measurement (Hambleton & Swaminathan, 1985). By locating items in order of the underlying skills required to answer each item correctly in relation to the person ability distribution, IRT models provide powerful tools to assess measurement properties of the scales and the ability range of the people who are assessed by the scales (Wu, 2003; Wu, Tam, & Jen, 2016). These IRT modelling procedures are elaborated in Chapter Five. These validation procedures, when used in combination, enable researchers to generate expert consensus regarding what is required as core competencies of a profession and to empirically derive the developmental pathways of the competencies.

What distinguishes this multiple-methods approach from the functional analysis, actuarial, and consensus approaches is the lower dependence on a small group of experts' intuitions, and the involvement of learning and measurement theories and statistical analysis in the process of constructing and validating targeted competencies. The empirical validation adds

educational values to the competency framework derived. This procedure produces not only core professional competencies expressed in observable behaviours and phrased according to developmental progressions of these competencies, but also enables the operationalisation of professional competencies along a progression. This developmental, multiple-methods approach is rooted in the concept of developmental continua applied in curriculum development, assessment design and competency frameworks and standards by Masters (1998, 2003) and Griffin and colleagues (2004, 2007, 2014, 2018), viewing competence as a progression of observable and measurable tasks ordered in increasing levels of proficiency. Viewing competence from the developmental perspective guided by learning and measurement theories, namely Glaser's (1963, 1981) criterion referencing, Vygotsky's (1978) Zone of Proximal Development (ZPD) and Rasch's (1960) mathematical model, Griffin, Gillis, and Calvitto (2004) put forward a series of principles to be applied to the construction of competency framework/standards. These principles are fully articulated as follows:

1. The development of competence can be defined by a set of underlying developmental continua, each of which defines a smaller, measurable, operationalisable construct. This is the assumption of criterion-referenced interpretations and latent trait theory.
2. The continua do not exist in and of themselves but are empirically constructed to assist in explaining observations of developmental behaviour.
3. Each continuum can be defined by a cohesive set of indicative behaviours representing levels of proficiency in the area of learning. These behaviours can be demonstrated through the performance of representative tasks and can be regarded as either direct or indirect indicators of proficiency.
4. Not all behaviours can be directly observed. Related, indirect observations can be used, along with directly observable behaviours, to describe competency.

5. The indicators (behaviours or task descriptors) may be ordered along a continuum according to the amount of proficiency, competence, or ability required for satisfactory performance or success on each task.

6. People can be ordered along the continuum according to the behaviours they are able to exhibit or the tasks that they are able to perform. The behaviours, in turn, can be interpreted to provide substantive meaning to a level of proficiency.

7. It is not necessary to identify or observe all possible indicators or behaviours in order to define the continuum. The continuum can be defined by any representative, cohesive sample of indicators that cover a range of levels on the continuum.

8. There is no one correct sample of indicators that exclusively defines the continuum or the domain. Although there may be a set of indicators that is generally agreed upon as important in defining the continuum.

9. While the indicators used to define the continuum are related, there is no causal or dependent relationship between them. It is neither necessary nor obligatory to observe lower order indicators in order to observe higher order behaviours. The existence of higher order indicators implies the ability to demonstrate lower order indicative behaviours. The relationship is probabilistic, not causal.

10. Progress along the continuum is monitored by the ordered accumulation of behaviours described by the indicators. That is, the individual is said to be progressing along the continuum and progress is monitored by the positions of the behaviours or tasks on the continuum that the individual can demonstrate.

(Griffin et al., 2004, p. 20-21)

These principles can guide the operationalisation of competence as a progression in a hierarchical structure and derived rubrics. According to Griffin (2014), the progression should be made up of five components that should be hierarchically organised. The first level is a learning

domain, a construct or a latent trait that can be described as consisting of various strands and each strand includes various statements of competencies. Each competency is expressed in concrete terms as a series of performance indicators written in the form of observable behaviours that collectively demonstrate evidence of a specific capability. Each of these performance indicators is then defined in terms of quality criteria that demonstrate different levels of performance quality. These quality criteria are specific, ordered categories that describe how well each of the indicative behaviours/performance indicators can be demonstrated. These are descriptions of varying levels of performance that are well-defined, research-based, context-specific and finite (Wilson, 2005). They are supposed to be ordered in increasing proficiency, such that – in the case of this project – the first indicator would be able to be demonstrated by novice item writers and the last indicators would be able to be demonstrated by experts. Specifically, they can be constructed based on developmental models of skill acquisition and learning (Wilson, 2005), such as the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986). The performance indicators are combined with quality criteria as rating scales to form the composite called a rubric which describes the amount of competence a person possesses; thus, they can be used to map one's performance along the developmental learning continuum, based on which competence levels are differentiated (Griffin at al., 2004). The continuum is to be expanded beyond the 'competent/not-yet-competent' dichotomy to seek information at further levels of performance (Griffin et al., 2007). This also enables teachers to identify their starting point of development on a progression, and how to move forward along the progression of increasingly complex knowledge and skills to achieve the next proficiency level (Griffin et al., 2004, 2018).

Overall, the developmental progression possesses three important properties as summarised by Griffin et al. (2018):

- It is *criterion-referenced*. The performance is not qualified by comparison with any group or expectation of group membership, as is the case with norm referencing. The performance is described purely for its level of quality, by reference to criteria that are situated within a continuum of increasing competence. It provides a series of levels that enables the teacher's development to be tracked and monitored. A content analysis of the performance statements/quality criteria that have similar levels will result in the development of summary descriptions of each level along a continuum of proficiency. These level descriptions have the potential for the teachers to identify their present performance level and what they are expected to be able to do to reach the next level.
- It is *developmental*. The levels identify the next step in development for a teacher, allowing intervention activities and efforts to be focused.
- It is *analytical*. Each descriptor of the performance is defined in ways that are recognisable and consistently judged.

This developmental multiple-methods approach could ensure that different types of evidence such as expert opinions, research findings, and contemporary conceptual models on competence, empirical data and psychometric measures have been employed to derive the competence model, which can overcome the criticisms faced by the functional, actuarial and consensus approach. This developmental multiple-methods approach has been applied and/or adapted in several competency standards and assessment instrument development research projects led or guided by Griffin and in a series of works on professional standards for teachers and school principals by Ingvarson and colleagues (2003, 2005). Some examples are:

- The developmental learning framework for school leaders in Australia (see Department of Education and Training Victoria, 2012),
- Australian professional standards for teachers and principals (AITSL, 2015),

- National professional standards for highly accomplished teachers of science (see Ingvarson, 2002);
- Primary school teacher standards in Vietnam (see Griffin et al., 2004), and
- Instruments to measure nursing competencies in the operating theatre (see Nicholson et al., 2013),
- Instruments to measure the process-oriented writing competence of Vietnamese EFL students (see Duong, 2015)
- Instruments to measure communication and literacy competence of students with learning disabilities (see Woods, 2010).

The approach presents a comprehensive process for developing competency frameworks which are theoretically, contextually and practically sound and empirically validated through item response theory procedures, namely the Rasch model (Rasch, 1960, 1980) and the Partial Credit Model (Masters, 1982). This approach transcends the traditional functional analysis and actuarial methods and has the potential to produce a more holistic and valuable competency framework from the educative perspective. The stances taken in the developmental approach to competence correspond with the rationales for and aims of the research. This study was driven by the significant and urgent demand from the researcher's work context for building teacher capabilities in large-scale English test item writing. What was to be addressed was not only an account of what constitutes teacher competence in in multiple-choice test item writing for English reading and listening skill tests, but also understandings of how the growth of these competencies could be developed, monitored, and nurtured.

Central to this methodological approach is the concept of development (Masters, 1998), which is also termed growth, progress, learning and improvement. The application of developmental models of learning in combination with a psychometric approach makes it a theoretically and methodologically sound approach to deriving competencies. Within this

approach, competencies required for a profession are holistically generated from research-based evidence and contextualised and iteratively refined by practitioners and experts. Subsequently, a psychometric model such as the Rasch model (Rasch, 1960, 1980) and the Partial Credit Model (Masters, 1982) is employed to establish the developmental progression of these competencies. These models are reviewed in more depth in Section 5.3 of the thesis. The approach offers the potential of unravelling the developmental trajectories of competence and revealing the Zone of Proximal Development (Vygotsky, 1960) which helps to scaffold teachers' competence acquisition and development. The developmental multiple-methods approach presents the potential for identification of competencies for test item writers and the acquisition and development trajectories of these competencies that is of significant value in the context of capability building and item writer workforce development in Vietnamese higher education settings.

2.4 Chapter Summary

The chapter has constructed an understanding of the conceptual and methodological approaches to modelling competence. There are three broad views of competence: the generic, behaviourist and integrated approaches. The integrated, developmental approach, which defines competence as a continuum of underlying attributes manifested in observable tasks ordered in terms of increasing proficiency, leads to a holistic understanding of the competence and the unravelling of its developmental pathways. In contrast, the others lead to atomistic views of competence, reducing competence to a checklist of skills, knowledge and attributes (Griffin & Gillis, 1999). Levels of increasing competence can be described by a range of expertise development models. However, Dreyfus and Dreyfus model of skill acquisition (Dreyfus, 2004; Dreyfus & Dreyfus, 1986) is pertinent to the nature of an integrated, developmental notion of competence that conceptualises competence as the ability to apply knowledge, skills and attributes in executing performance tasks in real-world contexts and see competence as

evolutionary, contextual and impermanent, requiring a commitment to ongoing development throughout one's career. With regard to the methodology, the developmental multiple-methods approach proposed and applied by Masters (1998, 2003), Wilson (2005), Ingvarson (2002) and by Griffin et al. (2004), which employs a combination of functional, actuarial, consensus methods used in combination with empirical procedures, would appear to be the most appropriate approach to develop a theoretically grounded, methodologically sound and empirically validated framework for teacher competence in multiple-choice test item writing for English reading and listening skill tests. The next chapter, Chapter Three, reviews and critiques existing models of item writing competence in light of their applicability to the developmental notion of and methodological approach to competence as discussed in this current chapter. In Chapter Three, the contemporary research on item writing competencies is also synthesised in order to formulate an understanding of what has been known and unknown about how these test item writing competencies are demonstrated, nurtured and developed.

CHAPTER 3: TEST ITEM WRITING COMPETENCE

The chapter analyses the two most predominantly researched models of item writing: the cognitive and knowledge-based model by Fulkerson and colleagues (2009, 2010, 2011), and the performance task-based model by Salisbury (2005, 2016). First, these models are reviewed and critiqued on theoretical, practical and methodological grounds and in light of their applicability to the developmental approach to defining competence presented in the previous chapter. Second, the chapter synthesises research in best practices in multiple-choice test item writing. The synthesis is organised around the five major competencies required of English test item writers as identified in earlier research by Salisbury (2005). Third, the chapter concludes by establishing the conceptual framework for operationalising teacher competence in multiple-choice test item writing for English reading and listening skill tests within this current research study.

3.1 Item Writing Formats

Test items can be presented in different formats, and how the test items are displayed is shaped by the dominant theoretical paradigms adopted by the examination boards in charge of developing the test (Shin, 2012). Four philosophical paradigms are currently in use: the psychometric, authentic, systematic, and critical approach.

The psychometric and authentic approach are at the extreme ends as the former favours item formats that can be objectively marked while the latter prefers behaviour and performance-based tasks, such as essay writing and oral presentation tasks that require expert judgement of quality performance. The psychometric approach has grown out of the positivist paradigm held the view that language testing should be quantitative, objective and outcome-oriented so that context free inferences could be drawn from well- crafted test items. The preferable language proficiency and achievement test format consists of multiple-choice, True-False, and matching item types and items are written based on the key principle “reproduce similar items in multiple-

choice format. Do it objectively" (McNamara & Roever, 2006, p. 1) The multiple-choice format has become a "psychometric mold" (McNamara & Roever, 2006, p. 1) into which test writers pour constructs of language knowledge to be tested. This approach has been widely employed not only in discrete point tests of grammar and vocabulary but also in standardised macro skills tests of listening, speaking, writing and reading. This tradition remains prominent in internationally recognised language proficiency tests; thus, test preparation resources are awash with multiple-choice exercises and test writers are well supplied with taxonomies of multiple-choice item writing guidelines (For example, Haladyna & Downing, 1989; Haladyna et al., 2002).

Alternative paradigms proposed by Hamp-Lyons and Lynch (1998), who view testing as a qualitative and subjective process, advocate developing diverse authentic test items and avoiding the multiple-choice format. Item writers tried to produce integrative, authentic, communicative and pragmatic items which simulate real life language use situations and to test practical language skills as a whole rather than testing discrete points of language knowledge. This approach is being exposed to criticism mostly from psychometric testers for its impracticality of test administration, subjectivity of rating scales and unreliability of raters as pointed out by Shin (2012). The communicative competence model of language testing theoretically put forward by Dell Hyme (1966), and later by Canale and Swain (1980), Bachman (1990) and Bachman and Palmer (1996) promoted this type of communicative language testing to assess productive skills for real-world communication. In light of this approach, communicative tests should utilise interaction-based, authentic, and behaviour-based performance tasks that best capture different communicative competences of language learners (Morrow, 1986; Weir, 1990).

Many language test writers who have positioned themselves as neither psychometric nor authentic purists have argued that no specific test item can make perfect inferences about

examinees' language proficiency. Some items are strong at assessing productive skills and some receptive skills. Buck (2009) emphasises that a set of comprehensively balanced and diverse test items featuring specific, desirable characteristics laid out in test specifications is likely to produce a valid test. This characterises the systematic approach, which can be considered as the blended version of the psychometric and authentic approach. The systematic approach is endorsed as evidence-centred design (ECD) (Fulcher & Davidson, 2007, p. 63), which takes its cue from the construct-centred approach offered by Messick (1994) and uses the structure of evidentiary arguments created by Toulmin (1958) in the legal sector as its conceptual groundwork (Mislevy, 2003; Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006; Mislevy & Yin, 2013; Riconscente, Mislevy, & Corrigan, 2016). In the ECD systematic approach, item writing is informed by a chain of reasoning from proof of what learners make, do and say in a testing situation to claims about their learning and what they can perform in the real world (Mislevy et al., 2003). In other words, information on the design of test items and expected responses are characterised and systematically planned and documented in the test specifications to construct "a priori validity" arguments and make inferences about one's knowledge, skills, and abilities (KSAs) from the particular things (s)he makes, does or says or from the particular tasks (s)he performs (Shin, 2013). In the field of language testing, ECD has been well received by exam boards and the testing community and applied in the development of items for the Test of English as a Foreign Language internet-Based Test (TOEFL iBT), the Test of English for International Communication (TOEIC) Speaking and Writing test, and the Advanced Placement (AP) Spanish Language Exam (Mislevy et al., 2003; M. K. Wolf et al., 2014). ECD principles are best illustrated in the procedures ETS used to develop test blueprints for the large-scale assessment of English language proficiency – the TOEFL iBT (Chapelle, Enright, & Jamieson, 2008). This approach embraces diverse methods of item development and validation and acknowledge the dynamics and complexity of item writing practices, as supported

by Davidson and Lynch (2002), who argue that the approach to test item writing is meant to be "inclusive, open, and reflective" and that virtually no test specifications "will fit all settings" (p. 2). The major principle for item writing in this approach is to do it systematically with the whole picture of the test in mind (Shin, 2012, p. 580).

The last approach is the critical approach that has not actually been put into practice in test item development because of its limitations in terms of practicality (Shin, 2012). This approach was put forward by some scholars (e.g., McNamara, 2008; Shohamy, 2001; Spolsky, 1995) who argue that socio-political dimensions of language testing and examinees' active participation in and critical contributions to the process of test design and use (Shohamy, 2001, pp. 131-132) need to be addressed and promoted. From the view of critical language testing, test items should be designed to allow for multiple meanings and multiple interpretations for different test takers. This view contrasts with the psychometric approach to item writing which recognises test items such as multiple-choice questions as meaningful and objective scientific entities as they carry one correct answer, i.e., one "truth", one "interpretation" (Shohamy, 2001, p. 24). However, as reported by Shin (2013), critical language theorists have not yet proposed an alternative protocol for developing critical language test items and they currently focus on socio-political aspect of language testing, namely ethics in language testing and test misuse. The critical approach may not yet replace the other testing practices but function as "an additional approach to making decisions and judgements about individual language ability" (Lynch, 2001, p. 361). The principle of ensuring tests is being used for what they have been designed to be used for is very important in powerful, large scale tests, since misuse is very likely to adversely affect populations for whom the test was not designed but who are required to take it. This happens with many of the large-scale tests.

Each approach has its own philosophical stance and its derived item formats have their own strengths. While authentic constructed response test items are pervasive, selected response

items from the psychometric traditions remain influential for their practicality and efficiency (Shin, 2012; Shohamy, 2001). The systematic approach is well connected with building validation arguments for test items and test design, focusing on how items generate inferences from the response evidence to the claims test developers make about the test construct. This systematic approach has professionalised the item writing process, and been employed by examination boards such as ETS and Cambridge Assessment English, which were established in 1947 and 1913 respectively. (ALTE, 2005; ETS, 2016). The systematic approach is also used by Vietnamese tertiary institutions to develop their large-scale standardised English tests. In these tests, a combination of authentic, performance-based tasks and objective, multiple-choice test items are made use of for assessing productive skills and receptive skills correspondingly.

This study aimed to focus its investigation on item writing competencies for multiple-choice test items as opposed to performance-based tasks, so as to ensure its scope would be manageable. The difference between multiple-choice test items and performance-based tasks is found not only in the surface level features but also in the skills targeted for testing as well as in the social and cultural contexts of use (Lim, 2009). For example, the nature of assessment of performance tasks may present the possibility of errors arising owing to the subjective judgement of human raters and topic selection, which are less likely to affect multiple-choice tasks. The typical procedures and approaches followed for effective multiple-choice item development are not totally applicable for development of performance-based tasks (Welch, 2006). This might be because of the variance in the underlying principles of development between these task formats, themes that pervade the research literature on multiple-choice test item and performance-based test item development are different. While central to multiple-choice test item writing are principles, rules, guidelines and taxonomies (Haladyna, 1997; Haladyna, 2004; Haladyna & Downing, 1989; Haladyna et al., 2002; Popham, 1997), the issues of interest and concern for performance tasks are more related to assessment of performance

(McNamara, 1996). Factors such as raters, rater variability, rating expertise, inter-rate reliability, development and validation of scoring rubrics and the effects of training and background characteristics on rating expertise are of greater interest in performance testing (e.g., Kuiken & Vedder, 2014; Lim, 2009; Yan, 2014; Yang, 2010; Zhang, 2011). In addition, multiple-choice testing is favourable in the Vietnamese context mainly for its practicality. When administered to large groups of students, multiple-choice testing is more efficient and economical in terms of test operation and human resources required. They can be automatically marked with the use of computer technology or can be objectively marked by trained clerical markers or by teachers. Nevertheless, research has suggested that it is challenging and time-consuming to construct multiple-choice test items. Masters, Hulsmeyer, Pike, Leichty, Miller, and Verst (2001) reported that it required approximately one hour to write a good multiple-choice question. The challenge becomes greater when test items are written to assess students' complex abilities (Crooks, 1988; Stiggins, Griswold, & Wikelund, 1989). Haladyna et al. (2002) urged that more guidelines be provided for writing items to better assess complex learning and higher order thinking skills.

Given the need for practicality and efficiency of administration of testing in Vietnam, this study has directed its focus towards expertise in multiple-choice test item development for large-scale English reading and listening skill tests. Multiple- choice test items encompass objectively marked or selected-response test items. These items are claimed to be the preferred format for virtually every standardised test because of their ease of administration, high-speed scoring and more reliable and content-valid interpretation of test scores (Haladyna & Downing, 1989). The multiple-choice questions may be in different formats, including 3-4 choice questions, dichotomous items (e.g., True/False; Yes/No), matching, information transfer (i.e., transferring materials from the text to a table, chart, form, or map), gap-filling, Cloze, and C-test. This study focuses on Vietnamese teacher competence in writing multiple-choice test items for large-scale English reading and listening skill tests.

3.2 Existing Models of Multiple-Choice Test Item Writing

Very few studies have attempted to investigate the expertise of multiple-choice item writing, for example the cognitive and non-cognitive processes that item writers undergo while creating objective test items, the knowledge structures involved and how these may contribute to the increased understanding of the professional practice of item writers and to the quality of written items. Among a few rare pieces of research focusing on this area are those conducted by Fulkerson et al. (2009), Fulkerson et al. (2010), Fulkerson et al. (2011) and Salisbury (2005, 2016). This section is devoted to discussing the two models of item writing that emerged from these studies and exploring how these models could contribute to the underpinnings of investigating test item writing competence for English teachers in Vietnam.

3.2.1 Cognitive, Knowledge-Based Models of Item Writing

Research studies led by Fulkerson (2009, 2010, 2011) focused on exploring what item writers think when they are writing multiple-choice test items for content science subjects. The test items written by the item writer participants in Fulkerson and colleagues' studies are similar to VSTEP® listening and reading test items in terms of item stem and distractor format; however, the input materials for these items are different in terms of content and genre. The script texts for multiple-choice items for content sciences in the studies led by Fulkerson were written descriptions of scenarios provided to test takers to observe different processes of sciences while the input texts/scripts for VSTEP® listening and reading tests are English conversations, discussions, narratives and lectures of various topics ranging from social issues to academic content subjects. The findings from these studies were generalised into a cognitive model of item writing. Fulkerson et al.'s (2009, 2010, 2011) approach is related to the functional analysis and the actuarial approaches discussed in the previous chapter (see Sections 2.3.1 and 2.3.2) in that they used a group of exemplary practitioners as their research participants. The difference is that the data was not directly elicited from the participants but through the authors' interpretations of

the participants' think-aloud statements. The research participants were a small group of science content specialists (3-9 people) who had a broad range of education, teaching and assessment writing experience. These specialists were required to complete a test item writing assignment given by the researchers and were requested to verbalise cognitive information they generated during the item writing process, known as think-aloud protocols. Their verbal statements were audio-recorded, analysed and coded into the categories of cognitive processes and knowledge structures that item writers made use of to write test items. In their first project in 2009, Fulkerson, Mittelholtz, and Nichols conducted a study of the cognitive process of three experienced item writers, and it was revealed that expert item writers typically engaged in a three-phase problem solving process (see *Figure 3.1*). In the initial representation/definition phase, item writers defined a problem in the form of a "mental model" (Fulkerson et al., 2009, p. 2). In the middle exploration/operation phase, item writers searched within their problem-solving space for content that functioned as a plausible solution to the item writing task. A sequence of micro processes was involved in this phase. In the final solution phase, item writers successfully accomplished writing items with desirable properties.

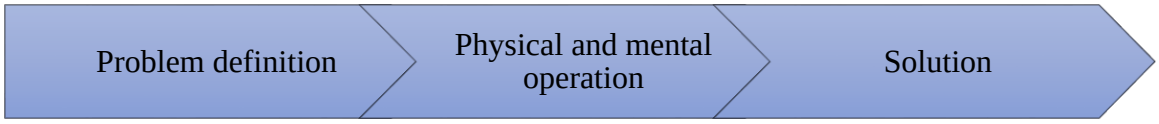


Figure 3.1 Model of item writing expertise based on cognitive processes

Adapted from "Expanding the model of item writing expertise: Cognitive processes and requisite knowledge structures" by D. Fulkerson, P. Nichols, & E. Snow, 2011, p.18. *Paper presented at the annual meeting of the American Educational Research Association*. Copyright 2011 by Pearson.

In 2010, Fulkerson, Nichols, and Mittelholtz extended this cognitive model of item writing expertise to develop an understanding of not only expert item writers but also novice ones. This second study involved the participation of both inexperienced and experienced item writers in an individual one-hour item writing assignment for storyboard scenes, during which

they were asked to verbalise the cognitive processes known as think-aloud protocols. The analyses of these verbal reports helped to explore how these item writers performed their tasks and to elicit data on the strategies, knowledge and logic they applied when they wrote test items. Table 3.1 presents 14 cognitive processes the participant item writers experienced as coded from their think-aloud statements. Figure 3.2 displays the framework employed in Fulkerson et al.'s (2010) study for recording the item writers' patterns of movement as they solved their item writing task. In this framework, 14 cognitive processes were arranged along a continuum with problem definition cognitive process placed in the middle. The cognitive processes that indicate forward movement through the problem space, namely Solution satisfy, Evaluation, Decomposition, Relaxation, Schema active and Operator are placed above the problem definition. Below the problem definition are those indicating backward motion which include Backtracking, Impasse, Constraining, Meta-clarify, Missing information, Nonconforming and Extraneous. The descriptions of these cognitive processes are provided in Table 3.1. The movements of the item writer participants were graphically recorded using scatter plots that showed their forward- and backward-movement through the problem space as they constructed the four storyboard scenes and multiple-choice items aligned to the scenes and the selected benchmarks. The results from this second study continued to confirm that, regardless of their experience, item writers went through a three-phase problem solving process as they wrote items for tests. The data analyses indicated that inexperienced item writers spent more time defining the task and made frequent backward movements in the problem space while experienced item writers moved forward at a higher frequency. The patterns of forward movement and backward movement were repeated across the three phases of the item writing process for experienced and inexperienced writers respectively. The less experienced writers used similar combinations of problem definition, operator, evaluation, and solution satisfaction statements but used these less frequently than experienced writers. This reflects the expertise development process proposed in

the Conscious Competence Theory and the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) (see Section 2.2) which state that inexperienced practitioners tend to adhere to rules and procedures while experts approach problems holistically by drawing on their breadth and depth of knowledge and experience (S. E. Dreyfus, 2004). Additionally, the findings implied that inexperienced item writers need coaching in ways to decompose the task in order to reduce the cognitive load and needed to be provided with a wealth of information relevant to their item writing assignment, templates, and other supports that help to structure their item writing process. Though this item writing framework offers important insights into the cognitive processes item writers go through, it does not provide a comprehensive framework of integrated abilities item writers demonstrate in each phase of the item writing process.

Table 3.1 *The Cognitive Item Writing Processes*

Cognitive process categories	**Description**
1. Problem definition	Describes an initial or subsequent problem representation that includes potentially useful knowledge elements. Includes seeking clarification about assignment or study procedure
2. Impasse	Refers to the state of mind in which the item writers feel that all options have been exhausted. Includes retreating toward earlier states or even to the beginning of the problem
3. Constraint recognition	Sets limits on the problem-solving phase
4. Relaxation and prioritisation	Expands the problem-solving options by relaxing constraints or selecting a preferred solution to an impasse or constraint.
5. Schema activation	Describes the application of mental structures drawing on experience
6. Operator	Reveals active searching for content and solutions. Includes mental and physical operators that advance the writer in the problem space
7. Evaluation	Describes evaluating an operator relative to some task requirement
8. Solution satisfaction	Describes meeting some desired goals
9. Extraneous	Represent statements that appear irrelevant to the assignments
10. Nonconforming	Represent statements relevant to the assignment that do not fit any coding categories
11. Meta-clarification	Examples of asking for clarification about the study procedure
12. Backtracking	Examples of retreating toward an earlier or intermediate state or even to the beginning of the problem.
13. Decomposition	Describe breaking-up problem-relevant chunks
14. Missing information	Represents recognizing and/or searching for clarification about the assignment

Note. Reprinted from *"What item writers think when writing items: Toward a theory of item writing expertise"* by D. Fulkerson, P. Nichols, & D. Mittelholtz, 2010, p. 8. Paper presented at the annual meeting of the American Educational Research Association. Copyright by Pearson.

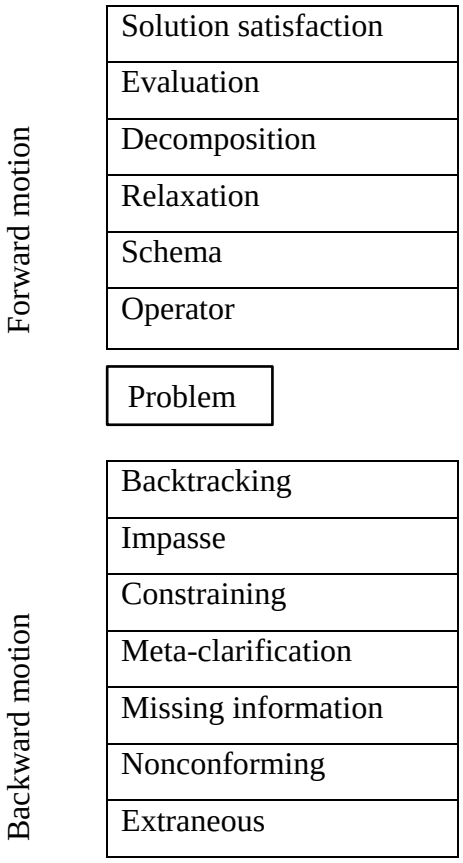


Figure 3.2 Cognitive process placement for item writers, indicating movement forward and backward in the problem space.

Adapted from "What item writers think when writing items: Toward a theory of item writing expertise" by D. Fulkerson, P. Nichols, & D. Mittelholtz, 2010, p.11. *Paper presented at the annual meeting of the American Educational Research Association*. Copyright 2011 by Pearson.

In their most recent expert study of item writers, Fulkerson et al. (2011) expanded the cognitive model of item writing expertise developed earlier by Fulkerson et al. (2009); Fulkerson et al. (2010). Fulkerson et al. (2011) incorporated requisite knowledge structures for item writing into their previously proposed model of the cognitive processes of item writers (Fulkerson et al., 2009, 2010) to propose a more comprehensive model of item-writing expertise. Using similar methods as in Fulkerson et al. (2009), the nine item writer participants were asked to verbalise cognitive information generated as they attempted to write multiple-choice items for the assigned tasks. Their verbal statements were coded into two broad categories: revised cognitive processes and knowledge structures. The results continued to confirm the findings from earlier work by Fulkerson et al. (2009, 2010) showing that item writers moved through a three-phased

problem-solving period with nine cognitive processes starting with problem definition, followed by operators, then solution satisfaction, with occasional application of constraint recognition, schema activation, and evaluative statements (see Table 3.2). The results from this study also coincided with those of the previous research showing that expert item writers were more efficient than novices as they moved forward at a greater pace in the problem space, and they more promptly recognised constraints and task priorities, and activated schema.

Table 3.2 *Cognitive Processes Employed by Item Writers*

Cognitive process categories	**Description**
1. Problem definition	Describes an initial or subsequent problem representation that includes potentially useful knowledge elements. Includes seeking clarification about assignment or study procedure
2. Impasse	Refers to the state of mind in which the item writers feel that all options have been exhausted. Includes retreating toward earlier states or even to the beginning of the problem
3. Constraint recognition	Sets limits on the problem-solving phase
4. Relaxation and prioritisation	Expands the problem-solving options by relaxing constraints or selecting a preferred solution to an impasse or constraint.
5. Schema activation	Describes the application of mental structures drawing on experience
6. Operator	Reveals active searching for content and solutions. Includes mental and physical operators that advance the writer in the problem space
7. Evaluation	Describes evaluating an operator relative to some task requirement
8. Solution satisfaction	Describes meeting some desired goals
9. Extraneous	Represents statements that appear irrelevant to the assignments

Note. Reprinted from "Expanding the model of item writing expertise: Cognitive processes and requisite knowledge structures" by D. Fulkerson, P. Nichols, & E. Snow, 2011, p.6. *Paper presented at the annual meeting of the American Educational Research Association*. Copyright 2011 by Pearson.

In addition, the findings revealed the use of five knowledge structures and the interaction between cognitive processes and knowledge structures among novice and expert item writers. The five knowledge structures used by item writers are general item writing knowledge, general

and pedagogical content knowledge, knowledge of content and items, knowledge of content and scoring, and context specific knowledge.

The analyses of transcripts of the participants' think-aloud sessions conducted in the study suggested that the item writers regardless of their experience used all the knowledge structures. This indicated that these knowledge structures were essential for item writers to produce quality items with the assessment content knowledge and general item-writing knowledge being the most frequently used and the context specific knowledge the least used. Novices expressed their application of general item writing knowledge slightly more frequently than experts. Expert writers demonstrated their exploitation of general and pedagogical content knowledge and knowledge of content and items at a higher proportion than novices. Though the study started to explore the novice - expert practice, it focuses on the novice - expert differences rather than the transition from novice to expert as proposed in the stage-based model of expertise development, the Dreyfus and Dreyfus model of skill acquisition. Thus, how a novice item writer raises their expertise to become an expert remains an area of active inquiry. The study also found the interactions between some cognitive processes and knowledge structures. The highest percentage of use was found with the combination of Operator/Knowledge of content and items, followed well behind by Problem definition/Not applicable by experts and Problem/General item writing knowledge by novices. The predominance of the two- category combination of Operator/Knowledge emphasises that regardless of item writers' expertise, the item writing process primarily starts with describing the aspects of an item writing task, followed by a quest for assessment materials to infer student ability.

The cognitive, knowledge-based model of item writing developed from the latest study led by Fulkerson (2011) shed light on the combined application of knowledge structures and cognitive processes by item writers and refined the three-phased item writing process they moved through with some marked differences between novice and expert item writers. However,

the model has its limitations with regard to its theoretical and methodological robustness and its practicality. While the model contributed an insight into the test item writing process that item writers underwent, it did not view test item writing from a competency-based perspective. This model conceptualises item writing expertise as a set of cognitive processes and knowledge bases rather than application skills which could be acquired, nurtured and developed. Thus, this model of item writing expertise has contributed very little to the training and practice of test item writers. The cognitive, knowledge-based item writing model consists of a list of cognitive processes and knowledge structures, which seems to be of limited usefulness for item writers to inform and support their practice. The model does not specify the application of knowledge structures and cognitive processes in observable performance tasks, delineating what item writers do, create, write and say at each stage of their item writing process. It also offers little understanding regarding the core competencies required of item writers and the performance indicators of these competencies. Moreover, the model is meant to be applicable to test item writing across disciplines rather than specifically to the EFL domain. It views competence from the generic perspective as discussed in Chapter Two (see Section 2.1.1); therefore, it fails to acknowledge the unique and distinctive bodies of knowledge and practice that define competence for such a specialised profession as English language test item writing. In terms of methodology, the model was generated based on the interpretation of a small group of item writers' verbalised thought processes, thus, might be limited in its generalisability.

3.2.2. Performance Task-Based Model of Item Writing

Of specific relevance to this research domain is the task-based model of item writing that emerged from Salisbury's (2005) study, which explored the sequential steps that expert and non-expert English for Speakers of Other languages (ESOL) listening test item writers undertake when writing the test items. Similar to Fulkerson and colleagues' methodological approach, Salisbury (2005) recruited five substantially experienced and five inexperienced item writers

who had knowledge of language assessment through their own learning and teaching experience. The participating item writers worked for the University of Cambridge Assessment English. These item writers were given the same test item writing task, resembling what is normally required in this field. The main components of the task were reading a given article, reviewing the test specifications, using the specifications to create a script by adapting the written text and writing eight multiple-choice questions. The study condition in Salisbury's (2005) research was rather different from the operational work of item writers. That is, in operational practice, item writers are not provided with the input materials, but they are required to search, select and adapt the input materials for their listening scripts. In the study, the participant item writers were asked to complete the task in 90 minutes and think aloud, that is, to say aloud into the audio recorder everything that crossed their minds when doing the task. The verbal protocols were transcribed and analysed using a modified grounded-theory approach to build the model of item writing directly out of the emerging data (Salisbury, 2005). The results demonstrated that there were three broad phases that all item writers went though, echoing the three-phased problem-solving process both inexperienced and experienced item writers engaged in as revealed in the studies conducted by Fulkerson and colleagues in 2009, 2010 and 2011.

While Fulkerson et al. (2009, 2010, 2011) categorised the process into cognitive problem-solving stages, Salisbury (2005) conceptualised these stages in terms of actions carried out by item writers. In the first phase, the *Exploratory* phase, item writers search and identify potentially exploitable input materials for tests from various written or oral sources such as magazines, newspaper, specialist journal articles and radio programs. In the second phase, the *Concerted* phase, they work intensively through five 'episodes' (Salisbury, 2005, p. 36) to produce test items for initial submission. These episodes are *Review test specifications, Read text, Devise context, Devise script, and Devise items*. These episodes represented key stages in the problem-solving path that the item writers took, and these item writers used diverse

approaches to their problem solving. Thus, the episodes may not necessarily happen in a linear sequence, as some episodes may be embedded, passed, and revisited or may not occur during the process of constructing test items. The final phases, the *Refining* phase, happens after either self-,peer- or editor-review. It is in this final phase that the item writers respond to a critical review by the writers themselves and/ or by others and improve the test items in an effort to make it comply with requirements.

Embedded in the three phases as described by Salisbury (2005) is a range of specific tasks at a macro-level namely *'Explore input materials, Review test specifications, Read text; Devise context; Devise script, Devise and Refine items'* (p. 61) employed by novice and expert item writers, which are the focus of exploration in the next section of the chapter. Unlike Fulkerson et al. (2009, 2010, 2011) who modelled test item writing in terms of cognitive processes, Salisbury (2005) adopted the behaviourist approach to conceptualising test item writing as a set of key performance tasks executed by item writers (see Section 2.1.2 for the behaviourist approach to competence). This performance task-based model is the only one found in the current literature that views test item writing from a competency-based perspective and contributes to demystifying the competencies required of item writers and expressing these competencies in observable behaviours. However, this theoretical approach to conceptualising test item writing competence in Salisbury's (2005) study was not considered an optimal one. As argued by Gonczi and Hager (2010) and Fish and de Cossart (2006), the behaviourist approach to competence that describes professional practice as a list of tasks could possibly hinder the development of professional expertise and result in the failure to fully capture the complexity of professional practice.

While Salisbury's (2005) model of item writing competence specified major item writing competencies exhibited by the IELTS listening test item writers, it did not enhance our understanding of the developmental pathways of these competencies so that item writers could

self-assess their own competence at different stages of their career, identify their professional learning needs and goals and scaffold their competence growth (Gillis & McPherson, 1995; Griffin et al., 2007). Based on Salisbury's (2005) model, item writers may be informed of the major job functions they are expected to perform, but it does not offer them an indication of how well they perform. There was a potential for conceptualising multiple-choice test item writing competence in light of the developmental approach reviewed in the previous chapter (see Section 2.1.3.3) so that the multiple-choice test item writing competencies identified from Salisbury (2005) could be further operationalised to the extent that they could be measurable. Specifically, the developmental theoretical and methodological approach as reviewed in the previous chapter (see Sections 2.1.3.3 and 2.3.3) would guide how these competencies could be further operationalised into corresponding indicative skills/performance indicators and their varying levels of proficiency. Thus, item writers' level of competence in multiple-choice test item writing at a particular stage could be identified through self-assessment, which in return could possibly inform the item writers of their next steps in developing their expertise in multiple-choice test item writing. Additionally, in terms of methodology, Salisbury's (2005) model was generated from the think-aloud protocols of ten IELTS listening test item writers, and it was not empirically validated. Thus, it was unknown if it could contribute to research and practice as a reliable and valid measure of the competence. There remained a great need for constructing a model of competence for multiple-choice test item writing which is underpinned by current holistic views of competence and developmental learning theories and guided by the developmental multi-methods approach as reviewed in the previous chapter (see Sections 2.1.3.3, 2.2.3 and 2.3.4).

The valuable contribution that the existing models of item writing made to the current study was that they provided a broad understanding of the item writing process and the major item writing job functions performed by item writers of different levels of expertise. Built up on

these findings, this present study aspired to extend our understanding of the constituents of the item writing competence and its development. Salisbury's (2005) analyses of the systematic approach to test item development taken by the IELTS listening test item writer participants revealed five distinctive tasks the item writers conducted in a three-phase process, resembling the three-phase problem solving process as revealed from Fulkerson et al. (2009, 2010, 2011). These findings were consolidated by Anderson and Morgan (2008), who elaborated description of the item writer's job as follows:

> Item writing requires attention to detail, creativity, intellectual rigor, depth of content knowledge, and good understanding of students' development in a learning area. Ideally, item writers should demonstrate the following traits: (1) They should show initiative and willingness to conduct an extensive search for interesting stimulus materials, and they should be able to develop high-quality stimulus materials; (2) They should have a capacity to accept feedback on their work and to comment on the work of other item writers with the same degree of critical detachment; (3) They should demonstrate a desire for excellence and a willingness to attend to details when developing and refining items (p. 52).

The functional and actuarial analyses used in the research studies led by Salisbury (2005) and Fulkerson and colleagues (2009, 2010, 2011) revealed that multiple-choice test item writing competence can be represented as a three-phase process made up of five core competencies representing five major job functions performed by item writers (see *Figure 3.3*).

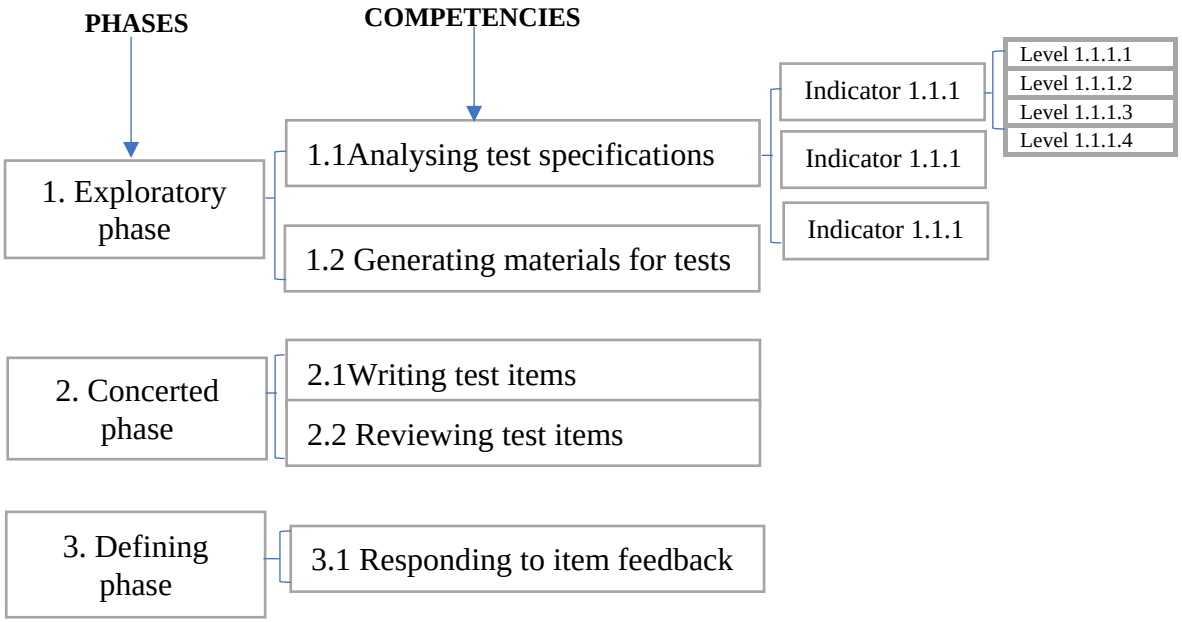


Figure 3.3 The conceptual framework for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests.
Adapted from "The edge of expertise: Toward an understanding of listening test item writing as professional practice", by Salisbury, 2005, London, UK: King's College. Copyright 2005 by King's College.

In summary, while the reviewed studies provided a general understanding of the process item writers experience and the major competencies they need to possess in order to carry out their test item writing tasks, they did not show how these competencies could be demonstrated and developed. There was a potential for describing test item writing competencies in light of the developmental theoretical and methodological approach to include descriptions of performance behaviours that demonstrate evidence of the identified test item writing competencies and the varying levels of the performance quality. Adopting the developmental approach to competence grounded on the theoretical and psychometric underpinnings contributed by Vygotsky (1978), Glaser (1981) and Rasch (1960, 1980) (see Section 2.1.3.3), the study aimed to describe Vietnamese EFL teacher competence in multiple-choice test item writing as a continuum of observable tasks manifesting the underlying ability in terms increasing levels of proficiency. The

description of increasing competence was guided by Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) (see Section 2.2.3). The Rasch modelling procedure was conducted to identify the mathematical connection between the competence of a teacher item writer and the demand of a task. Analyses of clusters of tasks of similar levels of difficulty were carried out so that the emerging developmental progression of Vietnamese EFL multiple-choice test item writing competence could be observed. This progression would present different levels of expertise in multiple-choice test item wring along the continuum an item writer could attain at different stages, illustrating a learning trajectory in which novices progress successively into experts, opening up opportunities and providing inspiration and motivation for teachers to monitor their competence growth (Abbott-Chapman et al., 1993; Griffin, 2007, 2014; Louden, 1994; Louden & Wildy, 1998; Snoek et al., 2019; Symeonidis, 2019). This approach enables defining multiple-choice test item writing competence in observable performance tasks as well as the description of how one develops their ability to complete specific tasks. This helps to unveil the nature of item writing competencies, the typical order of their development, complexity and levels of sophistication, which was not available the reviewed research. Most importantly, within this perspective, multiple-choice test item writing competence is not static, restricting itself to a list of attributes and tasks, but the competence is integrative and evolutionary and can be increased through learning, experience and mentoring/coaching from more competent colleagues (Griffin et al., 2018, p. 102). In other words, the adopted developmental approach to defining and operationalising item writing competence contributes a more dynamic understanding of item writing tasks, creating a potential for mapping professional learning trajectories and development paths for Vietnamese EFL item writers.

The following sections report on and critique research findings from the scholarly literature in order to help unravel the behavioural indicators and possible developmental trajectories of the test item writing competencies.

3.3 Test Item Writing Competencies

A few separate studies have been identified on how expert and non-expert item writers demonstrate evidence of their competence in analysing test specifications, generating input materials for tests, and approaching and judging item difficulty. However, these studies focused on distinguishing between novice and expert practice, limiting the competence into dichotomous levels and were thus not able to capture the progression of these competencies as item writers develop from novice to expert (Griffin et al., 2007). Moreover, these studies did not investigate the novice-expert transition in light of a developmental skill acquisition and expertise model, such as the Dreyfus and Dreyfus model. In terms of methodological approach, these studies mainly rely on either the functional analysis or actuarial approach. Similar studies on the other competencies are extremely scarce; instead, these competencies are mainly discussed in handbooks and manuals for test item development. It is argued that while current research and guidelines contribute a general understanding of how these item writing competencies are demonstrated, little is known about how item writers develop their competence in performing these competencies/tasks. There is a lack of empirical research that offers a comprehensive insight into how test item writing competencies are acquired, nurtured and developed. As illustrated in Chapter Two, there needs to be a developmental progression for item writing competencies that item writers can use to guide their item writing practice and monitor and scaffold their expertise development, and professional interventions can be targeted and implemented. The following section is organised around the five core test item writing competencies identified from Salisbury's (2005) research and is a synthesis from the scholarly literatures on what was known and what was unknown about how these competencies are evidenced, acquired and developed.

3.3.1 Analysing Test Specifications

It is common practice for a group of test developers to prepare test specifications through a series of steps. As defined by Davidson and Lynch (2002), a well written language test specification is a detailed and flawless document that allows for consistent interpretations by item writers who are often identified as teachers with similar training experiences and from similar teaching contexts. Moreover, it should allow for the production of homogeneous spec-fit test items. It is expected that by referring to test specifications, item writers can write items with relative ease. In other words, item writers are commissioned to write items as detailed in the test specifications upon the assumption that well-established test specifications yield high-quality test items. However, Kim et al. (2010) maintain that how item writers utilise test specifications presents a "potential mediating factor" (p. 161) in the process of item development. It was found in their case study that what item writers thought about test specifications impacted their ways of utilising them. Some item writers drew on their language teaching experience, actual item writing practices, and expertise as content matter experts to internalise the content of the test specifications and used it as a guide to develop their own internal guidelines for item writing. Some item writers described test specifications as "abstract", "mysterious", "less transparent", "less useful", and dense with descriptions of constructs and task types which they considered unnecessary; thus, they avoided reading the test specifications in detail or did not rely much on test specifications once they started writing items. As observed by Kim et al. (2010), besides their teaching experience, item writers perceived that additional resources were helpful, such as training materials, weekly briefing with the team, face-to-face communication, and intensive editing and revising sessions with item writing teams. These resources and activities provided them with a clearer understanding of the items they were to write and with well-tuned and specific "organic guidelines" (p.174) for developing each test item and a variety of vividly illustrated example items with desirable features and varied levels of difficulty.

It can be inferred that even though test specifications are tailored for item writers, as suggested by Alderson et al. (1995), there is a predictable gap between item writers' expectations and test developers' intentions. Kim et al. (2010) concluded that it could be beneficial to engage item writers in the test specifications design stage. This resonated with Lynch and Davidson (1994) and Salisbury (2005), who suggested that experienced item writers can be a valuable source of feedback for improving test specifications. Regarding expert versus non-expert interpretation of test specifications, Salisbury's (2005) research also revealed that the vast majority of item writers review the test specifications first and expert and novice writers devote a similar proportion of operators in this stage. While novices normally read and comment on such aspects as text type, text length, item type, sub-skills and marking, experts go beyond reading the specifications but internalise the significant facets of the specifications and draw on their experience with the targeted candidates to consider how to operationalise the test specifications. Expert item writers internalise test specifications more effectively, following them with ease and without needing frequent checks on them (Salisbury, 2005).

To sum up, contemporary research focuses on distinguishing novice from expert performance in test specifications analysis but fails to capture the pathway that progresses from novice and expert performance. Thus, little has been known as to how to raise one's expertise in analysing test specifications. There is potential for applying what was reviewed in Chapter Two on the developmental approaches to conceptualising and modelling competence in order to better understand test item writers' transition pathway from novice to expert in light of a stage-based model of learning and skill acquisition.

3.3.2 Generating Materials for Tests.

Item writers' next task is to generate appropriate materials that meet the test specifications and potentially yield desirable items. Multiple-choice test item writers for large scale English reading skill tests usually have their own bank of texts accumulated from their

daily reading for use in future tests, and test items should not be developed until the input materials are approved by the editing committee (Green & Hawkey, 2011). A few studies have been interested in exploring how item writers search, select and adapt input materials, in audio, oral or written forms, for their tests. Green and Hawkey's (2011) study focuses particularly on how item writers select and adapt texts for large-scale academic IELTS reading tests that are similar to VSTEP® reading skill tests in terms of text types and test item format required. The participant item writers were asked to write and audio record their reflections on the text generating process to provide data on their perception, selection and adaptation of source texts for IELTS reading tests. After that, how they refined the texts and items was observed at a focus group session. The results from qualitative and quantitative analyses have shown certain levels of competence and application required in selecting and adapting input materials for tests so as to align them with the test specifications in terms of task types, textual features, lexical density, readability and cultural appropriateness.

Regarding text search and selection, non-expert item writers in the study by Green and Hawkey (2011) usually used IELTS websites or published IELTS test materials and utilised IELTS practice tests as a source of information for text selection. They tended to web-search for topics of general or personal interest. Expert item writers did not use IELTS resources, but they browsed the web for materials on topics recommended in the item writers' guidelines and test specifications. They regularly browsed the bookshops and/or fruitful websites, searching magazines, newspapers or even peer-reviewed journals for materials. Some writers claimed that they "sort of live searching for text the whole time" (Green & Hawkey, 2011, p. 293) even when they are not commissioned to write tests and save potential texts on file. As experienced writers, they kept a record of "high-yield" sources and they had confidence and commitment to continuing to pursue texts from these sources. Green and Hawkey (2011) also found that novices focused on searching for texts of suitable length and may not be certain about the levels of text

difficulty and complexity. In contrast, experts were concerned with issues such as bias, cultural sensitivity, and testability of materials. Non-experts were generally satisfied with the text they found and they would ask a friend for their opinion on the text. They did not consider perceived characteristics of IELTS texts mentioned in the training materials or test specifications. Expert IELTS item writers also based selection on the perceived characteristics of IELTS texts, but they did not seem to be content with the guidance on the levels of text difficulty and complexity for each proficiency level so they tended to develop their own proficiency level criteria, which they called "text comprehensibility factors" (Green & Hawkey, 2011, p. 290).

According to these researchers, expert item writers arrived at their decision to select a text after a series of activities. They selected a longer than required text, read closely, highlighted points to test, made notes of each paragraph to identify redundancy and repetition, associated tasks with sections, and cut out unused chunks of texts. After writing test items, they removed non-highlighted sections to end up with the suitable text. Some others would work on the items and text in tandem, editing the text to make sure texts-fit-tasks and tasks-fit-specifications. This process often results in creating a new text catering for different test purposes. The item writers paid special attention to text density and lexical difficulty when considering a source text for their test items. They often saw items "leap out from the text" (Green & Hawkey, 2011, p. 293) as they evaluated the text. The expert item writers also considered avoiding texts related to potentially divisive or offensive issues and religious issues. Novice item writers tended to retain many of the original base text features while experts employed a repertoire of strategies to adapt original texts to meet the required textual features in terms of language use and style and their rhetorical functions. These strategies include:

- Cut redundancies and repetitions;

- Cut journalistic elements and make word substitutions to give the text a more formal style (e.g., substitute abbreviations and contractions with full forms; exchange colloquial synonyms);
- Delete or substitute technical vocabulary;
- Cut or gloss culturally specific knowledge
- Eliminate unsuitable and offensive content
- Use text analysis tools (e.g., English profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to measure text difficulty and complexity (Green & Hawkey, 2011)

When non-experts make changes to the text, they tend to largely change the text structure, and register, making the text more formal or academic by adding more academic words, using more passive forms, and hedging language. They also may replace technical words with more frequently used alternatives and add a gloss. Apart from making textual changes to the selected text, item writers frequently adapt or adjust the text to make the text fits the items they write or to make the items fit into the text. Experienced item writers often employed some research-grounded factors in the adaptation of text. The strategies used by the item writers from the studies by Salisbury (2005, p. 221) and Green and Hawkey (2011, p. 279) are presented in Table 3.3.

.

Table 3.3 *Strategies for Adapting Texts/Scripts Applied by IELTS Reading and Listening Test Item Writers*

Strategies used by IELTS listening test item writers	Strategies used by IELTS reading test item writers
• Text-item barter: swapping lexical items between text and item so that the higher-level synonym is in the text • Adding distractions to the script to reduce guessability of keys • Trimming text – reducing text in order to constrain the key • Placing keys at intervals throughout the text, often by adding text, to allow sufficient candidate processing time between items • Modifying keys so as to prevent listeners from identifying a cue from the item stem (to avoid dictation effects) • Shifting word forms • Moving distracting elements to subvert a predictable pattern and source of prediction	• Cutting texts to an appropriate length • Removing unsuitable content to make the text inoffensive • Cutting or amending the text to avoid candidates being able to get the correct answer simply by word matching, rather than by understanding the text • Glossing or removing cultural references if inappropriate • Deleting confusing or redundant references • Glossing, amending or removing parts of the text which require experience or detailed understanding of a specific topic • Moving ideas around • Paraphrasing and using antonyms

Text adaptation was one of the five macro-level tasks that the IELT listening test item writers in Salisbury's (2005) study executed. In order to explore the mental processes involved in each test item writing macro-level tasks, Salisbury (2005) identified 96 cognitive operators grouped into six categories, namely VERBS, OBJECTS, JUSTIFIERS, CONTROL, CONTEXT and SPECIFICATIONS and applied operator codes in the item writers' think-aloud scripts according to what she interpreted the participants did, to what object and for what reason. These operators were defined as mental operations or problem-solving strategies that the item writers applied as they progressed through the item writing task. Some example operators are

READ/RE-READ, REVIEW, REHEARSE and PARAPHRASE (VERB operators); TEXT CUT, TEXT TYPE MAIN TEXT POINT/GIST, and ITEM WORDING (OBJECT operators); COHERENCE, COHESION, and EXPLOITABILITY (JUSTIFIER operators); and LANGUAGE LEVEL, ITEM TYPES, ITEM ORDER and AUTHENTICITY (SPECIFICATIONS operators) (see Salisbury, 2005, p. 344-353 for the detailed list of operators with their sub-sets and description). In this stage of text adaptation, Salisbury's (2005) analyses revealed that experts used far more operators than non-experts, 24 operators and 3 operators respectively. Non-experts surveyed, read and understood the text quickly to establish the text content. Experts spent lots of time working on texts (e.g., specifying text context, analysing text, commenting on text type, text access technicalities, text exploitability, text-item match, mapping out the text's impact on items to be developed, text length, proposing text cuts, distilling, and evaluating text content) and establishing the most appropriate item types for the base text content. They were inclined to multi-task – working between text and tasks simultaneously on a split screen and made extensive use of paraphrase, synonymy and antonymy.

Current research has contributed to enhancing our understanding of expert versus non-expert practice in input text selection and adaptation (e.g., Green & Hawkey, 2011; Salisbury, 2005; Sydorenko, 2011). The research results could be demonstrated on a continuum in light of an underlying developmental learning theory to map out the developmental trajectories of these competencies, contributing a more useful and educative tool for item writing practitioners.

3.3.3 Writing Test Items

Test item construction is a complex and time-consuming process and the most crucial task for item writers is to confirm if each test item fits into the provided test specifications (Griffin & Care, 2014). Research has provided a repertoire of tips and techniques for item writers to embark on in producing test items. Griffin and Care (2014) recommend that it is helpful for item writers to refer to test specifications and other proper guidelines to ensure that

students' responses to test items provide inferable evidence of their proficiency in the constructs being tested. They also suggest that new item writers may use and/or adapt a pool of existing test items to their test purposes. In the same vein, Alderson et al. (1995) contend that textbooks and other learning materials are very useful sources of ideas for writing items. Besides, item writers should be able to "imagine" their item types and be the best imaginative and creative thinker they can be. It seems to be practical for test writers to consult past test papers or their own resources of previous contents for replication or construction of test items. However, as emphasised by Alderson et al. (1995), item writers should also adhere to the test specifications to ensure inclusion of the varieties of test tasks required. Haladyna and Rodriguez (2013) describe many important steps in item development, which aims to gather procedural validity evidence for a test. From the perspective of an item writer, item construction involves four main tasks: deciding on how to approach item writing, drafting item stems and distractors, providing answer keys, and estimating item difficulty.

3.3.3.1 Approaching item writing.

It appears that there is a scarcity of research on approaches undertaken by item writers to writing items for different language skills test. To date, Salisbury (2005) has been the first and only study that investigated practitioners' approaches to writing items for English listening tests. These English listening test items are similar to VSTEP® listening tests items in that they are multiple-choice test items with four answer options and the listening scripts for IELTS and VSTEP® listening tests both include conversations, discussions and lectures. Two distinct approaches to writing listening test items applied by expert and non-expert writers emerged from Salisbury's (2005) research data: the item-first approach and the script/text-first approach. The expert participants took the item-first approach, that is, the Devise item episode preceded the Devise script episode, or they started from the questions and then built a script to fit with these. In the meanwhile, the non-experts were inclined to adopt the script-first approach, that is, they

wrote the scripts before developing items. Non- experts seemed to adhere to their communicative approach to teaching and materials development, and therefore they prioritised the script. They tended to retain the original text but end up modifying it to accommodate the items in the later stage. On the contrary, recognising that items need to be carefully arranged to fit the script, the experts flexibly privileged items over scripts; thus, they lessened the need for rewriting or editing scripts later. Though a greater emphasis was placed on devising scripts and items by all participants, the experts embedded and distributed their attention more equally to all of the operation phases than non-experts to avoid being obsessively concerned with superficial matters such as text length or item type. While the non-experts in Salisbury's (2005) study usually demonstrated a distinct imbalance between devising items and devising scripts, the expert item writers were found to undertake two tasks concurrently on split screens. For instance, they adapted the text as they finalised the items or they reviewed the text as they devised the script. This is similar to the key findings from Fulkerson et al. (2009, 2011) in which the experts frequently combined problem definition, exploration, evaluation, and solution activities as they moved through the problem space toward the accomplishment of their item writing task. This reflects the ability to retrieve the information quickly and to multi-task intuitively as featured in the expert performance by Dreyfus and Dreyfus (1960). In the meanwhile, the non-experts' patterns of thought processes were erratic and they frequently moved backward in the problem space.

3.3.3.2 Writing item stems and distractors.

Each multiple-choice test item has three components: (1) a stem, which can be a complete question, or a partial statement completed by the choices or an instructional statement; (2) one or more correct options; and (3) one or more plausible but incorrect options called distractors. As reported by Rodriguez (2016) (p. 266), objective test item writing guidelines have been present in approximately 100 textbooks and book chapters since 1951 and reviewed

multiple times by such authors as Haladyna and Downing (1989), Haladyna et al. (2002), and Moreno et al. (2015). To date, the taxonomy of 22 guidelines on selected-response test item writing synthesised and reviewed based on a meta-analysis of previous studies on item writing guidelines by Haladyna and Rodriguez (2013) has been recognised as the most recent and comprehensive set of guidelines. The taxonomy consists of 22 self-explanatory statements which largely cover issues related to writing the distractors and issues about experience-based good item writing skills. The item writing guidelines included in Haladyna and Rodriguez (2013) are as in Table 3.4.

Table 3.4 *Item Writing Guidelines*

Content concerns
1. Base each item on one type of content and cognitive demand.
2. Use new material to elicit higher-level thinking.
3. Keep the content of items independent of one another (unless they are scored as an item set).
4. Test important content. Avoid overly specific and overly general content.
5. Avoid opinions unless qualified.
6. Avoid trick items.
Format concerns
7. Format each item vertically instead of horizontally.
Style concerns
8. Edit and proof items.
9. Keep linguistic complexity appropriate to the group being tested.
10. Minimize the amount of reading in each item. Avoid window dressing.
Writing the stem
11. State the central idea clearly and concisely in the stem and not in the options.
12. Word the stem positively, avoid negative phrasing.
Writing the options
13. Use only options that are plausible and discriminating. Three options are usually sufficient.
14. Make sure that only one of these options is the right answer.
15. Vary the location of the right answer according to the number of options.
16. Place options in logical or numerical order.
17. Keep options independent; options should not be overlapping.
18. Avoid using the options none-of-the-above, all-of-the-above and I do not know.
19. Word the options positively; avoid negative words, such as NOT.
20. Avoid giving clues to the right answer:
a. Keep the length of options about equal.
b. Avoid specific determiners, including always, never, completely and absolutely.
c. Avoid clang associations, options identical to or resembling words in the stem.
d. Avoid pairs or triplets of options that clue the test taker to the correct choice.
e. Avoid blatantly absurd, ridiculous options.
f. Keep options homogeneous in content and grammatical structure.
21. Make all distractors plausible. Use typical errors of test takers to write distractors.
22. Avoid the use of humour.

Note. Reprinted from *Developing and validating test items (p.91)* by T.M Haladyna and M.C. Rodriguez, 2013, New York: Routledge Taylor and Francis Group. Copyright by Routledge Taylor and Francis Group.

Rodriguez (2016) scrutinised the scarcity of empirical evidence in the literature to support most of the guidelines; however, as argued by Haladyna and Rodriguez (2013) research on item writing guidelines could be “pointless” (p. 89) as they are “widely held values” (p. 89) based on a consensus among testing experts. Nevertheless, there is some evidence to support the relevance and usefulness of these guidelines. First, these guidelines have been used in studies on

item-guideline violations (E.g., Ellsworth, Dunnell, & Duell, 1990) and effects of poorly written items on reliabilities and item difficulties (e.g.,Downing, 2002; Downing, 2005). One notable study is the one by Jozefowicz et al. (2002) in medical education which evaluated items written by trained writers versus items written by untrained writers. Five hundred and fifty-five science test questions, 92 of which had been written by item writers who were trained to write items for the United States Medical Licensing Examination, were collected. The quality of each of these questions was rated independently by three expert biomedical test developers without knowing the purpose of the study nor the identities of the item writers using a rating scheme with the maximum score of 5. The statistically significant difference between item quality for trained and untrained item writers has been found, with a mean rating of 4.24 out of 5 for the items written by trained writers and 2.39 out of 5 for those written by untrained writers. This could imply that training and adherence to item-writing guidelines contribute to the quality of test items. Rodriguez (2016) echoes that specific item writing guidelines employed in training, supervising and monitoring item writers should be closely followed and documented in the test specifications. He also mentioned that some evidence has been found for guidelines 12 “Word the stem positively, avoid negative phrasing”, 13 “Use only options that are plausible and discriminating; three options are usually sufficient”, 18 “Avoid using the options none-of-the-above, all-of-the-above and I don’t know” and 20 “Avoid giving clues to the right answer: a. Keep the length of options about equal” (Rodriguez, 2016, p. 268). Third, some evidence from test accessibility studies, for example, Beddow (2012), has confirmed the significance of using such guidelines. Beddow (2012, p. 98) reports that when strong item writing guidelines like those reported in Haladyna and Rodriguez (2013) are employed, the item functioning and examinees’ experience are enhanced. This is also supported by Elliott et al. (2010) and Kettler et al. (2011) (cited in Rodriguez, 2016, p. 269).

Apart from the guidelines for item writing above, researchers also presented other tactics to especially support new item writers who may experience "writer's block" (Haladyna &Rodriguez, 2013, p.23 & p. 144) when they start drafting test items. One solution is the use of retired items. Novice item writers can construct a new item by editing and improving the retired items or revising these items. These activities give them some hands-on practice and experience with item writing. Another remedy is the use of a variety of item shells (cloning items) which might be found in a good item writing guideline. Item shells are hollow items containing a set of syntactic structures derived from successfully performing items and the expertise of subject matter experts (SMEs) who develop these shells (Haladyna & Rodriguez, 2013). Item writers may use generic item shells made available to them in the test specifications or training materials. If these item shells are not provided, item writers may generate these item shells by themselves from successfully performing item stems in sample tests or past test papers. Examples of generic item shells are shown in Table 3.5. These item shells are organised around a classification system proposed by Haladyna and Rodriguez (2013).

Table 3.5 *Generic Item Shells*

Defining Concepts
Which is the best definition of …?
Which is the meaning of …?
Which is synonymous with …?
Which is the correct definition for …?
Which is like …?
Which is characteristic of …?
Which is an example of …?
Which distinguishes …?
Defining Principles
Which is the best definition of …?
Which is the principle of …?
Which is the reason for …?
Which is the cause of …?
Which is the relationship between … and …?
Which is an example of the principle of …?
Predicting Principles
What would happen if …?
If …, what happens?
What is the consequence of …?
What is the cause of …?

Evaluating Facts and Concepts
Which is the most or least important, significant, effective …?
Which is better, worse, higher, lower, farther, nearer, heavier, lighter, darker, lighter …?
Which is most like, least like …?
What is the difference between … and …?
What is a similarity between … and …?
Evaluating Principles
Which of the following principles best applies to …?
Evaluating Procedures
Which of the following procedures best applies to the problem of …?
Applying Concepts, Principles, Procedures
What is the best way to …?
How should one …?

Note. Reprinted from *Developing and validating test items (p.145)* by T.M Haladyna and M.C. Rodriguez, 2013, New York: Routledge Taylor and Francis Group. Copyright 2013 by Routledge Taylor and Francis Group.

Item writers may also follow a series of steps to generate item stems as presented by Haladyna and Rodriguez (2013, p. 146). These steps are:

1. Identify successfully performing items
2. Identify cognitive behaviours represented by the item
3. Identify the content the item aims to test
4. Remove a critical feature from the stem

The stem shells will then have much more detail and can be adapted to be more specific about the content to be tested. When item writers' capability builds up, they may be able to modify or even create new stems and options. When it comes to writing distractors for objectively scored items, it is advisable to adhere to the 22 guidelines on writing options. One important issue of concern is that all distractors should be plausible. A distractor is considered useless if it is not chosen by any test taker and test developers recommend that such a distractor should be replaced, removed or revised (Haladyna & Rodriguez, 2013). Another is that item writers should draw on their subject matter knowledge and their experience with the target examinees to write distractors that reflect common errors or misconceptions, offering diagnostic information of examinees' knowledge, skills and abilities. A measure of distractor analysis can inform item writers of how effectively a set of distractors perform. At the initial phases of item

development, consensus expert judgement is critical in evaluating the distractor discrimination via multiple reviews of test items, and key verifications. Item writers may try out the items as an examinee or ask a colleague to try them out so as to infer some information about the distractors' performance. They may pilot the items as an embedded section in a test. Research has shown that the best to identify distractor discrimination is through statistical analysis of responses to item distractors (Alderson et al., 1995; Haladyna & Downing, 1993) because distractor analysis could identify poor distractors unrecognised by SMEs. However, this is beyond the scope of an item writer's job as they are rarely involved in the statistical validation of test items, whether formal or informal. Therefore, discussions of tools to assess the statistical properties of responses to item distractors is not the focus in this study.

Item writers may employ other technology-driven item writing methods such as generic testlets and testlet conversions to produce items that fit testing purposes and the test task types required. The generic testlets are usually scenario-based and used to design items to test complex, multistep high order thinking skills. Testlet conversion is a recycling of the generic testlet; item writers retain the original scenario and develop a series of items that probe into sequential and logical steps toward the correct solution. It is also speculated that the future development of a new science in item writing called Automated Item Generation (AIG) will allow items to be created automatically by computers and more efficiently and productively than traditional item generation (Haladyna & Rodriguez, 2013). AIG promises to improve item development in terms of cost and efficiency and test security; however, it is beyond the scope of this study as the focus is on the engagement of item writers throughout the item development process. In brief, it can be deduced that there are evidence-based characteristics, quality and methods for item writing and that the increasing role of technology in the science on objective test item development is predictable. Haladyna and Rodriguez (2013) noted that item

development must be accompanied by a sound argument and recorded validity evidence to support it regardless of whatever direction it takes.

It could be concluded that there are ample checklists, tips, and do's and don'ts on writing test item stems and distractors. No research has examined the ordered difficulty of these strategies and whether their use varies as a function of expertise so that a developmental map of how to raise expertise in writing item stems and distractors could be created.

3.3.3.3 Providing answer keys.

As viewed from the psychometric paradigm discussed earlier (see Section 3.1), after writing item stems and distractors, item writers provide correct answers for multiple-choice items, whose correctness must be verified and determined by 100% consensus. Item writers can provide a reference from an authoritative source which may be underlined or from highlighted chunks of texts from reading passages or listening scripts, textbooks, curricula, or journal articles so as to ensure the correctness of the answer keys. Item writers may also provide justification for each option in the test items. Scoring of these items does not require expertise and they are often scored by clerical staff or by machines. The experienced item writers in Salisbury's (2005) research usually gave a clear, single-word answer key and spaced the answer key at appropriate intervals in the texts/scripts so as to allow the examinees sufficient time to process the language between items. In contrast, the non-experts frequently provided overly long and complex answer keys, which require high-inference criteria for marking; in other words, the items with long and complex answer keys may not be reliably clerically marked, but they may be required to be marked by subject matter experts (Salisbury, 2005).

3.3.3.4 Deciding on item difficulty.

Estimating the content and cognitive demand for each item or estimating difficulty levels of test items as specified in the test specifications is considered to be one crucial task item writers perform in their process of writing items (Alderson et al., 1995). While items are being

developed and before they are piloted, it is essential that item writers can establish fairly accurate judgements of item difficulty and discrimination. In some cases when it is costly to field test or pilot items, item writers' prediction of item difficulty is vital. Another way to estimate item difficulty is to adhere to item level descriptors provided in the test specifications or a framework of reference in use while constructing items. Alderson et al. (1995) found that experienced item writers predict item difficulty slightly better than inexperienced ones. Other researchers concluded that training can help to improve item difficulty estimates (Fortus, Corriat, & Fund, 1998; Hambleton & Jirka, 2006; MacGregor, Kenyon, Christenson, & Louguit, 2008). Sydorenko (2011) demonstrates that experienced item writers consider research-grounded factors influencing item difficulty when writing items so when they write items intended for a particular difficulty level, they are better aware of what affects item difficulty. The researcher reported four sources of factors affecting item difficulty from various relevant studies. They were:

- the items themselves (e.g., main idea items are found generally easier than detailed ones; easier items contain greater similarity of words used in the input and the items);
- the input materials (e.g., complexity of syntax and lexis are harder; popular topics reduce difficulty; faster speech rates can reduce comprehension and redundancy may increase comprehension);
- the linguistic knowledge or skills targeted (e.g., different components of communicative language ability constructs might be challenging for some learners and easy for others), and/or
- the response format (e.g., negative stems, some item formats are subject to possibility/higher chance of guessing the correct answer, True/False items are

easier than multiple-choice items; Multiple-choice items are easier than cloze, dialogue reconstruction and matching).

These findings demonstrate that there is a demand for empirical research to confirm whether the levels of competence in item writing are determined by the amount and/or type of experience accumulated.

3.3.4 Reviewing and Revising Test Items.

When new test items have been developed, they should be subjected to multiple purposeful reviews for content, editorial style, fairness and psychometric properties (Schmeiser & Welch, 2006). Item reviewing is also referred to as qualitative item validation or item moderation. The item-reviewing processes are to ensure the test items conform to one of the fundamental principles of a test development process which is validity, along with reliability and fairness. Item validation is one critical step in an item writing process as items are important sources of validity evidence. As defined by the Standards for Educational and Psychological Testing (2003), validity is "the degree to which all the accumulated evidence supports the intended interpretation of test scores for the proposed use" (AERA, APA, & NCME, 2014, p. 11). The process of validation entails collecting relevant evidence from various sources, namely test content, response processes, internal structure of test items and other variables (ibid., 2014). It is crucial that each item content is scrutinised by SMEs and testing specialists.

Approaches to item review may be of varying rigour and formality; however, the three main reviewing processes, namely editorial review, item content fit review and keys and distractor review, are mandatory. It appears that there has been abundant research and guidelines focusing on ensuring and improving item quality through vigorous item reviews and analyses. Nevertheless, no research to date has contributed a solid understanding on how item writers demonstrate their competence in self-reviewing their items before submission, in improving and polishing their test items after either self-, peer- or editor/SMEs- reviews, and in responding to

reviewers' feedback. Salisbury (2005) also admitted that her research collected very little data on this aspect due to the limitation of the research frame (p. 276). Furthermore, it is indicated in Salisbury (2005) that the task of test item writing involves a distinct mix of work modes (p. 231). Item writers carry out much of the initial test item production independently, for instance, reviewing test specifications, searching for input materials for tests and writing test items. Meanwhile, a significant portion of the item refining process is conducted collaboratively with peers or a panel of editors or SMEs. That might explain the scant interest in investigating item writers' capabilities in reviewing and revising test items as part of their individualised item writing process. Furthermore, it is also highlighted that the role and the extent to which the item writers are involved in the process after they submit their test items also vary (ALTE, 2005; Haladyna & Rodriguez, 2013; Lane et al., 2016)); for some item writers, their job finishes when they submit the items and the rest will be taken care of by the editing team and/or the test developers. For others, they may take part in editing meetings and engaging in robust discussions on their items, and revise their items based on received feedback. For yet others, they may even get informed of how their test items perform in the operational or pilot tests. This indicates that the extent to which item writers are involved in the team effort to review test items varies depending on the policy and practice of the examination board. It was revealed from Salisbury's (2005) research that though item writing appeared to be both individual and collective task, the item writer participants carried most of the test item production "in isolation" (p. 231). It was found that collective collaboration was only present in the last stage of item writing process (i.e., reviewing test items) in form of an iterative process of receiving and responding to written item feedback or taking part in item review meetings. Salisbury's (2005) study did not focus on the collective aspects of the item writing task but on the individual task performance. This was analogous to Fulkerson and colleagues' studies (Fulkerson et al., 2009, 2010, 2011) which investigated the item writers' individualist approach to item writing.

Similarly, this study adopted the individualistic conceptualisation of competence (see Section 2.1.3.3) and focused on the item writing tasks performed independently by individual test item writers to inform their self-development. It was decided that the collaborative aspect of test item review was included in the initial draft competency framework and discussed at the data collection events with the research participants. This was to affirm that the findings from the literature with regard to collaborative test item review were addressed and examined in the context of the study to ensure the proposed framework best reflected the targeted item writers' practice.

The following section discusses findings from the literature on the topics of item review and item analysis. The synthesis may provide some insights into how item writers acquire, demonstrate and develop their competence in reviewing and revising test items.

3.3.4.1 Editorial review.

Once written, firstly, test items must undergo a series of editorial reviews. The goals of the reviews are to improve item clarity without affecting the content and to make sure that the items are free from grammatical, spelling, punctuation and capitalisation errors. Another facet of item review is to eliminate bias related to cultural and ethnic background, gender, and stereotyping which may surface from the test materials. Lane, Raymond, Haladyna, and Downing (2016) recommend that the reviews should focus on:

- linguistic complexity, fairness, correctness of key or correct model answer, appropriate intended cognitive demand, and match to test specifications
- alignment between item content/difficulty and the test specifications
- sensitivity review to reduce or eliminate irrelevant variables related to cultural, ethnic, religious background and gender

Depending on the context and the size of the testing program, the editorial review is done by fellow item writers, item writers themselves, external SMEs and/or the intended examinees

but it is best accomplished by professionally trained editors using a standard editorial style manual (Lane et al., 2016). Downing (2006) strongly advocates a collaborative or team approach to item editing: the item writers and the editors work collaboratively as partners to undertake a comprehensive review of items. This resonates with Malau-Aduli and Zimitat's (2012) suggestion that peer review be included in the process of multiple-choice item writing. The findings from Malau-Aduli and Zimitat (2012) revealed that peer review processes produced an increased proportion of items with appropriate item difficulty and discrimination power and distractor functioning and a reduction in mechanical flaws. Whenever feasible, the items should be reviewed by the targeted examinees themselves to eliminate any possible ambiguous wording or item flaws. Proofing and editing items are important in the item writing process even for the most carefully crafted items by experienced item writers, as these activities partly contribute to encouraging item writers to "review, reflect and rewrite" (Baranowski, 2006, p. 352).

There seems to be such a fine line between editorial reviews and qualitative validation (also known as content reviews or moderation of test items) that Baranowski (2006) classifies it as a source of qualitative validity evidence. This study views editorial review and content review of test items as integral, important tasks in a qualitative validation process of test items. According to Wise and Blake (2016), the Standards for Educational and Psychological Testing requires that "the test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria" (p. 88).

3.3.4.2 Item content review

With regard to item content reviews, Osterlind (1998) characterised item content validation as a process of uncovering sources of errors through precise and systematic validation methods to provide item writers with valuable information to examine and improve the quality of their test items. The writer endorses two main approaches to validating items: through SMEs'

judgement using established rating scales and through appropriate statistical analyses of item responses piloted with a sample representative of the target population.

3.3.4.2.1 Qualitative item review.

Approaches to judgement involve seeking for agreed opinions about the congruency between test items and their defined set of descriptors in terms of content and levels of difficulty in the test specifications from a qualified panel of 4-15 experts (Osterlind, 1998). Typically, SMEs are asked to express their perceived judgements of how congruent each item is with its intended objective or descriptor and to rank their opinions on a Likert rating scale. The judging process can be open or anonymous and with or without the engagement of the item writers. The judges' ratings are then tallied and interpreted based on a certain criterion level (e.g., the minimum acceptable value for congruence is a correlation of .78 for the index score ranges from +1 to -1).

Alderson et al. (1995) discussed a similar process and termed it item moderation. According to Alderson et al. (1995), items in their operational test form may undergo a number of informal self- and peer-reviews and/or independent and anonymous reviews. Informal reviews should have taken place at an early stage of the test construction process. Individual item writers may look closely at their items by themselves to detect any obvious errors. Item writing team members and coordinators may cross check each other's items by actually doing the test, responding to the items as examinees would, reflecting upon what abilities are required to complete the items and then comparing what they think the items are testing with what is claimed in the test specifications. The best crafted items are often subject to a number of informal reviews within the groups until the final item drafting stage. Once the items are edited, they are assembled into a draft test paper to be evaluated by an editing committee including experienced item writers, SMEs, teachers who teach the target group of examinees and other testing experts. These committee members do not merely read through the test but actually

replicate the experience of a test taker attempting each item in the test in advance of the committee meeting. The committee members discuss to verify the congruency between the test items and the item blueprint, examine the conformity of the items to the recognised guidelines, check their levels of difficulty and identify possible technical problems related to test layout, wording of items and instructions, answer keys and the match between texts/scripts and questions. It is suggested that an effective editing committee is chaired by a firm chief examiner and that the feedback from the committee is recorded, followed upon, and incorporated into a revised test.

3.3.4.2.2 Statistical item analysis.

The other approach is to conduct statistical analyses of the test item responses. Item analysis has been extensively discussed in the testing and assessment literature (e.g., Alderson et al., 1995; Downing & Haladyna, 1997; 2006; Haladyna & Rodriguez, 2013; Anderson & Morgan, 2008; Osterlind, 1989; Raymond, 2015). It should be noted that this process elicits the data about the item performance by analysing test takers' responses to test items so it can only be conducted when the item responses or test results are available. Livingston (2006) recommends that it can be done either before or after the items are assembled into operational tests depending on testing authorities' policies and resources available or even right before the tests are actually scored. Research has shown that there is a discrepancy in judgements or estimates of item difficulty even among experienced item writers and testers, as reported in Alderson et al. (1995, p.73) and Syndrorenko (2011), and many items, though thoughtfully designed and carefully edited, turn out to be problematic, which is only identified during item pretesting. Pretesting may vary in scope. Item writers may try out their test items on their team members of whom at least one is a native speaker of English. They may administer the trial of the test items, either as a separate or embedded set of items in a live test, on a representative sample of the target examinee population. In contexts where pre-testing is constrained for a

certain reason, item analysis happens after the test has been conducted and functions as a mechanism for quality control. In this case, statistical analyses of items are done after the test is administered but before it is scored, and the data provided help item writers to make necessary amendment in the scoring key before the tests are scored by computers or clerical staff. Alternatively, it is done after the test results have been reported; at this stage, the data help item writers to select items for future use and to inform their practice. The main purpose of the analysis is to yield information about the psychometric properties of the items, rather than the test takers. The most useful psychometric properties of the test items that statistical item analyses can provide item writers with are item difficulty, item discrimination and distractor performance. The information helps item writers to identify items that should be revised or retired and to select healthily-performing items for operational tests (Livingston, 2006 in Downing & Haladyna, 2006). However, as discussed earlier (see Section 3.3.3.2), item writers rarely engage in the statistical validation of test items, mostly for test security reasons and because additional and different skill sets may be required; the ability to analyse and interpret the statistical properties of test items may not be one of the core competencies item writers must possess. Therefore, this was not elaborated in this study.

It can be seen that the literature is replete with informative discussions on how items can be qualitatively reviewed through experts' opinions and analysed through empirical procedures so as to improve item quality. It is also found that access to data on item and distractor performance and functioning is helpful for item writers to improve their items and inform future item writing practice (Alderson et al., 1995; Livingston, 2006 in Downing & Haladyna, 2006). Nevertheless, no research to date has provided a solid understanding on how item writers tend to respond to item feedback and exploit the feedback from reviewers to review and revise their test items.

In summary, there is a scarcity of empirical research providing an understanding of the ordered difficulty of the strategies used by item writers of varying levels of expertise when applying their skills and knowledge to produce multiple-choice test items. It is important that these competencies be operationalised on a developmental continuum so that item writers' acquisition of and development in these competencies can be better understood for recruitment, training and accountability purposes. The aim of this study was to enhance understanding of how EFL test item writing competencies are demonstrated and developed. The first primary research question of this study was: "What is the typical developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests?" Following the procedure for constructing developmental frameworks suggested by Griffin et al. (2018), the findings from this chapter's review of the test item writing competency literature contributed an insight into core competencies in test item writing and their behavioural indicators. This produced a hypothesised conceptual framework for later construction of the developmental progression of the competencies and for its subsequent empirical testing. *Figure 3.4* depicts the information in the competency framework for multiple-choice test item writers developed thus far. This framework is limited to the core competencies and performance indicators only, and there is no hypothetical developmental progression yet.

Phases	*Core competencies*	*Performance indicators*
The exploratory phase	Analysing test specifications	Reviewing test specifications
		Utilise test specifications
	Generating input test materials	Searching for input test materials
		Evaluating the materials for text-fit-specifications
		Making textual changes to the selected materials for tests
		Adapting texts for texts-fit-items and items-fit-texts
The concerted phase	Writing test items	Determine the approach to item writing tasks
		Write item stems
		Write item distractors
		Provide answer keys
		Decide on item difficulty
	Reviewing test items	Check item content fit
		Verify answer keys and distractors
The refining phase	Revising test items	Respond to item feedback

Figure 3.4 The core competencies and corresponding performance indicators required for multiple-choice test item writers.

3.4 Chapter Summary

This chapter has reviewed the scholarly literature and research on test item writing competencies. A large part of the literature on test item writing competencies seems to be scattered across several case studies and an indefinite number of manuals, guidelines and handbooks for test development. The contemporary literature has contributed a solid understanding of expertise in item writing which can be modelled as a three-phase process delineating five major competencies in terms of core, observable tasks performed by item writers

regardless of their levels of expertise (Salisbury, 2005) (see Section 3.2). There is an absence of empirical research into the developmental pathways of these competencies and the factors contributing to the development of these competencies to inform item writers' professional learning and development of expertise. Current research has provided limited understandings into how the five test item writing competencies are acquired, nurtured and developed. Building on the existing literature on test item writing, there is a possibility to describe the development progression of these competencies in light of the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) following the developmental, multi-methods approach as discussed in Chapter Two (see Section 2.3.3), which is the focus of Chapter Five. The next chapter, Chapter Four, will explore factors contributing to teachers' professional competence growth to illuminate how expertise in test item writing can be developed and what educational and professional interventions could be targeted to extend teachers' expertise in test item writing.

CHAPTER 4: FACTORS AFFECTING TEACHER EXPERTISE DEVELOPMENT

There has been an absence of empirical research into the developmental trajectories of the test item writing competencies required of teachers and, in particular, a scarcity of investigation into the factors promoting this competence growth. However, the substantive research that has examined factors affecting the development of professional expertise, teachers' professional competence growth in the workplace, teachers' teaching and testing and assessment expertise informed the conceptual design of this study. Given that professional learning is the main mechanism that drives teachers' professional competence growth, models and measures of professional learning in the field of teacher education and development are reviewed. Existing measures of professional learning are contextualised and adapted to be used to define and measure English test item writers' engagement in professional learning. The chapter also explores educational and professional background characteristics that have been identified to potentially influence teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests within the context of EFL large-scale testing and assessment. At the end of the chapter, the propositions underpinning the investigation are described.

4. 1 Professional Learning: A Mechanism for Teacher Competence Growth

Research in the field of development of competence has suggested a series of factors that account for higher levels of competence that distinguish experts from novices as in Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) (see Chapter 2, Section 2.2.2) in both well-structured domains of professions (i.e., the domains where practice leading to improved performance is recognisable and observable, such as sports, performing arts and sciences) and in ill-structured domains (i.e., the domain where it is challenging to define what constitutes improved performance, such as the teaching profession) (Dunn & Shriner, 1999).

Several theoretical models have been put forward to explain professional competence growth from different perspectives. Of interest to this current study are those models situated in the social theories of learning since they propose to explain the development of competence in terms of acquirable or malleable factors and limit the potential impacts of innate capacity, individual giftedness and affective characteristics. These models of learning also emphasise the inherent developmental and context-dependent nature in the process of becoming experts, which aligns with the conceptual framework of competence as a continuum adopted for this study. Among the most influential models are deliberate practice by Ericsson et al. (1993) and Ericsson and Charness (1994), self-regulated learning by Pintrich (2000), experiential learning by Kolb (1984) and reflective learning (Schön, 1987). In order to argue that professional learning empowers teachers' professional competence growth, first, social theories of learning are briefly discussed. Secondly, the connections between these theories are examined. Third, an account is given regarding how the learning models have been applied in research studies to construct an understanding of the extent to which deliberate, self-regulated learning from and through work experiences contributes to development of higher levels of competence in different work contexts.

Deliberate practice activities that are performed and maintained with prolonged personal effort are supposed to be directly related to improving the level of competence (Bransford & Schwartz, 2009). Deliberate practice as described by Ericsson et al. (1993) comprises the goal-oriented practice activities that are conducted in three highly-structured phases. In the first phase, professionals are already engaged in regular practice and demonstrate rapid progress that is recognised by others. In the second phase, professionals practice intensively with clearly set goals guided by more expert mentors who demand enormous effort and perfect performance. In the third phase, professionals are themselves committed to excel, to overcome difficulties and perfect their performance through working with master experts to critically analyse their and

others' performance. Ultimately, professionals move beyond the existing body of knowledge and contribute to their profession. Deliberate practice is repetitive in nature and requires effort, concentration and attention. Engaging in deliberate practice activities has been shown to provide individuals with optimal learning opportunities based on which their performance is improved (Ericsson et al., 1993).

Research has also shown there is an association between self-regulated learning and effective competence building and performance improvement (Zimmerman, 1998, 2000, 2006). Self-regulated learning is a process of planning, monitoring and evaluation as summarised by Pintrich (2000). The planning phase involves goal setting, knowledge activation and strategy planning. During the performance phase, actions and results are tracked to make adjustment to goals, plans, and strategies if required. The final phase focuses on the evaluation of and reflection on the outcome and the implementation procedure with reference to the goals and the action plans set at the beginning. These phases are not necessarily sequentially ordered as they may be dynamically entwined and complementary to each other.

A vast literature from expertise development, adult learning, and professional learning in higher education (Boud & Garrick, 1999; Eraut, 2000) suggests that professionals learn from their experiences. The process of experiential learning starts with the identification of a problem to be solved. Insightful analyses of the problem through reflection or self-inquiries prompt professionals to plan to conduct a task in a different way and reflect on the results. The whole process is deliberately mediated by reflection, resulting in new knowledge and understanding, and leading to more highly effective future performance. Kolb (1984) summarised the process of experiential learning as "a cycle of concrete experience, reflective evaluation, abstract conceptualisation and active experimentation" (p.38). In the same vein, Schön (1983, 1987) emphasised the critical role of individuals as reflective practitioners in their development of competence. He claimed that practitioners are characterised by their ability to engage in a

deliberate, retrospective analysis of action which he termed ‘reflection-on-action’ and the ability to interpret problems and experiment solutions as the problems occur, ‘reflection-in-action’ (as cited in Dunn & Shriner, 1999, p. 632).

The theories of deliberate practice, self-regulated learning, experiential learning and reflective learning resemble each other in that they emphasise the active self-directed nature of learning and focus on learning from professional experience. It is defined that professional learning is a variation in knowledge, skills, attitudes and values that enhances performance and brings about higher levels of competence (Boshuizen et al., 2004).

A number of studies have attempted to apply the concept of deliberate practice and self-regulated learning to work settings and to analyse the relationship between deliberate practice and competence development. For example, Dunn and Shriner (1999) conducted a study on the influence of deliberate practice on teachers’ professional competence and Sonnentag and Kleine (2000) did a similar study on insurance agents. Both studies aimed at identifying the deliberate practice activities that may impact the higher levels of competence growth. In these studies, lists of relevant activities as depicted in the theory of deliberate practice and self-regulated learning were recorded, gathered, and reflected upon by professionals in diary logs and structured interviews. The activities which were regarded as deliberate practice activities had to satisfy the criteria for such activities as proposed by Ericsson et al. (1993). For example, the activities are highly relevant to improving teaching effectiveness or results in performance improvement; the activities are regularly performed; the activities are highly optional and indirectly related to financial rewards (Dunn & Shriner, 1999; Sonnentag & Kleine, 2000). Both studies revealed that professionals’ engagement in self-regulated, deliberate practice is related to their performance. Dunn and Shriner (1999) investigated teacher activities that potentially impact the development of higher competence levels for teachers who have achieved satisfactory competence levels. Eight volunteer teachers who on average had 18 years of experience teaching different grade

levels participated in the study. The teachers were asked to keep a diary of their teaching related activities within school and outside of school in 15-minute increments on a daily basis for 14 days and attended qualitative interviews to elaborate on the activities and the outcomes of activities as perceived by the teachers themselves. Every day, they kept a record, in a given coding booklet, of their activities using a coding scheme developed through a pilot study. They posted the log data to the researchers every other day. Several common and routine aspects of teaching featuring deliberate practice as presented by Ericsson et al. (1993) were teaching activities that involved frequent planning, evaluating the effectiveness of teaching strategies, revising and restructuring the teaching strategies for improvement. It was found that teachers' engagement in more planning, more thorough self- evaluation and revision frequently for years led to and maintained their teaching expertise. In teaching, mental planning, reading educational literature, attending workshops, discussions with colleagues, developing instructional strategies and activities, and informal evaluation of students through observation were considered deliberate and self-regulated activities that teachers carried out. However, it was indicated from the study that teachers seemed to engage in the activities in order to improve instruction and student learning outcomes.

In the study by Sonnentag and Kleine (2000), 100 sales agents from German insurance companies who had a mean experience of 11.7 years (Standard Deviation, SD = 9.3) and a mean age of 40.4 (SD = 9.4) participated in a structured interview which was designed to assess their work activities, the goals associated with each activity and the current and cumulative time spent on ten activities. These activities were identified as deliberate practice activities based on a set of five criteria. The activities were preparation, mental simulation, exploring new strategies, asking for feedback, consulting colleagues, consulting domain experts, formal meetings, informal meetings and private conversations. In order to minimize socially desirable responses, the interviewees were informed that the study focused on insurance agents' work activities and

procedures. In order to evaluate the validity of the interview data, the participants were requested to fill in the diary sheets the duration of time they spent on 10 specified activities during the following week. Time spent on the deliberate practice activities during the week was then summed up and correlated with the interview data. It was concluded that the validity of the interview data was satisfactory. Professional competence of these sales agents was measured based on the performance data obtained from their supervisors. The supervisors were asked to rate the agents' performance with respect to "meeting the sales goals" and "acquisition of new clients" on a 9-point Likert scale (1 = extremely low performance; 9 = outstanding performance). The regression analyses when years of experience and number of cases were controlled showed that current time spent on and the amount of deliberate practice accounted for significant percentages of explained variance in the agents' job performance. However, cumulative time spent on deliberate practice was not detected to be significantly correlated with performance. This finding was inconsistent with previous studies, which used retrospective reports of deliberate practice (e.g., Chanrness et al., 1996; Ericsson et al., 1993; Stakes et., 1996). Sonnentag and Kleine (2000) explained this inconsistency as possibly being attributable to the methodological shortcomings of using retrospective reports as valid measures of cumulative practice and the differences in competence development between professional domains. This suggests further empirical research be conducted to validate the findings.

In the same vein, van de Wiel, Szegedi, and Weggeman (2009) investigated the development of competence from a self-regulated or deliberate practice perspective in the professional domain of consultancy. The authors followed the procedure used by Sonnentag and Kleine (2000) to explore whether recognised top strategy and organisational consultants applied more self-regulated learning activities than those of similar experience. It was found that top professionals invested more time in reflection, talking to colleagues and updating activities such as attending a conference, reading professional literature and taking part in innovation. They also

perceived these activities as being most important for competence building and professional development. The activities of reflection and consulting knowledgeable others are strongly recommended in theories of deliberate practice (Ericsson et al., 1993) and self-regulated learning (Pintrich, 2000).

In summary, professional learning conceptualised as engagement in learning activities that are deliberate, self-regulated, and stimulating reflection, self-evaluation and collaboration is integral to developing competence. Compelling evidence from research (e.g., Dunn & Shriner, 1999; Ericsson et al., 1993, 1996; Sonnentag, 1998; van de Wiel et al., 2004;) has shown that these deliberate professional learning attempts/endeavours exert positive impacts on competence development. As synthesised by Boshuizen et al. (2004), professional learning as a self-directed and self-managed attempt to enhance competence from and through reflection and experience in working contexts bridges the gaps between novices and experts. In other words, professional learning is regarded as transitions between novices and experts.

4.1.1 Defining Professional Learning for EFL Teacher Item Writers

Given the role of professional learning in empowering teacher competence growth and professional development, research on teacher education and development has been interested in conceptualising professional learning. Attempts have been made to define critical features of professional learning activities that contribute to teacher competence development and to propose different models of professional learning based on which measures of teacher engagement in professional learning can be developed.

It is worth mentioning the shift in the use of the word ‘professional development’ to ‘professional learning’ in the field of teacher education. This is more than a change in terminology. Durksen, Klassen, and Daniels (2017) made a useful distinction between professional development and professional learning. According to the authors, professional development refers to activities arranged for teachers while professional learning emphasises

teachers' self-motivation and responsibility for their learning. The researchers quoted a comprehensive definition of teacher professional learning from Avalos (2011) and Richter, Kunter, Klusmann, Lüdtke, and Baumert (2011), who view teacher professional learning as:

> a complex process, which requires cognitive and emotional involvement of teachers, individually and collectively, the capacity and willingness to examine where each one stands in terms of convictions and beliefs, and the perusal and enactment of appropriate alternatives for improvement and change… (Avalos, 2011, p. 10).

With a focus on learning, the concept of professional learning moves the emphasis from passive development to active learning. Professional learning is acknowledged to be sociocultural, focusing both on individual learning and collaborative learning. Professional learning is conceptualised as continuing, active, social, and related to practice. This is aligned with the nature of deliberate, self-regulated, reflective and experiential learning that has been found to positively influence the development of professional expertise discussed earlier in this chapter.

Teacher professional learning encompasses a vast range of activities and interactions that keep them up to date and enhance their skills and knowledge, and thus improve their practice and student learning outcomes and experience. These activities can range from formal, specified, and mandated learning opportunities known as a main element of the 'training model' (Little, 1993, p. 129) or the 'traditional view' (Lieberman, 1995, p. 591), on professional development to everyday, informal learning opportunities (Desimone, 2009; Richter et al., 2011). As Richter et al. (2011) put it, teacher professional learning is a complex process which includes "the uptake of formal and informal learning opportunities that deepen and extend teachers' professional competence, including knowledge, beliefs, motivation, and self-regulatory skills" (p. 116). The traditional model assumes that teachers update their skills and knowledge by attending discrete activities such as formal term long courses, half- or full-day workshops, and

local and national conferences, in which knowledge that can be applied at the workplace is transferred to them by experts (Feiman-Nemser, 2001).

In contrast, informal professional learning is embedded in teachers' everyday work lives (Desimone, 2009) and includes individual activities such as reading books, journals, teaching materials or subject-matter literature, and observation and collaborative activities such as informal conversations and discussions with colleagues, mentoring and coaching activities and study groups. These activities can be engaged in independently at teachers' own pace and will. Among these activities, Putnam and Borko (2000) emphasise that one of the most powerful teacher learning occurs in activity at individual levels, such as engagement in self-examination of and reflection on their lessons and teaching practice, engagement in online educative activities (Ingvarson, Meiers, & Beavis, 2005), and their independent exploration (Guskey, 2000).

Thus, teacher learning can happen in myriad informal and formal contexts at both the individual and collaborative levels, as suggested by Cobb (1994), who argued that "learning should be viewed as both a process of active individual construction and a process of enculturation into the…practices of wider society" (p. 13). Variations have been found in teacher participation in these professional learning opportunities. The vast majority of teachers (98.2%) engaged in certain types of professional learning according to the largest nation-wide study in America which examined roughly 54,200 teachers' uptake of different professional learning opportunities by Choy, Chen, and Bugarin (2006). Most teachers used formal workshops, conferences or training courses, and fewer teachers participated in informal collaborative activities (Choy et al., 2006). Richter et al. (2010) examined teacher engagement in formal and informal professional learning activities across their five-stage career lifespan. In order to assess teachers' formal learning opportunities, Richter et al. (2010) asked teachers to recall all in-service training activities they had participated in over the last 24 months. The total number of in-service training activities indicated the teacher participation in formal professional learning.

Informal learning opportunities were measured by two different indicators: teachers' use of professional literature and their collaborative practice. Teacher collaboration was measured by asking teachers to rate a set of items which examined how well teachers collaborated with their colleagues in selecting teaching strategies, developing materials, and planning lessons on a Likert scale ranging from (1) never to (4) very often. Results based on the data from 1939 German secondary teachers in 198 schools revealed that mid-career teachers most frequently used formal in-service training activities whereas teacher participation in informal professional learning varied across their teaching career. Both younger teachers and older teachers invested their time in professional learning, and the older teachers preferred self-directed learning, namely reading the professional literature while their younger peers tended to collaborate more (Richter et al., 2010).

Considerable research has suggested that pre-service formal training and in-service professional development focusing on assessment or language assessment improved teachers' levels of testing and assessment literacy respectively (Griffin et al., 2013; Koh, 2011). Assessment literacy refers to apprehending the principles of sound assessment (Popham, 2004; Siggins, 2002) by integrating theories and practices to enhance teaching and learning (DeLuca & Bellara, 2013). Thus, an assessment-literate teacher is "one who creates, chooses, administers, interprets, responds to, records and reports assessment information in such a way that those decisions can be shown to be adequate and appropriate" (Brown, 2008, p. 286). Fulcher (2012) defines teachers' language testing and assessment literacy as:

> The knowledge, skills, and abilities required to design, develop, maintain or evaluate, large-scale standardised and or classroom-based tests, familiarity with test processes, and awareness of principles, and concepts within wider historical, social, political and philosophical frameworks in order to understand why practices have arisen as they

have, and to evaluate the role and impact of testing on society, institutions, and individuals. (p. 125).

This definition of teachers' language testing and assessment literacy shows that test item writing competence is imbedded in the broad definition of language testing and assessment literacy (i.e., develop large-scale standardised or classroom-based tests). The studies in teachers' language testing and assessment literacy served as the relevant references for this current study which aimed to explore teacher competence in test item writing. A number of studies have found that formal training in assessment significantly contributes to the development of teachers' assessment literacy.Levy-Vered and Alhija (2015) examined the inter-relationships between assessment literacy and assessment training, self-efficacy, and perception of assessment based on the data collected from 187 inductee and 140 in-service Israeli teachers through three questionnaires and a test in assessment literacy. The results from the Structural Equation Modelling (SEM) analyses demonstrated that there was a direct and positive correlation between training in and perception of assessment and assessment literacy, and this correlation accounted for 68% of the variance in assessment literacy with a strong indication of correlation magnitude referred to as an effect size (Cohen, 1992), $f^2 = 2.13$. Training in assessment accounted for 23% of the variance in conceptions which represents a medium-large effect size ($f^2 = .30$). Training in assessment and assessment literacy explained 33% of the variance in teachers' assessment self-efficacy with a large effect size, $f^2 = .49$.

In this study by Levy-Vered and Alhija (2015), training is defined as acquiring knowledge about assessment tools, grading, interpretation, reporting and ethics; practising assessment; and modelling of quality assessment practice by teacher educators (Stiggins, 1999; Volante, 2006). The results showed that formal pre-service training or in-service professional development workshops in assessment competence had a positive influence on teachers' assessment activities and resulted in higher levels of assessment literacy. These results resonate

those from other studies on the correlation between assessment literacy and the acquisition of knowledge, practising, and modelling assessment throughout the training period (e.g., Chen, 2005; Sato, Wei, & Darling-Hammond, 2008; Zhang & Burry-Stock, 2003). The data reveals that if training is of higher quality, novice teachers' assessment literacy is higher, and the perceptions and understanding are better. Assessment literacy is predicted by training and conceptions of assessment rather than self-efficacy. The findings underline the role that the training institutions play in nurturing assessment literacy among novice teachers. Novice teachers who received high-quality, formal training in assessment appeared to be more assessment-literate than those who did not. These studies did not investigate the relationship between training and the assessment literacy levels of experienced teachers.

Similarly, McGee and Colby (2014) investigated the assessment literacy of teachers before and after finishing an assessment course in their teacher training program using the Assessment Literacy Inventory developed by Mertler and Campbell (2005), which assesses teachers' seven core assessment capabilities, namely choosing assessment methods, sound design of assessment, scoring, using results, grading, communicating results and assessment ethics. The data collected from pre-test and post-test procedures showed significant positive change in all seven assessment capabilities. The teacher participants achieved their increased mean scores for sound design of assessment and communication of results on the post-test and sustained a high level of competence in ethical assessment, scoring and choosing assessment methods. It was concluded that an assessment-focused course contributes to the assessment literacy of teachers; thus, formal training in assessment is a vital part of teacher education programs (McGee & Colby, 2014, p. 530).

Pertinent to the field of multiple-choice test item writing competence for English tests is research conducted in language testing and assessment literacy. Numerous studies have shown that professional training in language assessment led to improvement in in-service language

teachers' language assessment literacy (Mahapatra, 2016; Montee, Back, Donovan, & Thompson, 2013; Nier, Donovan, & Malone, 2013; Riestenberg, Di Silvio, Donovan, & Malone, 2010; Walters, 2010; Yastibas, 2018). Montee (2013) studied the development of language teachers' assessment literacy after they participated in a seven-week training course in language assessment which blended online components and live face-to-face workshops. The course was designed as an introductory course to language assessment and focused on supporting the participants develop language assessment materials. The analysis of the assessment materials produced by 29 course participants showed that the face-to-face workshop element of the course was effective in increasing the participants' assessment knowledge and applying these to assessment tasks. In the same vein, Mahapatra (2015) conducted a case study of ten foreign language teachers with the aim to track the development of language assessment literacy after they participated in an online professional development course on language assessment using Web 2.0 tools such as Google Document, Google Drive and Google Hangouts. The data were collected from the teachers' stories, reports, weekly hangouts and their assessment tasks. The data analyses revealed improvement in the quality of designed assessment tasks, the teachers' understanding of the concept of language competence and assessment principles and their interpretation of assessment results. This finding is consistent with the result from the study on promoting teachers' language assessment literacy through an online course by Nier, Donovan and Escobar (2012). Multiple-choice test item writing competence is closely related to one of the seven testing and assessment capabilities, which is sound design of assessment, indicated in the Assessment Literacy Inventory developed by Mertler and Campbell (2005).The findings from the studies on the correlation between pre-service and in-service training teachers' levels of testing and assessment literacy may shed light on the hypothesis that there could be a similar correlation between training and teachers' levels of competence in multiple-choice test item writing.

In summary, there are two models of professional learning: the traditional model, or the training model (Liberman, 1995; Little, 1993) and the newer, more complex and dynamic model advocated by various researchers in the current professional development literature (e.g.,Borko, 2004; Greeno, 1997; Greeno, Collins, & Resnick, 1996; Putnam & Borko, 2000). The 'training' model conceptualises professional learning as a range of formal, structured, and topic-specific training seminars offered to in-service teachers. The newer model emphasises the dynamic nature of professional learning viewing it as a range of informal, and individual learning activities teachers engage in in their everyday lives (Ingvarson et al., 2005; Lieberman, 1995). Both informal and formal learning activities at individual and community levels have been consistently found in the scholarly literature as powerful mechanisms for teacher growth and development (Desimone, 2009).

This study hypothesised that teacher item writers' professional learning as teacher engagement in self-regulated professional learning (i.e., the newer model of professional learning or informal professional learning) and the institutional induction and training offered to teachers (i.e., the traditional 'training' model or formal professional learning) are two primary factors affecting their competence in multiple-choice test item writing for English reading and listening skill tests. The self-regulated professional learning activities that teacher item writers are hypothesised/supposed to effortfully and mindfully engage in are those that meet the requirements prominently found in theories and studies of deliberate practice (Ericsson, 1996, 2009), self-regulated learning (Pintrich, 2000; Zimmerman, 2000), reflective learning and experiential learning (Kolb, 1984; Schön, 1987). It was hypothesised that the induction and training offered to Vietnamese EFL teachers item writers by their institutions should bear the critical features of high-quality professional development as characterised in the contemporary related literature.

The following sections review existing frameworks for measuring teachers' self-regulated professional learning and institutional training and development and contextualise these frameworks in the target domain of English test item writing. The adapted frameworks were used in this study to measure English teacher item writers' levels of engagement in both formal and informal professional learning.

4.1.2 Measuring English Teacher Item Writers' Professional Learning

4.1.2.1 Measuring teacher engagement in self-regulated professional learning: The categorical model.

It is recognised that Kwakman (2003) is one of the first authors to have investigated teacher engagement in professional learning at the workplace through a survey instrument. According to the author, teacher learning is governed by three learning principles. First, teacher learning takes place as teachers participate actively in professional development activities within the community of practice. Second, teacher learning is individual and collaborative. Both individual and collaborative levels of learning play significant roles in stimulating teacher learning, and these notions are widely addressed in the professional development literature. At the individual level of learning, teachers take ownership of initiating, planning, implementing and assessing their own learning activities. At the collaborative level of learning, teachers are engaged in collegial interactions in which they are exposed to peer feedback, open discussion, and new ideas. Such collaboration is assumed to nurture a learning culture, and stimulate and support teacher learning. Thirdly, teacher learning is aimed for teachers to develop professionally.

From these principles, Kwakman (2003) distinguished four main philosophical categories of professional learning. The first three individual activities address the core responsibilities of professionals: they are reading to keep up with new professional knowledge; experimenting, that is, applying new ideas and putting effort into in improving their professional practices; and

reflecting. The fourth one addresses collaboration, or the collaborative level of learning which provides teachers with feedback, new ideas and professional support. Subsequently, Kwakman (2003) derived from her qualitative study twenty-one professional learning activities organised in these four categories, namely reading, experimenting, reflecting and collaborating. The activities were formulated as items which made up a measure of teacher professional learning. The items were reviewed by experts and pilot tested before they were embedded into a questionnaire to collect the data for the analysis of the measure's reliability and the items' validity. The responses from 542 teachers across 10 schools in The Netherlands who used the scale to self-report their participation in professional learning were analysed using factorial analysis. The analysis resulted in the removal of three items due to their low factor loadings, but Kwakman (2003) could not empirically confirm the categories of professional learning. Later, Kuiper-Rinsema (2007, as cited in Evers, Kreijn, & Van der Heijden, 2016)) used the same list of items and replicated Kwakman's (2003) study. His analysis led to the deletion of five other items as they were considered not to match the defined theoretical categories.

The categorisations initially postulated by Kwakman (2003) were used by Evers et al. (2016) as the theoretical framework for their design of the Teacher Professional Development at Work (TPD@Work) survey instrument to explore teacher participation in professional learning. The authors incorporated the theoretical categories from Joyce and Calhoun (2010) and extended the categorisation to include five theoretical themes: (1) keeping up-to-date, (2) experimenting, (3) reflecting and asking for feedback, (4) collaborating with colleagues with the aim of improving lessons, and (5) collaborating with colleagues with the aim of improving school development. Evers et al. (2016) built up on the items from both Kuiper-Rinsema (2007) and Kwakman's (2003) with some rephrasing, and adjustment of theoretical categories. The authors added some ICT-related indicators to reflect the substantial role of Information and Communication Technologies (ICT) in professional learning that has emerged from more recent

international studies. This adaptation produced the TPD @Work survey comprising twenty-four items in the five categories. Evers et al. (2016) also made use of Kwakman (2003)'s measurement of how frequently teachers engaged in each learning activity using a self-reported 4-point rating scale: 1= hardly ever, 2= sometimes, 3= fairly often, and 4= often. The TPD @Work survey was pilot tested with educational experts before it was administered in 24 Dutch schools with the population of 2385 teachers. The data collected from 692 teachers who completed the survey was divided into two equal subsets. The first subset was analysed using exploratory factor analysis, and the second used confirmatory factor analysis for cross validation. These analyses confirmed the categorical model of professional learning and showed that the categories identified in their empirical research was seemingly consistent with those in the literature: (1) Keeping up-to-date, (2) Experimenting, (3) Reflecting and asking for feedback, (4) Collaborating with colleagues with the aim of improving lessons, and (5) Collaborating with colleagues with the aim of improving school development. In addition, two items were removed since they loaded on an unexpected factor. The Cronbach's α for all the subscales in the survey showed that the subscales yielded sufficient to good internal consistency (Cronbach's α ranged from .67 - .80, Evers et al., 2016, p. 172). Table 4.1 presents the final list of indicators that were used in the survey to measure teacher participation in professional learning (Evers et al., 2016)

Table 4.1 *The List of Activities for Teacher Professional Development at the Workplace*

Category	Activities
Keeping up to date: Reading	1.Studying subject matter literature
	2.Reading educational/subject matter pedagogical literature
	3. Visiting educational sites on the Internet
Keeping up to date: Work-related training	4. Participating in a training course that centre around subject matter pedagogy
	5. Participating in one day conference or study day that centres around subject matter pedagogy
Experimenting	6. Trying out new applications in ICT in my lesson
	7. Testing alternative teaching materials in class
	8. Applying and evaluating other forms of assessments
	9. Inquiring into new teaching methods in class
	10. Trying out new teaching methods in my lesson
Reflecting and asking for feedback	11. Reflecting on my strong and weak points
	12. Adapting my teaching methods in response to pupils' reactions
	13. Asking pupils for feedback on the way I teach
	14. Inviting colleagues to attend my lesson
Collaborating for improving lessons	15. Preparing lessons with colleagues
	16. Making agreements with colleagues about the pedagogical practices
	17. Discussing the teaching approaches I use in class with colleagues
Collaborating for improving school	18. Giving an opinion together with colleagues about school organisational matters to the school management
	19. Assembling a school working group or committee with colleagues
	20. Thinking about the design and method of pupil counselling with colleagues
	21. Discussing ideas about educational improvement and innovation in my school with colleagues

Note. Reprinted from "The design and validation of an instrument to measure teachers' professional development at work" by A. Evers, K. Kreijn, and B. van der Heijden, 2016, *Studies in Continuing Education*, 38, p. 170-171. Copyright 2016 by Routledge Taylor & Francis Group.

In summary, teachers' engagement in self-regulated professional learning for competence development can be reflected in their participation in four categories of professional learning activities, namely keeping up to date with the professional literature, experimenting with newly learnt skills and knowledge, engaging in reflective practice and collaboration.

4.1.2.2 Measuring institutional induction and training for teacher item writers: The RITCME framework.

It is claimed by testing and assessment services such as ETS, British Council and Cambridge ESOL Examinations (Cambridge English for Speakers of Other Languages Examinations, currently known as Cambridge Assessment English) that in order to ensure the creation of high-quality test items, it is essential that all their item writers receive extensive training on the rigorous principles of and best practices in good item writing. Research has also shown that proper training substantially enhanced item quality (Iramaneerat, 2012; Jozefowicz et al., 2002; Naeem, van der Vleuten, & Alfaris, 2012). The induction and training may be delivered in different modes, such as an online three-month training course (British Council), a one-day short course (ALTE) or a series of training workshops (ETS). The training is designed either for the sake of professional development or for a certification examination. The content of the induction and training centres around principles of and guidelines on test item writing, analysis of what makes good items, collaborative test item writing and exchanges of tasks, peer feedback and lessons to take away.

As described in Chapter 1 (see Section 1.1), the VSTEP® examination boards in Vietnam inclined to adapt the recruitment, induction, training, coordination, monitoring and evaluation processes implemented by Cambridge Assessment English, so the processes used with VSTEP® item writers that were employed by Vietnamese universities could bear some resemblance to the RITCME framework. Additionally, Cambridge Assessment English could be the one testing organisation which has published a large number of serial publications and reports that address

the most significant issues in language test development aiming to provide the stakeholders with useful insights into the standards and practices applied to their training and development of writers of high-stakes Cambridge Assessment English assessments. The RITCME framework may well be the most prominent training model for language test item writers authored and implemented by Cambridge Assessment English, one of the global leaders in language test development and research. The framework for the training and development offered to Cambridge Assessment English item writers, as described by Ingham (2008) has an acronym RITCME: Recruitment, Induction, Training, Co-ordination, Monitoring and Evaluation. The application of the RIT (i.e., Recruitment, Induction and Training) part of this standards process is to make sure that novice item writers have appropriate qualification, relevant experience and are provided with training in test item writing skills and information on the work processes. The CME (i.e., Co-ordination, Monitoring and Evaluation stages) provide opportunities of face to face dialogues between Cambridge Assessment English and the item writers. The RITCME framework for the training and development of item writers at the University of Cambridge Assessment English is elaborated in Ingham (2008). In general, the induction and training for item writers often consists of two full-day sessions, covers general item writing issues and includes hands-on practice on text selection, text adaptation and critical analyses of sample and live test items. Trainees are provided with an overview of tests delivered by Cambridge Assessment English and an introduction to the test development and production principles. They are given information on typical test question formats (e.g., multiple-choice, productive, matching) and to the fundamental terminology (e.g., stem, key, distractors) which is presented in the *Item Writer Guidelines*. The duration for each session on the item writing techniques is approximately 120 minutes and is balanced to permit both input from the trainer and contributions of ideas and experience from the participants in group activities. The content focus of the training sessions is on analysing the strengths and drawbacks of test item types so as to

identify the features of sound test items and the training sessions are conducted in the way that promote the participants' active learning. Typically, each of test item types is introduced and it is open to discussion on components and critical features of the test item, its strengths and weaknesses and important implications for writing such an item. After that, examples of the items are given for the participants who work in groups to critically evaluate the negative and positive aspects of the item. The training also covers a basic overview on writing test items for specific skills papers and provides the participants hands-on practice on specific item writing tasks such as text selection and adaptation which the participants may be requested to prepare before the training session. The item writers are asked to give feedback on the session and to choose the skills paper they would prefer to write. Based on their preference, the participants are allocated into skills specific training teams for the next day. A skills paper-specific training event is also provided and often led by the Chair of the item writing team who has substantial item writing experience with the participation of both existing and new item writers. The training centres around three main tasks: selection of texts, adaptation of written sources, and feedback on live test. First, the Chair pre-selects a range of texts from different sources. Item writers of varying experience are allocated into groups to discuss each text and its potential for types of test items on the paper they could be used. The item writers engage in fruitful discussion on factors affecting their choice of texts and item writing techniques. Experienced writers may demonstrate how swiftly they evaluate a text and annotate the text where it would yield desirable test items. The discussion is also to focus on techniques to adapt written texts to create a meaningful and authentic listening experience for test takers. These techniques are often elicited from the participants and exemplified with concrete examples for source texts. At the end of the discussion, they may reach a consensus on the best texts to use. Then, the training is also devoted to considering performance of test items from the latest operational test material. The participants review all items and are requested to speculate which items candidates may find

most difficult, and which distractors functioned best. The participants are given classical item statistics related to each test item and the Chair explains to them - especially to the new item writers - some of the basic measurement characteristics of test items such as facility values, discrimination values, point-biserial correlation, and the acceptable range for these values targeted by Cambridge Assessment English. The Chair identifies some items to investigate more carefully, in order to communicate appropriate implications for item writing. The induction and training activities for English reading and listening test item writers at Cambridge Assessment English as reported by Ingham (2008) are summarised in Table 4.2

Table 4.2 *Induction and Training Activities for English Reading and Listening Test Item Writers at Cambridge Assessment English*

Induction
1. Provide background information on the test production process and documentation on what is expected of Cambridge Assessment English item Writers
Pre-training
2. Provide a pre-training pack including tasks relating to text selection and adaptation and a list of typical features of spoken English with examples
Training
3. Introduce typical test tasks and the basic terminology used to describe test items
4. Provide training on techniques of writing multiple-choice test based on input from the trainer and group activities drawing on the ideas and experience of the participants
5. Provide specific training on adapting written sources, selecting and sourcing texts, and approaches to item writing
6. Consider the strengths and drawbacks of multiple-choice as a test type and to identify the features of sound multiple-choice items
7. Consider and discuss in groups examples of multiple-choice items
8. Consider the most recent live test paper, discuss and give feedback on how well the item had performed
9. Work in groups of new and experienced item writers on pre-selected a range of texts from different sources and discuss the best texts to use in terms of topic, content and style of writing and practice writing items for the selected text
10. Work in groups to revise and adapt sample texts
11. Offer a training weekend with Cambridge Assessment English item writer trainers

Note. Adapted from "The Cambridge ESOL approach to item writer training: The case of ICFE listening" by K. Ingham, 2008, Retrieved from http://www.cambridgeenglish.org/images/23151-research-notes-32.pdf (p.5-7). Copyright 2008 by UCLES.

It can be seen that the training undertaken within the RITCME framework clearly reflects the five critical features of professional training programs synthesised by Desimone (2009), namely content focus, active learning, coherence, duration and collaboration. Desimone (2009) reviewed various case studies, correlational, quasi-experimental, and experimental research of the past several decades in the field of teacher professional development and concluded that there is a research consensus on the core characteristic of effective professional development programs. He put forward the core conceptual framework for high-quality professional development which consists of five critical features: content focus, active learning, coherence, duration, and collective participation. The framework is said to reflect a new paradigm for professional developing, rejecting the ineffectiveness of the traditional "drive-in" workshop model (Stein, Smith, & Silver, 1999). These core features of an effective professional training program are echoed in the research reviewed by Darling-Hammond and Richardson (2009). In sum, it is suggested that professional training should be intensive, content-rich, provide opportunities for teachers to acquire new knowledge through active, hands-on learning, apply it to practice, reflect on the result with colleagues and promote collegial and collaborative learning (Darling-Hammond & Richardson, 2009, p. 6).

To date, the only measure of professional development at institutional levels is the one called the Observation Checklist for High-Quality Professional Development Training (HQPD Checklist developed by Erickson, Noonan, Brussow, & Carter, 2017). The Checklist is a set of observable indicators of effective professional development training initially developed through the authors' extensive review of the scholarly literature on features of first-rate professional development. The Checklist was reviewed by professional development providers and evaluators before it was field tested by observing professional development trainings in four states in America. It was then beta-tested through the observation of 99 diverse trainings in America, further refined and subjected to an inter-rater reliability process. The result of this four-year

project was the Checklist being presented in six domains with 22 dichotomous choice items. This is claimed to be a valuable tool for institutions to support the learning of participants, inform their design of and reflection on the content of the training and for assessors to monitor the professional development quality (Erickson et al., 2017, pp. 687-688). The Checklist is presented in the following table (see Table 4.3).

Table 4.3 *The Observation Checklist for High-Quality Professional Development Training (HQPD Checklist)*

Preparation
1. Provides a description of the training with learning objectives prior to training
2. Provides readings, activities and/or questions to think about prior to the training
3. Provides an agenda (i.e. schedule of topics to be presented and times) before or at the beginning of training
4. Quickly establishes or builds on previously established rapport with participants
Introduction
5. Connects content to participants' context (e.g. community, school, district)
6. Includes the empirical research foundation of the content (e.g. citations, key researchers)
7. Content builds on or relates to participants' previous professional development
8. Aligns with school/district/state standards or goals
9. Emphasizes impact of content on student learning outcomes
Demonstration
10. Builds shared vocabulary required to implement and sustain the practice
11. Provides examples of the content/practice in use (e.g. case study, vignette)
12. Illustrates the applicability of the material, knowledge or practice to the participants' context
Engagement
13. Includes opportunities for participants to practice and/or rehearse new skills
14. Includes opportunities for participants to express personal perspectives (e.g. experience, thoughts)
15. Includes opportunities for participants to interact with each other related to training content
16. Adheres to agenda and time constraints
Evaluation
17. Includes opportunities for participants to reflect on learning
18. Includes discussion of specific indicators – related to the knowledge, material or skills
19. Engages participants in assessment of their acquisition of knowledge and skills
Mastery
20. Details follow-up activities that require participants to apply their learning in a new setting or context
21. Offers opportunities for continued learning through technical assistance and resources
22. Describes opportunities for coaching to improve fidelity of implementation

Note. Reprinted from "Measuring the quality of professional development training" by A. Erickson, P. Noonan, J. Brussow, & K. Carter, 2017, *Professional Development in Education,* 43, p. 687. Copyright 2017 by Routledge Taylor & Francis Group.

The induction and training activities for Cambridge Assessment English reading and listening item writers described in the RITCME framework by Ingham (2008), the HQPD Checklist developed by Erickson et al. (2017) for evaluating high quality professional training and the critical features of professional training identified by Desimone (2009) and Darling-Hammond and Richardson (2009) provide insights into the features and types of institutional induction and training activities that could support teachers' development of expertise in multiple-choice test item writing.

4.2 Educational Background Characteristics Affecting Teacher Competence in Test Item Writing

Educational background characteristics have also been found to potentially influence teacher item writers' competence. These characteristics are levels of academic qualification, degree training in TESOL/TEFL, applied linguistics and other related disciplines and pre-service uptake of a course in language testing and assessment.

According to the language testing and assessment standards of practice, it is suggested that item writers exhibit two main attributes: language teaching experience "at the level the assessment is aimed at" (EALTA, 2006, p. 3) and knowledge of "current language testing theory and practice" (ILTA, 2007, p. 3). Buck (2009) adds that professional item writers are highly trained, often having a master's degree in applied linguistics, TEFL/TESOL or a related discipline, and that well qualified item writers are often sought after. To be recruited as item writers for the Cambridge Assessment English, candidates must possess a Bachelor degree, a relevant qualification and five years of experience in language teaching, and some experience with material production and involvement in preparing students for Cambridge Assessment English exams would be desirable (Ingham, 2008). No research to date has attempted to investigate if a degree in language teaching would predict test item writers' competence. Mixed results have been found with regard to the levels of academic qualification of teachers and their

influences on their assessment knowledge and practice. While Chapman (2008) found no correlation between teachers' academic qualifications and their assessment literacy, King (2010) reported that the assessment knowledge possessed by teachers with PhD degrees was considerably higher than that of those who had bachelor or master's degrees. However, these studies considered the qualification levels only and did not consider the specialised disciplines that the participants enrolled in. The case study of EFL reading test item writers by Kim et al. (2010) revealed that the item writers' educational background influenced their preference of the topics and content of the test materials they chose to work on. For example, item writers self-reported to be more comfortable writing reading texts related to their university majors. The interviews with the principal investigators and two coordinators of an item writing project for an online English proficiency test for adult English language learners regarding selection criteria for item writers in Kim et al. (2010) revealed pre-service academic training as one major criterion. It was found that it was most important that the item writer had prior training in related majors, namely TESOL/ TEFL, and applied linguistics, and possibly a language testing and assessment course in their degree study. It was assumed that such training would acquaint item writers with key concepts and issues in specifications-driven testing (Kim et al., 2010, p.169). The item writers in Kim et al.'s (2010) study confirmed that the knowledge of test specifications and the spec-driven item writing process that they gained in language testing courses they took as part of their degree program greatly contributed to their growth as item writers.

Thus, it was assumed that levels of academic qualification, pre-service degree training in TESOL/TEFL, applied linguistics and other related disciplines and uptake of a course in language testing and assessment may influence English teachers' item writing proficiency.

4.3 Professional Background Characteristics Affecting Teacher Competence in Multiple-Choice Test Item Writing

The length, scope and quality of work-related experience have been of research interest for decades. Research in the early 1980s showed that length of experience is a significant variable when comparing students with professional practitioners (Alderson, 1981; Jeffries, Turner, Polson, & Atwood, 1981). Some studies examining how professionals differed with regard to their length of experience found that performance and working processes executed by experienced professionals were different from those by less experienced colleagues (Koubek & Salvendy, 1991; Wiggins & O'Hare, 1995). Other research did not find any significant relationship between length of experience and level of performance (Charness, Krampe, & Maye, 1996; Ericsson et al., 1993; Sonnentag, 1995). There has been a plethora of research exploring the role that length of experience plays as a critical role for differences in performance levels among professionals. Sonnentag (1998) and Sonnentag and Kleine (2000) used job performance and peer-nomination measures to explore the relationship between length of experience and excellent performance, and the findings revealed that it was not length of experience but the quality of continuing professional activities that influenced performance. The regression analysis result from Dunn and Shriner (1999) corroborated Sonnentag's (1998) finding that years of experience did not contribute, but the nature and the amount of effort devoted to deliberate practice activities predicted levels of proficiency.

No empirical research to date has investigated the relationship between length of experience and test item writing expertise in detail. However, this issue has been discussed in some case studies and publications on test item writers. It was observed by Alderson et al. (1995) that experience as a teacher provides item writers with insightful understanding of test takers' cultural background, interests, the skill areas they are strong and weak at, and the progress they may make at different stages of their language acquisition process. According to

Alderson et al. (1995), item writers who are experienced subject specific teachers with relevant qualifications are likely to produce better test items than those without such experience (Alderson et al., 1995, p. 40). Their survey with EFL examination boards showed that their selection criteria for item writers were appropriate EFL/ESL qualifications, relevant teaching and examining experience and hands-on experience in preparing students for English language exams. Item writers' ESL/EFL teaching experience was another crucial criterion and was considered to be a good resource for writing items. The item writers valued the experience as ESL/EFL teachers as it helped them understand ESL/EFL learners, who are the prospective test takers, and know the difficulties they have and the areas that they frequently make mistakes in. This resonates with the findings from Green and Hawkey (2007) which showed that the participants reported recalling their previous relevant experience or drawing on their experience when writing test items (p.290). Additionally, the non-native item writers in Kim et al. (2010) indicated that their previous test taking experience might be beneficial for their item writing because they are familiar with the test types, which enabled them to perceive how test takers would respond to the items. In other words, with their previous experience as test takers, they can look at the items from the examinees' perspective. Seemingly, EFL teacher item writers draw much on their teaching as well as their experience in test item development. The length and types of their prior experience may influence their proficiency in test item writing.

It was found that the item writers' native and non-native status seemed to affect their item writing process (Kim et al., 2010). While non-native item writers were reported to have no significant difficulty creating authentic reading materials and test items compared to native ones, they did focus more on the grammatical accuracy of the language use in both texts and items and needed their materials to be reviewed and further edited for naturalness and authenticity by native colleagues. Native item writers often took roles as editors and reviewers and found this to be in their comfort zone. In the EFL context of Vietnam, the assumption could be that teacher

item writers' English language proficiency levels may affect their item writing practice and the quality of their items.

4.4 Research Questions Revisited

The study proposed to achieve two main goals associated with two primary research questions. The first goal was to develop and validate a developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. The second goal was to explore the factors that contribute to the development of Vietnamese teachers' perceived competence in multiple-choice test item writing for English reading and listening skill tests. The two primary research questions are as follows:

1. What is the typical developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests?

2. What teacher-related factors predict Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

The literature reviewed in this chapter has indicated the need for examining a set of factors potentially contributing to Vietnamese teachers' perceived competence in multiple-choice test item writing for English reading and listening skill tests. Three sets of teacher-related factors, namely the teachers' engagement in professional learning, and their professional and educational background characteristics, were hypothesised to make an explanatory contribution to the teachers' judgements of their multiple-choice test item writing competence. The subsidiary questions for the second primary research questions are:

2.1 To what extent does professional learning *(i.e., self-regulated professional learning and access to institutional induction and training)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.2 To what extent do teachers' professional background characteristics *(i.e., length of experience in teaching English language skills, teaching English exam preparation courses,*

constructing English tests from published resources, and writing multiple-choice test items) predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.3 To what extent do teachers' educational background characteristics **(***i.e., their highest degree's major, uptake of coursework in language testing and assessment and their English language proficiency)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

The teacher-related factors that are hypothesized to predict teachers' perceived levels of competence in multiple-choice test item writing for English reading and listening skill tests can be graphically presented in *Figure 4.1* as follows:

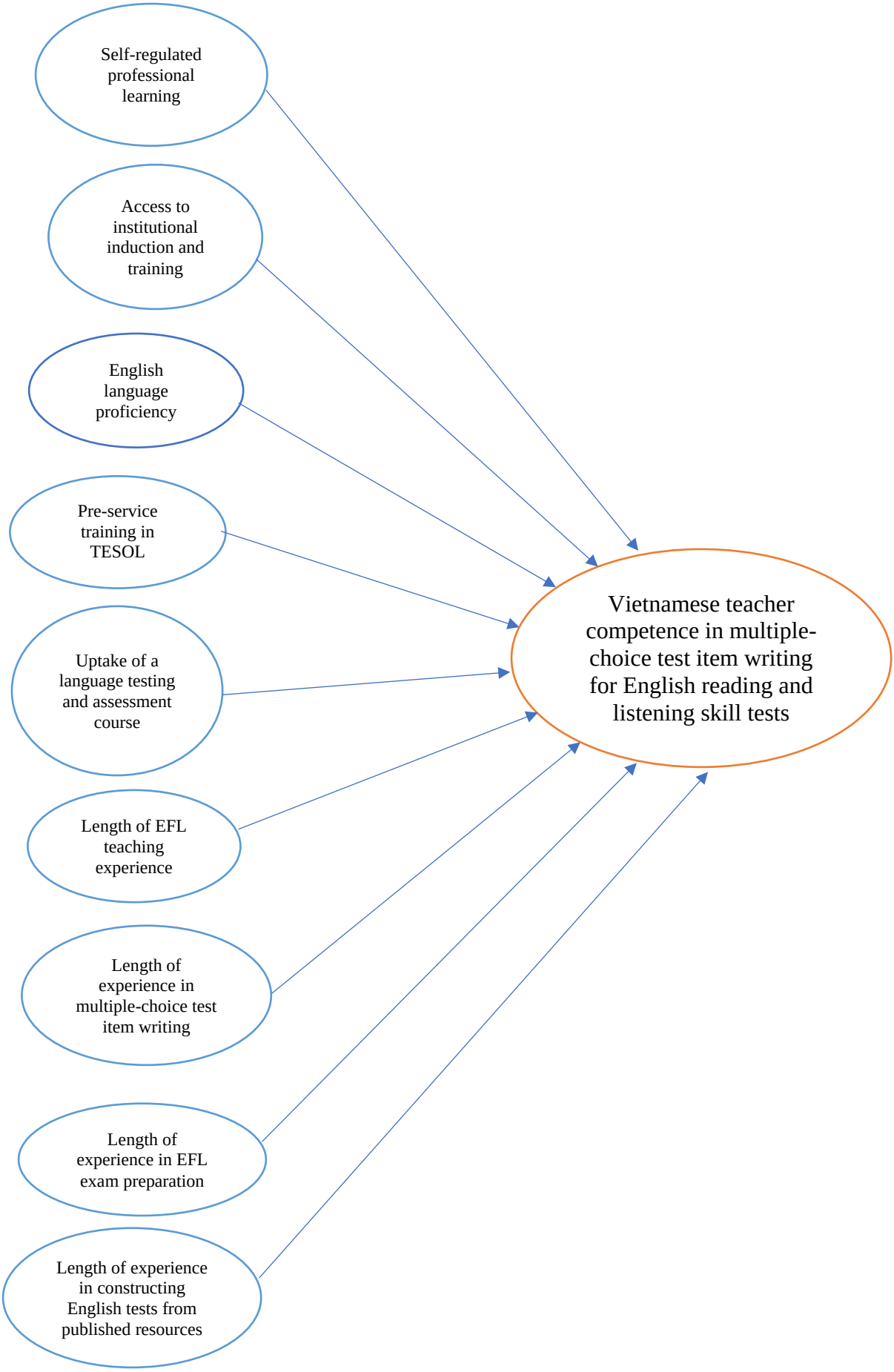


Figure 4.1 Model for prediction of Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.

4.5 Chapter Summary

The chapter focused its discussion on professional learning as a mechanism which drives teacher professional competence growth. Models and measures of professional learning in teachers' workplaces from prior research were examined, and contextualised. The items in existing measures of teacher participation in professional learning were adapted to suit the practice of teacher item writers. The lists of indicators were produced for use in the adaptation of the scales to measure English teacher item writer engagement in self-regulated professional learning and their access to institutional induction and training (see Table 4.1 and Table 4.2). The chapter also identified educational background characteristics (i.e., pre-service training in TESOL, uptake of a course in language testing and assessment, and teachers' English language proficiency), and the professional background characteristics (i.e., length of experience in teaching English language skills, teaching English exam preparation courses, constructing English tests from published resources, and writing multiple-choice test items) as other factors that affect Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. At the end of the chapter, the second primary research question was revisited and graphically presented. It is hypothesised that Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests could be predicted from their engagement in self-regulated professional learning, their access to institutional induction and training, and their educational and professional background characteristics.

CHAPTER 5: RESEARCH METHODOLOGY

This chapter discusses details of the research participants including sample composition and method of recruitment, and the development of the data collection instruments. It also elaborates on the methods and procedures for collecting and analysing the data for the investigation into teacher competence in multiple-choice test item writing for English reading and listening skill tests in Vietnamese higher education settings.

5.1 Scales and Questionnaire Development

This section details the development and adaptation of the three scales to measure three latent variables in the study. The first scale was developed for Vietnamese EFL teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests (the IWCOMP scale). The second scale was adapted from the Teacher Professional Development at Work (TPD@ Work) survey instrument developed by Evers et al. (2016) for the teacher participants to self-report their engagement in self-regulated professional learning (the PROLEARN scale). The third scale was adapted from part of the Cambridge Assessment English's RITCME framework (Ingham, 2008) for the teachers to self-report their access to institutional induction and training (the INTRAIN scale). The development and adaptation of the scales were based on an extensive review of the scholarly literature, scrutinised at focus groups and workshops with subject matter experts, and reviewed by a panel of experts. All the items in the scales along with the questions about the teachers' professional and educational background information were subsequently collated into a questionnaire. The questionnaire comprised three main parts. The first two parts included 60 items from the IWCOMP, PROLEARN and INTRAIN scales developed and adapted in the study. The third part included 10 items to collect the participants' background variables. Given that all the participants were experienced English teachers who had numerous years teaching English at university level and the development and

adaptation of the scales also involved English native experts, the scales and questionnaire were written in English.

The following section describes the process employed to develop the IWCOMP scale (i.e., the dependent variable) as well as those employed to adapt the PROLEARN and INTRAIN scales (i.e., the independent variables). It also outlines the process to design the items to collect the participants' demographic data.

5.1.1 The Development of the Progression Scale for Vietnamese EFL Teacher Competence in Multiple-Choice Test Item Writing (The IWCOMP Scale)

5.1.1.1 Scale Development Stages

The basic approach to the development of the IWCOMP scale and subsequent calibration of this scale was based on both theoretical and psychometric approaches to scale development (Gillis, 2003). The development of the IWCOMP scale underwent four main stages: drafting, focus groups, workshops with SMEs and expert review. *Figure 5.1* outlines the components of the framework for Vietnamese EFL teacher competence in multiple-choice test item writing. Table 5.1 summarises the methods employed in the development of the competency framework from which the items for the IWCOMP scale were generated. The informants and the activities involved in each stage of the development process of the IWCOMP scale are described in the next subsections.

Domain: Multiple-choice test item writing for English reading and listening skill tests targeted as competence Vietnamsese EFL teachers need to develop.

Phases: Prior research agreed upon and defined that multiple-choice test item writing is a three-phased problem solving process.

Competencies: Major job functions that multiple choice test item writers for English reading and listening tests perform as identified from research and practice.

Performance indicators: The behavioural indicators are the things that a teacher does, makes, says or writes, and from which their competency can be inferred.

Quality criteria: These criteria describes 'how well' each performance indicator can be typically demonstrated.

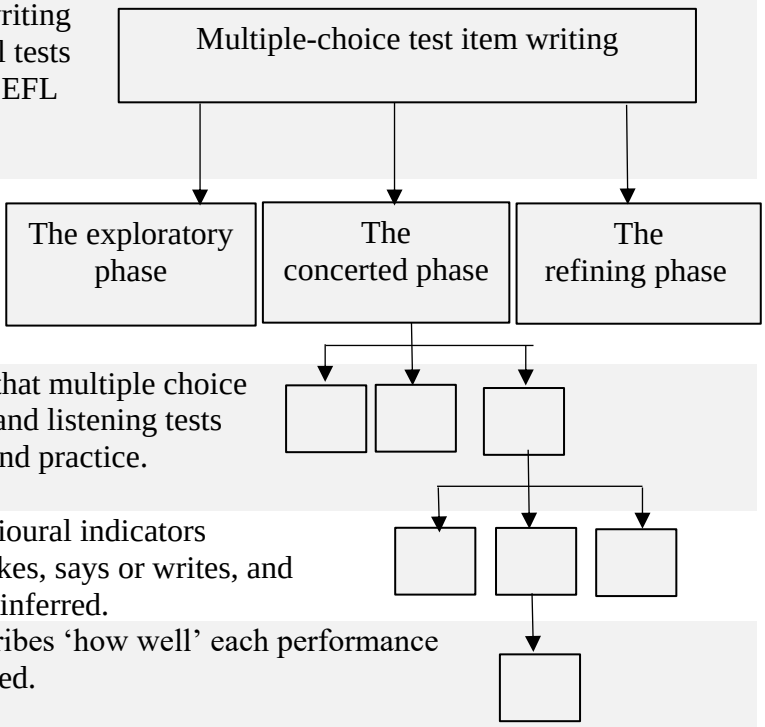


Figure 5.1 Hierarchical organisation of the components of the framework for Vietnamese EFL teacher competence in multiple-choice test item writing.

Adapted from "Standard-referenced assessment for vocational education and training in schools" by P. Griffin, S. Gillis and L. Calvitto, 2007, *Australian Journal of Education, 51,* p. 25. Copyright 2007 by SAGE Publications and from "Judgement-based assessment", by P. Griffin, M. Francis and P. Robertson, in P. Griffin (Ed.), *Assessment for teaching* (p.98), 2018, Cambridge: Cambridge University Press. Copyright 2018 by Cambridge University Press.

Table 5.1 *Methods Employed in the Development of the Draft Framework for Vietnamese EFL Teacher Competence in Multiple-Choice Test Item Writing*

Methods	Purposes
Literature review	• To generate a draft framework for teacher competence in multiple choice test item writing for English reading and listening skill tests
Focus groups	• To examine the appropriateness and relevance of the draft competency framework in a Vietnamese higher education setting
Workshops with SMEs	• To review and agree upon the phases, competencies and performance indicators identified from the literature, examined at and amended following the focus groups • To write the quality criteria for the competency framework and compare and contrast them with the draft ones
Expert review	• To improve the conciseness of and review the ordering of the quality criteria for each performance indicator

5.1.1.1.1 Drafting.

The main sources of material came from extensive literature reviews of different models of item writing (Fulkerson et al., 2009, 2010, 2011; Salisbury, 2005), comprehensive handbooks (AERA et al., 2014; Haladyna & Downing, 1989; Haladyna et al., 2002; Lane et al., 2016), guidelines for item writers (ALTE, 2005; EALTA, 2006; ETS, 2003; ILTA, 2007) and research in item writing practices (Alderson et al., 1995; Baranowski, 2006; Beddow, 2012; Davison & Lynch, 2003; Green & Hawkey, 2011; Kim et al., 2010; Livingston, 2006; Lynch & Davison, 1994; Maau-Aduli & Zimitat, 2012; Moreno et al., 2015; Osterlind, 1998; Raymond, 2015; Rodriguez, 2016; Sydorenko, 2011; Wise & Blake, 2016). This literature provided both specific content for the framework and vital information about the requirements, expectations and processes involved in test item writing. This conceptual approach was initially used to derive a draft framework for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. The configuration of the framework was based on the

hierarchical structure of competence put forward by Griffin et al. (2007) and revisited by Griffin, Francis, et al. (2018) (see *Figure 5.1*).

The literature review in Chapter 3 served as the basis for assuming three distinct stages in the domain of multiple-choice test item writing: the exploratory phase, the concerted phase and the refining phase. The three phases were revealed in five competencies expressed in terms of major job functions: analysing test specifications, generating input materials for tests, writing test items, reviewing and revising test items based on feedback (see *Figure 3.3*). When the distinct competencies had been defined, fourteen observable performance indicators were derived from various sources in the literature on models of item writing and research in item writing practice as reviewed in Chapter 3 of the thesis (see *Figure 3.4*). The quality criteria for each performance indicator were generated from the relevant literature and drafted in accordance with the principles underlying the design of the framework, the rules for rubrics writing established by Griffin et al. (2004) and revisited in Griffin, Francis, et al. (2018), and the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) (see Section 2.2). The items within the IWCOMP scale to enable Vietnamese teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests were made up from the performance indicators and the quality criteria within the framework. Table 5.2 displays the competencies established as representative components of the construct of test item writing, and the performance indicators for each of the competencies as sourced from the literature. Table 5.3 presents an example of the draft performance indicators and quality criteria for the first competency. The components in the framework were coded as follows: the first digit refers to the phase, the second digit refers to the competency, the third digit refers to the performance indicator, and the fourth digit represents the quality criterion. Appendix E provides the first draft of the framework in which the provenance of each item is specified.

Table 5.2 *Draft Competencies and Performance Indicators for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests*

Phases	Competencies	Performance Indicators	Sources						
			Alderson et al. (1995)	Davison and Lynch (2003)	Green and Hawkey (2011)	Kim et al. (2010)	Lynch and Davison (1994)	Salisbury (2005)	Sydorenko (2011)
1. The exploratory phase	1.1Analysing test specifications	1.1.1 Review test specifications	☑	☑		☑	☑	☑	
		1.1.2 Utilise test specifications	☑	☑		☑	☑	☑	
		1.2.1 Select source texts/materials for the test			☑			☑	☑
		1.2.2 Evaluate source texts to make decisions			☑			☑	☑
	1.2 Generating materials for tests	1.2.2 Evaluate source texts to make decisions			☑			☑	☑
		1.2.3 Pre-edit the generated texts for item writing			☑			☑	☑
		1.2.4 Adapt source texts			☑			☑	☑

Phases	Competencies	Performance Indicators	Sources						
			ALTE (2005)	Alderson et al. (1995)	Beddow (2012); Haladyna and Downing (1989)	Haladyna et al. (2002) Haladyna and Rodriguez (2013)	Moreno et al. (2015): Rodriguez (2016)	Salisbury (2005)	Sydorenko (2011)
2. The concerted phase	2.1 Writing test items	2.1.1 Approach item writing						☑	
		2.1.2 Write item stems	☑		☑	☑	☑		
		2.1.3 Write distractors	☑		☑	☑	☑		
		2.1.4 Provide answer keys						☑	
		2.1.5 Estimate item difficulty		☑					☑
			AERA et al. (2014) ; Alderson et al. (1995)	Baranowski (2006); ALTE (2005)	Haladyna and Rodriguez (2013)	Lane et al. (2016); Livingston (2006)	Maau-Aduli and Zimitat (2012)	Osterlind (1998)	Raymond (2015); Wise and Blake (2016)
	2.2 Reviewing test items	2.2.1 Review item content fit	☑	☑	☑	☑	☑	☑	☑
		2.2.2 Verify distractors and keys	☑	☑	☑	☑	☑	☑	☑
			Alderson et al. (1995)	Salisbury (2005)					
3. The refining phase	3.1 Revising test items	3.1.1 Respond to item feedback	☑	☑					

Table 5.3 *Draft Quality Criteria Levels for Competency 1.1: 'Analysing Test Specifications'*

Competency 1.1: Analysing test specifications				
Performance Indicators	Quality Criteria	Sources		
		Davidson and Lynch (2002)	Kim et al. (2010)	Salisbury (2005)
1.1.1 Reviewing test specifications	1.1.1.1 Check the test purposes, the test format, the distribution of test items and their targeted levels within the constructs	☑	☑	☑
	1.1.1.2 Identify targeted test item content, and levels of item difficulty	☑	☑	☑
	1.1.1.3 Recall important aspects of the test to be written/ Generate ideas about what and how to write the items for the test	☑	☑	☑
	1.1.1.4 Use mental models of target examinees to trigger operational consideration of test specifications or Consider/contextualise operational implications of the test specifications	☑	☑	☑
1.1.2 Utilising test specifications	1.1.2.1 Follow the instructions from the test specifications	☑	☑	☑
	1.1.2.2 Recall my prior (item writing) experiences to interpret the test specifications	☑	☑	☑
	1.1.2.3 Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications	☑	☑	☑
	1.1.2.4. Recall previous experience and discuss with colleagues to internalise content of test specifications	☑	☑	☑
	1.1.2.5 Refer to own internal guideline for item writing based on the given test specifications and/or other training materials/resources and prior experiences	☑	☑	☑

5.1.1.1.2 Focus groups.

After the first draft of the competency framework was generated from the literature, the focus groups were conducted with the subject matter experts to gather evidence of the content validity of the framework. The purpose of the focus groups was to gain insights into the item writers' tasks and to test out the draft competency framework derived from the scholarly literature in a rich communicative context and in a contemporary Vietnamese higher education setting. The description of focus groups as 'structured eavesdropping' by Powney (1988 cited in Barbour, 2008, p. 134) may capture the core essence of the method. Such discussions stimulate direct interactions among participants and allow the researcher as a facilitator to engage with participants and encourage them to elaborate on their perspectives and experiences. There were several good reasons for why focus groups were used in the exploratory phase of this research project to inform the development and refinement of the research instruments. First, focus groups offered the researcher a great degree of control over the composition of the group and the main focus of the discussion. Second, focus group discussions provided a valuable opportunity for generating collective input, sharpening and refining ideas, and elaborating experiences through face to face social interaction between participants (Wilkinson, 1999). This helped to penetrate into cognitive processes, knowledge structures and schema activated by item writers when they perform their task. It may also illuminate similarities and differences in individual item writers' perceptions, and practices. Third, the 'spontaneity' arising from the social context of focus group discussions (Finch, Lewis, & Turley, 2014) triggered more in-depth and reflective discussions, and elicited rich qualitative data for the research.

A group of twelve participants was recruited from the researched institution which is one of the leading foreign language universities in Vietnam. These participants were identified as the key workforce involved in the creation, development and operation of VSTEP®. Of the twelve participants, one was the Executive Director of VSTEP® specifications development project and

one was the Director of the VSTEP® examination board. These were classified as the executives. The other ten participants were classified as the item writers. Five item writers were the permanent staff of the institution's language testing and assessment centre, which was the first VSTEP® examination board officially authorised by the government of Vietnam to develop operational VSTEP® tests, to deliver VSTEP® examinations, and award VSTEP® certificates to Vietnamese learners. The other five item writer participants were the EFL teachers working at the university faculties who had been commissioned to write items for VSTEP® reading and listening tests in the last two years since the initial launch of VSTEP®. The five item writers who worked full-time at the centre for language testing and assessment were classified as internal item writers and the five EFL teachers from the faculties who were commissioned to write items for VSTEP® were identified as external item writers. The roles of the participants and their background information have been presented in Table 5.4.

Table 5.4 *Focus Group Participants' Background Information*

Participants' roles	Numbers of participants	Background information		
		Title/ Qualification	Major discipline/ Country of study	Years of experience in teaching English
Executive Director of VSTEP® project	1	Professor	Applied Linguistics (Vietnam)	30 years
Director (the VSTEP® examination board)	1	PhD	Applied Linguistics (Australia)	18 years
Item writers	10	1PhD	Language Testing and Assessment (Vietnam)	14 years
		1 PhD	Language Testing and Assessment (Australia)	16 years
		1PhD	Applied Linguistics (Vietnam)	17 years
		1 MA	Applied Linguistics (Australia)	7 years
		1 MA		
		1 MEd	Education (Korea)	10 years
		1 MA	TESOL (Australia)	8 years
		1 MA		
		1 MA		
		1 MA		

Four focus group discussions were conducted at the researched institution in Ha Noi, Vietnam in July 2017.Three focus group discussions were conducted with the item writers, and the other was with the executives. Each discussion lasted for approximately from 60 to 90

minutes. The focus groups with the item writers conducted in this project not only provided a valuable means of capturing a panorama of item writers' tasks but also enabled derived competencies in the draft framework for Vietnamese EFL teacher competence in multiple-choice test item writing to be reviewed in a context relevant to Vietnamese universities.

The purpose of the focus groups with the item writers was to examine the appropriateness and comprehensiveness of the competencies and performance indicators derived from the review of the scholarly literature. It was aimed to elicit important information "from-the-inside" perspective that sought to describe the practice as the ways different item writers experienced their item writing task (Marton, 1981, p. 181). It helped to gather input from teachers with advanced knowledge in EFL acquisition and teaching methodology, many years of EFL teaching experience, and hands-on practice in writing English test items.

At the focus groups, the item writers were asked to reflect on and describe the process they undertook when writing multiple-choice test items for VSTEP® reading and/or listening skill tests. The item writers' reflections on the major tasks they performed and the ways they executed the tasks were given special attention because these ideas and behaviours could be expected to reflect the practice of English multiple-choice test item writing conducted in a real work context and to be relatively consistent with what had been identified from the scholarly literature. The focus groups helped develop a perspective on item writers' actual practice within the EFL context in Vietnam. This perspective was formatively used to elaborate the performance indicators of item writers' competencies and adjust the language used to describe certain concepts and determine the level at which statements in the framework are targeted.

Following the reflective discussion, the item writers were requested to look at the draft competency framework and to review and critique its comprehensiveness, its relevance in the context of Vietnam and accessibility in terms of the language use for Vietnamese item writer participants. The part of the draft competency framework that presented the phases, major job

functions/competencies for writing multiple-choice test items for English Reading and Listening skill tests and their corresponding performance indicators (see Appendix E, p. 334) were distributed to the participants. In light of the focus group discussion, the participants were asked to give comments on whether the draft competencies and indicators in the initial framework captured the most significant aspects of their item writing practice in the context of Vietnam. The participants discussed with each other, signed off the competencies and performance indicators in the framework if they were relevant to their practice and made comments on the language use of these items. This was followed up with a discussion on the discrepancies between what the participants had said based on their practical experience and what had been identified from the literature; in light of the discussions, it could be seen how the framework might need to be modified. In conclusion, the analysis of the literature and relevant materials shed light on the initial formation of the framework. The information from the focus groups helped to confirm the relevance of the identified competencies and indicators and to improve the specificity and consistency of the language use of these items. The summary of the changes made to the competencies and indicators in the initial framework as recommended by the focus group participants can be found in Appendix F.1.

5.1.1.1.3 Workshops with Subject Matter Experts (SMEs).

After the focus groups, two four-hour workshops[1] with subject matter experts (SMEs) were conducted at the researched institution in Ha Noi, Vietnam. The fifteen SMEs participating in the workshops were considered by the faculty deans to be experts in teaching English as a Foreign Language at the researched tertiary institution. They had an average of ten years of teaching experience. All the participants were identified to be among the first groups of English teachers in Vietnam who pioneered and were engaged in the practice of writing items for high-stakes standardised English reading and listening tests since the onset of

[1] The workshops went longer than anticipated with the participants' consent.

VSTEP®. Ten of the workshop participants were those attending the focus groups. The other five participants had recently been commissioned to write items for VSTEP®. At the time the workshops were organised, they were 15 out of 231 teachers from the university's English faculties who were commissioned to write items for VSTEP®. Five of the internal item writers had stopped teaching at the English faculties since they were delegated to work for the VSTEP® examination board. The others taught full time at the faculties and were commissioned to write items for VSTEP® reading and listening skill tests. Of the fifteen participants, three had a PhD degree, two in language testing and assessment and one in applied linguistics; two had an MA degree in applied linguistics, one had a Master of Education and nine had an MA degree in TESOL. Their years of teaching experience ranged from 7-17 years and their years of experience in writing items for VSTEP® reading and listening skill tests were from 0.5 to 2.5 years. Given their educational and professional background, these SMEs were identified as expert English language teachers who were also practising as multiple-choice test item writers for English reading and listening skills.

The main purpose of the workshops was to develop the quality criteria for the framework for Vietnamese teacher competence in multiple-choice item writing for English reading and listening skill tests. Across the workshops, firstly, the SMEs were introduced to the background of and aim of the research on Vietnamese teacher competence in multiple-choice item writing for English reading and listening tests. The main findings from the focus groups were summarised and presented to all the SMEs. The phases, competencies and performance indicators of the draft framework identified from the literature and amended based on the recommendations from the focus groups with the item writers were presented to the SMEs for their review (see Appendix G, p. 361). The three components of the revised draft framework were distributed to the participants and they worked collaboratively in groups of five experts to discuss, evaluate and reach consensus on the comprehensiveness and relevance of the proposed

phases, competencies, and the performance indicators for each competency required of English reading and listening skill test item writers. After that, the researcher proceeded to introduce the main objective of the workshops, which was for the SMEs to write the quality criteria for each indicator in the framework and compare and contrast these criteria with those drafted based on the scholarly literature.

The quality criteria writing was scaffolded as follows to ensure that the quality criteria reflected increasing levels of expertise as guided by the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986). Before engaging in the activity, the SMEs were introduced to rules and techniques for writing quality criteria (Griffin et al., 2004) and were trained to use the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) to describe what it means to progress along the continuum of increasing competence. Some quality criteria extracted from the Vietnamese primary school teacher standards (Griffin et al., 2004, 2006) and the developmental learning framework for Victorian school leaders (Department of Education Victoria, 2007) were used as examples. The SMEs were given time for analysing and discussing the examples alongside the document summarising the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986).

After that, the SMEs worked in five groups of three participants. The groups were formed based on two criteria: (1) each group included at least one listening and one reading test item writer; (2) each group had three SMEs of three different levels of experience in writing items for VSTEP® ranging from low to high based on their length of experience in writing items for VSTEP®. Each group was responsible for writing the sets of quality criteria for the indicators for one competency. In their allocated groups, the SMEs were instructed to discuss and brainstorm what they would typically see an experienced item writer do, say, write, and create when performing each task and how that would be different to what would be done by someone who was new and/or was developing expertise in this area. The SMEs were reminded

to arrange the quality criteria for the indicators allocated to them along progressing continua of expertise. Each group was given A0-sized sheets of paper and markers to keep records of the ideas generated within their group. When they finished the activity, the researcher put the groups' poster notes on the wall and asked the SMEs to engage in the 'gallery walk' activity. They visited each other's posters, marked the statements they agreed or disagreed with, made comments, additions and deletions to the original content using different coloured markers.

After the poster display, the researcher distributed to the relevant group the performance indicators and quality criteria for each competency as cut-outs from the draft competency framework based on the review of the literature (see Table 5.3 as an extract). The SMEs compared and contrasted the two versions and identified the similarities and discrepancies. The researcher's tasks were to answer questions on the draft version, remind the participants of important theoretical and contextual considerations, and use these to resolve any disagreements. Then, the SMEs were engaged in the discussion on the content of the criteria that describe varying levels of performance quality for each indicator, focusing on the discrepancies between the criteria for each indicator written by the SMEs and those derived from the literature. All the recommended changes to the quality criteria were noted in the electronic version of the draft simultaneously as the discussions proceeded. The amendments, deletions and additions recommended by the SMEs and the reasons for their suggestions are summarised in Appendix F.2. In this appendix, all the quality criteria were arranged in the order that the SMEs agreed upon.

5.1.1.1.4 Expert review.

A round of expert review was conducted in order to seek consultation on improving the conciseness of and to confirm the ordering of the quality criteria for each performance indicator. The expert panel included six members. Three of them were identified as SMEs who were English teachers at three leading foreign language universities in the southern and the central

regions of Vietnam and had been engaged in writing items for VSTEP® reading and writing skill tests for approximately two years. These SMEs held doctoral degrees, had twelve to eighteen years of experience in teaching English at university level and previous research experience in developing assessment instruments to measure capabilities and skills. They worked outside the researched institution and did not participate in either the focus groups or workshops. The other three participants were research fellows at an Australian tertiary institution whose areas of interest and expertise was in the development, validation, and evaluation of assessment materials and instruments to assess capabilities and skills. Two of them had PhD degrees in education and the other had a PhD in assessment and evaluation. These research fellows also taught undergraduate and graduate courses in testing and assessment and were experienced in writing items for high-stakes tests in Australia, training item writers and developing instruments to measure skills and capabilities.

The participants were contacted, briefed about the research project via email and invited to participate in the panel of experts to review the competency framework. When they consented to participate, the competency framework derived from the literature review and examined and consolidated at the focus groups with the item writers and workshops with SMEs in Vietnam was emailed to them. They were requested to focus on ordering the quality criteria in terms of increasing proficiency and on making comments on whether these criteria were ordered in compliance with Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) and the guiding principles for writing quality criteria described by Griffin et al. (2004). They were also asked to check the comprehensiveness and appropriateness of the constructs and indicators identified and to check for language issues and made specific suggestions on wording, expressions, and the use of action verbs for describing increasingly higher-order thinking skills. This was meant to lead to improved clarity and conciseness both in terms of the communication of ideas and the implications for developmental learning underpinning the items. The panellists

used track changes to make comments on the document and sent it back to the researcher via email. After that, online meetings were held between the researcher and the Vietnamese panellists and face-to-face meetings were organised between the researcher and the Australian panellists. During the meetings, the panellists elaborated on the comments that they made in the written feedback, explained the reasons for their comments and proposed their suggestions. The amendments made to the wording and ordering of the quality criteria in the framework are summarised in Appendix F.3 and Appendix F.4. The final draft the framework is in Appendix G.

5.1.1.2 The IWCOMP Scale

The 14 performance indicators and 45 quality criteria in the final competency framework, that were derived from the scholarly literature and validated through the focus group discussions, workshops with SMEs and expert review, were converted into 14 multiple-choice items forming the IWCOMP scale. The scale was to examine the actual item writing practice of individual teachers with varying levels of expertise, and thus enable Vietnamese EFL teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests. Table 5.5 below displays the items within the IWCOMP scale.

Table 5.5 *Items within the IWCOMP Scale*

The Exploratory Phase	
Competency 1: Analysing test specifications	
Items (i.e., Performance indicators)	***Item Categories (i.e., Quality criteria)***
1. Review test specifications	1.0 Check aspects of the test specifications (e.g., the skills/subskills to be tested, text length, text type, item formats, quantity, and item difficulty levels) 1.1 Retrieve from prior experience important aspects of the test to be written. 1.2 Consider operational implications of the test specifications with the target examinees in mind
2. Utilise test specifications	2.0 Follow the instructions from the test specifications 2.1 Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications 2.2 Use experience and available resources to internalise the test specifications 2.3 Develop own internal guideline for item writing in accordance with the given test specifications and informed by other training materials/resources and prior experiences
Competency 2: Generating materials for tests	
Items (i.e., Performance indicators)	***Item Categories (i.e., Quality criteria)***
3. Search for input materials for tests	3.0 Browse materials of required length on recommended topics 3.1 Identify texts of specified complexity and difficulty from newspaper, magazines and books written for non-native English readers 3.2 Locate texts from previously used high-yield sources 3.3 Source texts from a diverse range of resources with consideration of the target examinees
4. Evaluate the materials for text-fit-specifications	4.0 Use text analysis tools (e.g., English Profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine text-fit-specifications in terms of difficulty and complexity 4.1 Assess texts based on defined characteristics of text contents and difficulty levels 4.2 Use multiple sources of evidence regarding text density, lexical difficulty, text potentiality for item development to judge the relevance of input materials
5. Make textual changes to the selected input materials for tests	5.0 Restructure and cut texts to required length 5.1 Make lexical and syntactic changes to achieve the target register 5.2 Employ an extensive repertoire of strategies to adapt the original text to meet the required textual features in terms of language use, style and their rhetorical functions

Items (i.e., Performance indicators)	*Item Categories (i.e., Quality criteria)*
6. Adapt texts for texts-fit-items and items-fit-texts	6.0 Design items closely adhering to the generated text/script 6.1 Use research- grounded ruses (e.g., use of paraphrase, antonyms and synonyms to avoid dictation effects, add distraction to text/scripts to reduce guessability, word-form shift, secondary text supplement) to adjust texts to fit items 6.2 Call upon a range of strategies to transform source texts to fit specified test purposes and test item requirements
The Concerted Phase	
Competency 3: Writing test items	
Items (i.e., Performance indicators)	*Item Categories (i.e., Quality criteria)*
7. Determine the approach to item writing tasks	7.0 Follow a standardised text/script-first item writing process 7.1 Focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items 7.2 Multi-task between text/script adaptation and item construction
8. Write item stems	8.0 Replicate items from sample tests provided in the test specifications 8.1 Construct items from a pool of item shells/item stems derived from established items in past test materials 8.2 Rely on different sources to create items accommodating test specifications and input materials 8.3 Build on various resources to innovate items
9. Write item distractors	9.0 Follow the format of the distractors given in the test specifications or sample tests 9.1 Write distractors with reference to guidelines on Dos and Don'ts for writing test item options 9.2 Utilise a range of paraphrasing strategies to construct distractors 9.3 Use a flexible repertoire of strategies to write distractors with consideration of typical errors or misconceptions of the subject content by test takers
10. Provide answer keys	10.0 Assign an answer key for each item 10.1 Provide a reference for each answer key 10.2 Justify correct and incorrect answer options based on the evidence from text/script
11. Decide on item difficulty	11.0 Align items with the descriptors defined in the item blueprint and/or in the six-level English language proficiency framework for Vietnamese learners 11.1 Consider the text complexity and cognitive demand/sub-skills required to answer each item 11.2 Draw on a range of factors affecting item difficulty from experience, expert colleagues and contemporary research to make judgements of item difficulty

Competency 4: Reviewing test items	
Items (i.e., Performance indicators)	***Item Categories (i.e., Quality criteria)***
12. Check item content fit	12.0 Examine the conformity of the items to the test specifications regarding item content, wording of items and instructions 12.1 Consider abilities required to complete the items to compare what the items are thought to be testing with what is claimed in the test specifications 12.2 Attempt items on own to make judgements of item-fits-specifications for content complexity and difficulty levels
13. Verify answer keys and distractors	13.0 Try out items on own to check the original keys 13.1 Analyse responses from a self-conducted pilot with a sample of students to identify any miskeys and/or poor performing distractors
The Refining Phase	
Competency 5: Revising test items	
Items (i.e., Performance indicators)	***Item Categories (i.e., Quality criteria)***
14. Respond to item feedback	14.0 Incorporate received feedback 14.1 Consider the test purposes to make decisions regarding implementing feedback 14.2 Use information from experience and research to justify decision to implement or not to implement received feedback

5.1.2 The Adaptation of the Professional Learning Scales (the PROLEARN and INTRAIN scales)

As discussed in Chapter Four, the analysis of the literature in expertise development and teacher competence growth provided an insight into potential factors affecting teacher expertise development in test item writing. The primary factors identified as influencing their development of expertise were elements of teachers' professional learning, which was further categorised into teacher engagement in self-regulated professional learning and institutional induction and training. The other contributory factors were teachers' professional and educational background characteristics. Existing conceptual models and measures of these two constructs were extensively reviewed to identify those that could be adapted to the research domain under this study (see Chapter 4, Section 4.1).

The reviewed literature provided the conceptual models and measures used in previous research that were suitable to adapt and construct items for the PROLEARN and INTRAIN scales. The PROLEARN scale was adapted from the Teacher Professional Development at Work (TPD@ Work) survey instrument developed by Evers et al. (2016) to assess teacher participation in four categories of professional learning put forward by Kwakman (2003) (see Table 4.1). The INTRAIN scale was adapted from part of the RITCME framework used by Cambridge Assessment English to recruit, induction, train, co-ordinate, monitor and evaluate test item writers reported by Ingham (2008) (see Table 4.2 for the training activities for Cambridge Assessment English reading and listening skill test item writers). The INTRAIN scale was also adapted to address five critical features of professional training programs identified by Desimone (2009) and Darling-Hammond and Richardson (2009) (see Section 4.1.2.2) and the indicators for high-quality professional training in the HQPD checklist developed by Erickson et al., (2017) (see Table 4.3). These original measures were chosen because they were the measures of teachers' professional learning at individual and institutional levels available in the literature at the time the research was conducted. In addition, these measures and their underlying conceptual models had been empirically validated and confirmed as sufficiently reliable through exploratory and confirmatory factor analysis (see Section 4.1.2).

As for the adaptation, the categorical models informing the original measures and their response scales were retained. The indicative activities for teacher professional learning at individual and institutional levels were adapted to reflect those targeted for developing Vietnamese EFL teacher expertise in multiple-choice test item writing. These indicative activities were produced in light of the reviewed literature and the focus group discussions with the item writer and executive participants (see Table 5.4 for the participants' information). These focus groups provided the information on the professional learning activities that the item writers were engaged in and were provided with to support their expertise development in

multiple-choice test item writing. This information in combination with the findings from the related reviewed literature informed the adaptation of the items for the PROLEARN and INTRAIN scales. The items within these scales were subsequently reviewed by a panel of experts (see Section 5.1.1.1.4 for the participants' background information). The panellists were requested to check the appropriateness of the items and underlying constructs and to check for conciseness and consistency of the items. The details of the PROLEARN and INTRAIN scales are presented in the next subsections.

5.1.2.1 The PROLEARN Scale

The PROLEARN scale consisted of 15 items. Ten of the items were adapted from 10 out of 21 items in the TPD @ Work survey by Evers et al. (2016). Of the other 11 items in the TPD @ Work survey, seven items (i.e., item 6, 7, 9, 12, 14 and 16) were not used due to their irrelevance to the confidential nature of test item writers' job and the other items (i.e., item 18, 19, 20 and 21) were related to teacher collaboration to improve their school, that was not pertinent to the subject of this research. The other five items of the PROLEARN scale (i.e., item 6, 7, 9, 10 and 15) were developed to address the individual professional learning activities at the lowest and highest levels of engagement as informed at the focus groups with the item writers and executives in Vietnam. The scale was designed to measure the frequencies of teachers' engagement in four categories of professional learning: keeping up to date, experimenting, reflection, and collaboration (Kwakman, 2003). As discussed in Chapter 4, this four-factor structure was confirmed through exploratory and confirmatory factor analysis. The reliabilities of the subscales were sufficient to good (Cronbach's α ranged from .67 - .80, Evers et al., 2016, p. 172) (see Section 4.1.2.1). The informants were required to indicate the engagement using a 4-point rating scale: 1= never, 2=rarely, 3= sometimes, and 4= often. Table 5.6 presents the four categories of professional learning, the original items in the TPD @ Work survey and associated

item numbers in the PROLEARN scale. Table 5.7 presents the items within the PROLEARN scale.

Table 5.6 *Categories of Professional Learning and Associated Items in the TPD @ Work Survey Instrument and in the PROLEARN Scale*

Categories of professional learning proposed by Kwakman (2003)	Items in the TPD @ Work survey by Evers et al. (2016, p. 170-171)	Item number in the PROLEARN scale
Keeping up to date: Reading & Work-related training	1.Studying subject matter literature	3
	2.Reading educational/subject matter pedagogical literature	1
	3. Visiting educational sites on the Internet	2
	4. Participating in a training course that centre around subject matter pedagogy	4
	5. Participating in one day conference or study day that centres around subject matter pedagogy	5
Experimenting	6. Trying out new applications in ICT in my lesson	10
	7. Testing alternative teaching materials in class	
	8. Applying and evaluating other forms of assessments	11
	9. Inquiring into new teaching methods in class	
	10. Trying out new teaching methods in my lesson	12
Reflecting and asking for feedback	11. Reflecting on my strong and weak points	8
	12. Adapting my teaching methods in response to pupils' reactions	
	13. Asking pupils for feedback on the way I teach	
	14. Inviting colleagues to attend my lesson	
Collaborating	15. Preparing lessons with colleagues	14
	16. Making agreements with colleagues about the pedagogical practices	
	17. Discussing the teaching approaches I use in class with colleagues	13
	18. Giving an opinion together with colleagues about school organisational matters to the school management	
	19. Assembling a school working group or committee with colleagues	
	20. Thinking about the design and method of pupil counselling with colleagues	
	21. Discussing ideas about educational improvement and innovation in my school with colleagues	

Table 5.7 *Items within the PROLEARN Scale*

	Items
1.	I study guidelines provided to item writers.
2.	I read item writing manuals from online resources.
3.	I read research publications on multiple-choice test item writing.
4.	I participate in compulsory training workshops organised by my institutions for teacher item writers.
5.	I attend professional learning opportunities outside my institutions.
6.	I conduct/co-conduct training sessions for item writers.
7.	I tend to write items in a similar way to how I have written them in the past.
8.	I reflect on/recall prior item writing experiences to suggest lessons learnt for future tasks.
9.	I record/document my own best practices in multiple-choice test item writing.
10.	I utilise my previously used strategies for multiple-choice test item writing.
11.	I adjust my test item writing strategies and approaches based on lessons learnt from past experiences.
12.	I experiment with new strategies and approaches to multiple-choice test item writing.
13.	I talk about problems with multiple-choice writing test items with colleagues.
14.	I share resources and ideas about multiple-choice test item writing with colleagues.
15.	I support or coach colleagues in multiple-choice test item writing

5.1.2.2 The INTRAIN Scale

The INTRAIN scale contained 21 items related to the induction and training organised by the institution in supporting teachers' professional learning in the domain of multiple-choice test item writing. Respondents were required to indicate the presence of the institutional induction and training activities using a 2-point rating scale '0= No, 1= Yes'. The 21 items in the INTRAIN scale were constructed based on the 11 training activities for Cambridge Assessment English reading and listening test item writers (Ingham, 2008) and other indicative activities that were pertinent to the context of item writer training in Vietnam as informed by the focus group participants. These items were adapted in alignment with the five critical features of professional

training (Darling-Harmond & Richardson, 2009; Desimone, 2009) and 12 out of 22 indicators for high-quality professional training in the HQPD Checklist (Erickson et al., 2017). The 10 indicators in the HQPD Checklist (i.e., indicator 4, 5, 8, 9 and 17-22) were not used because these are related to either a specific school context or long-term follow-up training activities. Table 5.8 displays the core features of professional training, the associated items in the original HQPD Checklist, in the list of induction and training activities for item writers at Cambridge Assessment English and in the INTRAIN scale. Table 5.9 presents the items arranged in order within The INTRAIN scale.

Table 5.8 *The Core Features of Professional Training and Associated Items in the Original Instruments and in the INTRAIN Scale*

Critical features of professional training (Darling-Harmond & Richardson, 2009; Desimone, 2009)	Original items in the HQPD Checklist (Erickson et al., 2017)	Induction and training activities at Cambridge Assessment English (Ingham, 2008)	Items in the INTRAIN scale
		Induction	
		1. Provide background information on the test production process and documentation on what is expected of Cambridge Assessment English item writers	1. I am provided with a documentation of the task description. 2. I am provided with sample tests. 3. I am provided with additional guidelines for item writers. 4. I am briefed about the test specifications. 5. I have opportunities to discuss the test specifications with colleagues for further clarification.
		Pre-training	
	1. Provide a description of the training with learning objectives prior to training 2. Provide readings, activities and/or questions to think about prior to the training 3. Provide an agenda (i.e. schedule of topics to be presented and times) before or at the beginning of training	2. Provide a pre-training pack including tasks relating to text selection and adaptation and a list of typical features of spoken English with examples	6. I am provided with a training agenda (i.e., schedule of topics to be presented and times) before the training. 7. I am provided with a description of the training with learning objectives prior to training 8. I am provided with readings, activities and/or questions to think about prior to the training.

Critical features of professional training (Darling-Harmond & Richardson, 2009; Desimone, 2009)	Original items in the HQPD Checklist (Erickson et al., 2017)	Induction and training activities at Cambridge Assessment English (Ingham, 2008)	Items in the INTRAIN scale
		Training	
Content focus	6. Include the empirical research foundation of the content (e.g. citations, key researchers) 7. Content builds on or relates to participants' previous professional development 10. Build shared vocabulary required to implement and sustain the practice	3. Introduce typical test tasks and the basic terminology used to describe test questions 4. Provide training on techniques of writing multiple-choice test based on input from the trainer and group activities drawing on the ideas and experience of the participants 5. Provide specific training on adapting written sources, selecting and sourcing texts, and approaches to item writing	9. I am trained about the principles of multiple-choice test item writing. 10. I have opportunities to engage in discussions with colleagues on multiple-choice item writing techniques. 18. I have opportunities to learn about basic measurement characteristics of test items (e.g., item difficulty, and item discrimination indices). 19. I am trained about using different tools and resources (e.g., CEFR, text analysis tools) to support my item writing.
Active, hands-on learning		6. Consider the strengths and drawbacks of multiple-choice as a test type and to identify the features of sound multiple-choice items	11. I have opportunities to engage in discussions with colleagues on item writing techniques. 14. I have opportunities to work with colleagues to identify features of sound items.
Collective participation	11. Provide examples of the content/practice in use 12. Illustrate the applicability of the material, knowledge or practice to the participants' context	7. Consider and discuss in groups examples of multiple-choice items 8. Consider the most recent live test paper, discuss and give feedback on how well the item had performed.	12.I am provided with example test items to analyse. 13. I have opportunities to consider the strengths and weaknesses of the items with colleagues. 16. I have opportunities to analyse recent live test papers. 17. I have opportunities to engage in feedback on live test item performance with colleagues.

Critical features of professional training (Darling-Harmond & Richardson, 2009; Desimone, 2009)	Original items in the HQPD Checklist (Erickson et al., 2017)	Induction and training activities at Cambridge Assessment English (Ingham, 2008)	Items in the INTRAIN scale
Application to practice, reflection on results and collaboration with colleagues	13. Include opportunities for participants to practice and/or rehearse new skills (Engagement) 14. Include opportunities for participants to express personal perspectives (e.g. experience, thoughts) 15. Include opportunities for participants to interact with each other related to training content	9. Work in groups of new and experienced item writers on pre-selected a range of texts from different sources and discuss the best texts to use in terms of topic, content and style of writing and practice writing items for the selected text 10. Work in groups to revise and adapt sample texts	15. I have opportunities to engage in hands-on item writing practice tasks (e.g., select and adapt texts, write item stems, write distractors, estimating cognitive demands for answering each item …) with colleague item writers of varying experience.
Duration and quality of trainers	16. Adhere to agenda and time constraints	11. Offer a training weekend with Cambridge Assessment English item writer trainers	20. I am trained by relevant trainers/have opportunities to work with and learn from experienced/ well-qualified trainers. 21. I have opportunities to participate in training and development workshops for teacher item writers 1-2 working weeks a year.

Table 5.9 *Items within the INTRAIN Scale*

Item
1. I am provided with a documentation of the task description.
2. I am provided with sample tests.
3. I am provided with additional guidelines for item writers.
4. I am briefed about the test specifications.
5. I have opportunities to discuss the test specifications with colleagues for further clarification.
6. I am provided with a training agenda (i.e., schedule of topics to be presented and times) before the training.
7. I am provided with a description of the training with learning objectives prior to training.
8. I am provided with readings, activities and/or questions to think about prior to the training.
9. I am trained about the principles of test item writing.
10. I am trained about techniques for writing different objective test item types.
11. I have opportunities to engage in discussions with colleagues on item writing techniques.
12. I am provided with example test items to analyse.
13. I have opportunities to consider the strengths and weaknesses of the items with colleagues.
14. I have opportunities to work with colleagues to identify features of sound items.
15. I have opportunities to engage in hands-on item writing practice tasks (e.g., select and adapt texts, write item stems, write distractors, estimating cognitive demands for answering each item …) with colleague item writers of varying experience.
16. I have opportunities to analyse recent live test papers.
17. I have opportunities to engage in feedback on live test item performance with colleagues.
18. I have opportunities to learn about basic measurement characteristics of test items (e.g., item difficulty, and item discrimination indices).
19. I am trained about using different tools and resources (e.g., CEFR, text analysis tools) to support my item writing.
20. I am trained by relevant trainers/have opportunities to work with and learn from experienced/ well-qualified trainers.
21. I have opportunities to participate in training and development workshops for teacher item writers 1-2 working weeks a year.

5.1.3 Professional and Educational Background Characteristics

The literature review on professional and educational background characteristics that affect teacher expertise development (see Sections 4.2 and 4.3) and the findings from the focus groups with the item writers and the executives informed the inclusion of the following characteristics hypothesised as factors contributing to Vietnamese EFL teachers' judgements of their expertise in multiple-choice test item writing. These were teacher qualifications, formal training in TESOL/TEFL and applied linguistics, and in testing and assessment, experience in

teaching English language skills, teaching exam preparation courses, constructing English tests from published resources and writing VSTEP® test items and teachers' English language proficiency. Ten items were constructed to elicit the participants' professional and educational background information and were included in the questionnaire. Table 5.10 includes background information sought about the participants.

Table 5.10 *Professional and Educational Background Information Sought about the Participants*

Background characteristics	Variable name	Data sought	Measure level	Values
Highest academic qualification held	HAQ	Bachelor/ Master/Doctor	Nominal	0 = BA 1 = MA 2 = PhD
Highest degree major	MAJ	TESOL/TEFL Applied Linguistics (AP) Education (E) Testing and Assessment (T&A) Others	Nominal	0 =TESOL 1= AP; 2= E 3 = T&A 4 = Others (Recoded into 0 = Non-TESOL and 1= TESOL
Years of teaching English language skills	TEACHEXP	Numbers of years	Continuous	0 -30 years
Years of teaching EFL exam preparation courses	EXAMEXP	Numbers of years	Continuous	0 -30 years
Years of constructing English tests from published resources	TESTEXP	Numbers of years	Continuous	0 -30 years
Number of VSTEP test writing commissions	VSTEP	State the number of VSTEP tests written	Continuous	1 – 150 tests
Uptake of language testing and assessment course	LTA	Yes/No	Nominal	0 = No 1 = Yes
Levels of the course's usefulness for multiple-choice test item writing	USEF	Very useful Useful Somewhat useful Not useful at all	Nominal	0= Very useful 1= Useful 2 = Somewhat useful 3= Not useful at all
English language proficiency levels	ELP	Self-report IELTS bands or equivalents	Ordinal	5.0 - 9.0

The recorded variables were used to examine the correlations of professional background characteristics and the participants' competence scores. In all analyses, the original variables were used except for the highest degree major (MAJ) which was recoded into Non-TESOL and TESOL.

The items from the three scales along with the questions about the teachers' professional and educational background information were subsequently collated into a questionnaire. Ten questionnaires were distributed for trialling among EFL teachers who engaged in writing items for VSTEP® reading and listening skill tests in two foreign language universities in Ha Noi and Da Nang, Vietnam before the official online survey administration. The final version of the questionnaire is included in Appendix D.

5.2 Data Collection Procedures

5.2.1 The Participants and the Sampling Framework

The desired population was defined as EFL teachers at universities across Vietnam who had been commissioned to write items for VSTEP® (Vietnamese Standardised Tests of English Proficiency). As VSTEP® had been launched in 2015 after one and a half years from its creation and development, this indicated that EFL teachers in Vietnam had engaged in multiple-choice test item writing for large-scale English tests for 3-5 years at the time the questionnaire was conducted. The presence of VSTEP® marked the first attempt ever recorded in Vietnam for EFL teachers to write items for tests, not copy items from test books, which was presumably their common test writing practice for both formative and summative assessment. Thus, the teachers who had experience and/or training in VSTEP® test item writing were deliberately targeted. A combination of snowball, and criterion recruitment strategies was used to select a sufficient number of teachers for participation in the survey. First, a criterion was set that the teacher participants had constructed items for VSTEP®. The number of test items the teachers had

written, and their professional characteristics did not matter, as the researcher intended to choose participants of various experience and expertise. These teachers often gathered in Ha Noi for professional development and training for item writers during their summer break. Second, the contacts were made with the university rectors or vice rectors and/or university testing centre directors, who would suggest the names and contact details of VSTEP® item writers. The procedure proposed by Warren (2001) was also followed: "sampling begins with acquaintances and moves to strangers" (p.88). Potential participants were suggested by the university leaders and testing centre directors and identified by the researcher herself. As the researcher had been involved in a number of government-funded projects as part of the Project 2020, she worked with many teachers from different foreign language universities across the North, Central and South of Vietnam. Potential participants were contacted; in turn, they suggested other teachers to contact. Third, as the targeted population was limited, all the available participants who had experience with VSTEP® item writing and agreed to complete the questionnaire were recruited. Table 5.11 provides the total numbers of the research participants by location.

Table 5.11 *Summary of Survey Participants from Major Universities by Location*

Location	Total
Ha Noi	43
Da Nang	22
Thai Nguyen	36
Thanh Hoa	11
Nha Trang	11
Son La	8
Other provinces & cities	19
Total	150

Survey sample size is one significant factor contributing to the validity of the study (Coates, 2005). The determination of sample size for the current study was informed by the

research aims and context, the analyses to be performed, statistical standards, and measurement considerations beside feasibility considerations.

The sample was chosen to be sufficiently large for psychometric calibration of the framework. Rasch item response modelling (Rasch, 1960; Wright & Masters, 1982) was the main psychometric method used, and partly formed the basis of decisions about the sample size. The Rasch model is less demanding on sample size than other item response models because it only estimates a single parameter per informant (Lord, 1983). It is suggested that a sample size of around 150 could generate reasonably precise parameter estimates for the items (Linacre, 1994b). An alternative heuristic practice is to gather around 30 responses for each performance level/quality criterion, which would indicate a sample size of around 120 (30*4 levels) item writers (Coates, 2005).

Additionally, as this study involved use of multiple regression, the cases-to-independent variables ratio had to be substantial. Hair, Black, Babin, and Anderson (2010) believe that the sample size directly affects the statistical power of multiple regression and the generalisability of the result. They suggest the desired ratio is between 15 and 20 observation for each independent variable, (thus, a sample size of 120 (15*8) or 160 (20*8) was deemed to be sufficient for the regression analysis within this proposed study. The authors contend that with 8 potential independent variables, the sample size of 150 is sufficient to detect the relationship with R^2 values of approximately 18% at a power, which is the probability of detecting a level of R^2 for a sample size as statistically significant of .80 with the significance level set at .01 and 13% with the significance level relaxed to .05 (Hair et al., 2010, p. 174)). Tabachnick and Fidell (2013, p. 123) recommended the simple formulas for testing the multiple correlation and for testing individual predictors which are $N \geq 50+8m$ (where m is the number of IVs) and $N \geq 104+m$ correspondingly. These rules assume a medium-size relationship between the IVs and the DV at $\alpha = .05$ and $\beta = .20$. This study hypothesised nine predictors, 50+8(9) = 122 cases to test

regression, and 104+9 = 113 cases for testing individual predictors would suffice. It was pertinent to the research interest to explore both the overall correlation and the unique contribution of each predictor variable, the larger number of cases was chosen (Tabachnick & Fidell, 2013). Thus, the sample size met both the guidelines from Tabachnick and Fidell (2013) and Hair et al. (2010), as it was sufficient to identify statistically significant relationships between teacher competence in multiple-choice test item writing and their professional learning and background characteristics and to ensure the generalisability of the findings to the entire Vietnamese population of EFL item writers, particularly when using a stepwise estimation procedure.

5.2.2 Response Rate

Emails with the link to the online questionnaires were sent to the entire population of 239 teachers. Completed responses were received from 150, yielding a return rate of 63%. This represented a volunteer population of EFL test item writers in Vietnam.

5.2.3 Questionnaire Administration

Emails were sent and/or telephone calls were made to the targeted teachers to introduce and briefly outline the research and invite them to participate in the research. Once the participants agreed to complete the survey, an email containing the link to the online questionnaire and instructions was sent to them. Following the ethics approval by the Humanities and Applied Sciences Human Ethics Sub-Committee of the University of Melbourne (see Appendix C), the online survey links included a plain language statement, a consent form, and the survey. The participants were instructed to check the boxes to express their consent to participate in the survey before proceeding to the survey. This data collection procedure was conducted at one institution at a time within a time frame of two weeks per institution, allowing the participants some flexibility in when and where they would like to complete the form.

Repeated e-mail-outs were done to ensure the highest possible response rate from every single participant.

The link containing the following materials was emailed to the participants:

- The Questionnaire (see Appendix D. The Questionnaire)
- The Plain Language Statement explaining the aims of the project (see Appendix A. Plain Language Statements)
- The Consent Form explaining the voluntary nature of participation (see Appendix B. Consent Forms)

The participants were requested to complete four parts of the questionnaire. For the first part, the participants were asked to self-report their test item writing competence by selecting the options that best described their typical behaviours in their item writing practice. In the second part, the participants rated their frequencies of engagement in four categories of self-regulated professional learning using a 4-point rating scale ‘never, rarely, sometimes, and often’. In the third part, the participants indicated if they were accessible to the induction and training opportunities offered by their institutions, using a 2-point Likert scale “Yes and No”. In the final part of the questionnaire, the participants were requested to supply educational and professional background information regarding their educational qualification, pre-service training, lengths of work-related experience and English language proficiency.

5.3 Data Analysis

5.3.1 Item Response Modelling

After the scales had been developed and administered via an online questionnaire, the data were exported into SPSS 25, screened for anomalies and prepared for analysis. Item response modelling programs were run to empirically calibrate each scale using the computer software program CONQUEST 4.0 developed by Adams et al. (2015). Calibrating is the process of determining “the accuracy of units on a common measurement scale”(Gillis, 2003, p. 105).

Calibrating items using item response models is applied to generate the relative difficulty of items and the relative person ability, and to position them on the same linear scale of measurement through certain item response theory (IRT) probabilistic models (Hambleton & Swaminathan, 1985). The application of IRT models in education has provided instrument developers with an extensive understanding of the strengths and weaknesses of test item construction (Carlson & Davier, 2013). By locating items in order of the cognitive demands or underlying skills required to answer each item correctly in relation to the person ability distribution, IRT models provide powerful tools to assess measurement properties of the scales and the ability range of the people who are assessed by the scales (Wu, 2003; Wu et al., 2016). This IRT modelling methodology has been used widely for the development of proficiency scales to measure cognitive abilities and competence (e.g.,Duong, 2011; Duong, Nguyen, & Griffin, 2011, 2019; Griffin, Nguyen, Gillis, & Mai, 2006; Lai, 1998; Woods, 2010; Wu, 2003). This methodology guided the development of the instruments measuring the three underlying constructs in this current study. An iterative process was applied in the calibration of the three scales in this study involving four steps: (1) the IRT program is run, (2) the model fit is checked, (3) the data or model is revised, and (4) the IRT program is re-run.

As previously indicated, the dependent variable, Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests (IWCOMP), and two of the independent variables, the teachers' engagement in self-regulated professional learning (PROLEARN) and their endorsement of their institutional induction and training (INTRAIN) were not directly measured. They were defined by a set of rubrics or indicators in the questionnaire. These indicators were measured using ordered categories. It was proposed that the Rasch's Simple Logistic Item Response Model for dichotomous items and the Partial Credit Model for polytomous items should be applied to calibrate the scales developed and adapted in this study using CONQUEST 4.0 (Adams et al., 2015).

Within the PROLEARN scale, the items were rated on a 4-point rating scale, but the response patterns were recoded as either ‘Rarely’ or ‘Often’ using a dichotomous score of one and zero correspondingly. The INTRAIN scale consisted of 21 dichotomous items and the response patterns were either ‘Yes’ or ‘No’. Therefore, the Rasch’s Simple Logistic Item Response model (hereafter referred to as the Rasch model) was applied to calibrate the scales, as opposed to a rating scale analysis to facilitate the interpretation of these variables (Wright & Masters, 1982).

The Rasch model is predicated upon the assumption that a latent trait or unobserved ability of a person can be predicted based on how well the person performs on items designed to measure that trait or ability. The Rasch model theorises that probability of an item response is dependent on the response patterns, not on which persons respond to the item. According to Rasch (1960), by assuming just one item difficulty parameter and one person ability parameter, the model stipulates that a person’s item response is merely determined by their ability to respond to the questions and not by other factors such as guessing, motivation or any personal characteristics. Mathematically, item difficulty is calculated as the ability of people whose chances of achieving the item has reached 0.5% probability. The result is joint calibration of people by ability and items by difficulty onto a common, interval-level scale, marked out by difficulty measures called logits, or ‘log of odds units’ (Wu et al., 2016, p. 110). Through this process, it is possible the estimates for the ability and difficulty parameter are simultaneously mapped on one scale which enables a substantive interpretation of the measurement scale obtained through the content analysis of clusters of items adjacent to the person’s position on the scale (Griffin, 1997).

There are two important features of the Rasch model. First, the model assumes that item/task difficulty is the only item/task attribute which influences that candidates’ responses, hence it is also named a one parameter logistic model (1PL).The Rasch model aims to

investigate the relation between "the demands of the tasks and the person's ability to perform them" (Griffin et al., 2007, p. 89) in the form of the person's probability of success on the tasks. Second, it posits that a person's latent ability is the point where the person has a 50% probability of getting an item of given difficulty correct and item difficulty is defined as the point where a person of an ability has a 50-50 percent chance of answering the item correctly (McNamara, 1996). If a person's ability is less than the item difficulty, then they will have less than 50% chance of responding correctly to the item. The Rasch model stipulates that people with higher capability have a higher probability of success on an item than those with low capability. According to Wu et al. (2016), the Rasch model is considered to have reliable measurement properties because "under the Rasch model, item difficulty order is well defined across the whole range of ability" (p.117).

The simple logistic Rasch measurement scaling approach to calibrating the PROLEARN scale meant that there was one engagement parameter θ_n for each person, and one level of engagement parameter δ_i for each item (Wright, 1977). According to the simple logistic model for dichotomously scored items by Rasch (1960), the probability of a specific item response does not rely on which person responded to the item, but rather on the response pattern. These parameters represented the position of persons and items on the latent variable they shared. Each of these parameters governed the likelihood of a teacher with a level of engagement, θ_n, selecting 'often' for the item (i.e., obtaining a score of 1 rather than 0). The analysis modelled the relationship between the teacher engagement and the level of engagement parameters of each item. The Rasch model was also applied to calibrate the 21 dichotomous items in the INTRAIN scale.

The IWCOMP scale comprised 14 items, each of which consisted of three to four categories describing varying levels of performance quality. The PCM was employed to calibrate this behaviourally anchored rating scale. As described by Masters (1982), the partial

credit model (PCM) is derived from the application of Rasch's (1960) dichotomous model to observations recorded in more than two adjacent response categories on an item given that these score categories should be ordered to reflect increasing proficiency. The PCM assumes that people with higher ability are more likely to be assigned to higher score categories for an item (Wu & Adams, 2007). The item step difficulties or delta ($\delta_{1, 2...}$) parameters generated from an item response theory analysis of PCM data represent the boundaries between adjacent score categories (Masters, 1982). They can be conceptualised as points along an ability continuum where the persons have equal chances of answering two adjacent categories correctly. Due to the polytomous nature of the items, the probability is not 0.5 anymore because the possibility is to be in score categories other than k-1 and k. As the delta (δ) parameters can be disordered, the Thurstonian thresholds (γ**)** can be used as indicators of item categories difficulties (Wu et al., 2016). The Thurstonian threshold is the point at which the probability of obtaining a score category or higher equals 0.50. They are cut-points for dividing the ability continuum into "score regions" (Wu et al., 2016, p. 171), reflecting the "cumulative achievement" of item categories (Wu et al., 2016, p. 170). For an item, the more score categories into which we can divide the participants to form ordered ability groups, the more discriminating the test will be, and the higher the test reliability will be. When the data fit the PCM model, the maximum scores of a PCM item correspond to the discriminating power of the item (Wu et al., 2016).

It is important to note that PCM is not a sequential step model (Wu & Adams, 2007). Thus, the PCM does not assume that a person must succeed in all lower score tasks to succeed in higher score tasks. This study chose to report the Thurstonian thresholds (γ) as indicators of item difficulty in the IWCOMP scale. For dichotomous items in the PROLEARN and INTRAIN scales, the Thurstonian thresholds are equal to the delta parameters (δ).

Item response modelling enables the establishment of the levels of person ability along a continuum; thus, it is frequently used to build developmental progressions using item-person

maps or Wright maps (Wilson, 2011). The spread of the items on the variable maps produced by CONQUEST 4.0 provides a vision of how well items are sufficiently spread out to distinguish levels along a variable. The Wright map has been extended containing partial-credit steps of increasing difficulty (Masters, 1982) where each step represents a threshold to a new ability level. The map often referred to as an item-threshold map indicates that it is the threshold difficulties or steps of the items that are ordered, not the items. Whether the instrument measures the person's ability range can be seen from a judgement of whether the range of item difficulty covers the range of the person's ability on the logit scale. In other words, it is necessary to check whether the items are spread out or cluster along the continuum of the latent trait. The clustering of items along the difficulty continua plays a crucial role in defining the constructs (Wright & Masters, 1982; Wright & Stone, 1979). The items should not only be spread, but those with the same cognitive/attitudinal demand should be close to one another on the continuum. Ideally, each cluster of items should convey a common theme of all the items in the cluster so that the cluster represents a proficiency level which forms the continuum. In this way, the clusters become useful for educational purposes because explication of the latent ability in this way assists identification of skills that persons are just beginning to acquire, where achievement likelihood has reached 50%. Once the continua formed by the items in measuring the three latent variables were found, the content analysis of the item clusters located along each scale was conducted. This analysis provided statements summarising the skills/characteristics of the items. These summary statements are descriptions of a progression for the population (Wu et al. 2016, p.117). Based on these descriptions, holistic interpretations of the teacher item writers' varying competence, their intensity of engagement in professional learning and attitudes toward the training and development opportunities their institutions offered to support their development of expertise could be made.

The next section casts more light on how empirical statistics under these models were step-by-step derived to generate reliability estimates of the scales and to analyse to what extent the scales designed could be used to make criterion-referenced interpretation (Glaser, 1981) of teachers' test item writing competence, engagement in and access to professional learning. Subsequently, the relationships between the dependent variable, IWCOMP, and the latent independent variables, PROLEARN and INTRAIN and other professional background variables were investigated using multiple linear regression.

5.3.1.1 Fit of the Model.

In order to evaluate the precision and accuracy of the scales, the measures employed to evaluate how well the scale was constructed were the standard error of measurement and the fit of the items to the Rasch model. The standard error of measurement (SE) supplies the information on the precision of the item difficulty/parameter estimates. Fit statistics, as Wright, Linacre, Gustafson, and Martin-Löf (1994) put it, describe the structural validity of a measure. It is suggested that the fit statistics including OUTFIT and INFIT mean squares statistics and t-statistics should be used to examine how the items fit the Rasch model (Bond & Fox, 2001; Boone, Staver, & Yale, 2014; de Ayala, 2009). These fit statistics are reported by CONQUEST 4.0.

Item FIT MSQ statistics are based on the difference between expected and observed scores or residuals which are standardised to obtain a Z-score by dividing the difference between the observed and expected scores by the standard deviation. The squares of each of the resultant values are distributed approximately like chi-square variates (Wu & Adams, 2007). The average of these squared values is called the OUTFIT known as the *"outlier sensitive mean squared residual goodness of fit statistic"* and the INFIT referred to as the *"information weighted mean squared residual goodness fit of statistic"*. The INFIT is easily affected by unexpected, small residuals (Linacre, 1994a), and both the INFIT and OUTFIT are sensitive to very large sample

sizes (Wright et al., 1994). Their values vary from zero to infinity. When the item fits the Rasch model, both INFIT and OUTFIT statistics are expected to approximate to 1. There are not fixed rules on the reasonable mean-square fit values (Wright & Linacre, 1994) but some acceptable ranges have been introduced in the Rasch literature. For example, de Ayal (2009) and Boone et al, (2014) recommended that 0.5-1.5 could be a productive range for measurement while Adams and Khoo (1996) postulated that the acceptable range of item mean-square fit statistics in practice in 0.77 to 1.3 for the sample size that is equal to 200 participants. Wright et al. (1994) suggest the reasonable ranges of INFIT and OUTFIT MNSQ values are 0.8-1.2 and 0.6-1.4 for high-stakes multiple-choice tests and Likert rating scale questionnaires respectively. The farther the fit statistics are from 1, the more different the items are from most others in measuring the construct.

T-statistics, which are the transformation of the fit statistics into normal deviates with a mean of zero and a standard deviation of 1, should be in the range of – 1.96 to +1.96 (Wu & Adams, 2007). Mean-square statistics are more affected by the sample size than t-statistics, because with larger samples, their variations from their mean (=1) become smaller and all of them become fit. It has been shown that the mean-square fit statistics become better (items measure the latent trait better) when the sample size is larger, while t-statistics are less sensitive to sample size. In general, the ranges for fit statistics are not absolute and the best strategy is to review both mean-square statistics and t-statistics when deciding the fit of items (Wilson, 2005)

Regarding confidence intervals and mean-square fit statistics, the acceptable range for the mean-square fit statistics given errors of measurement with a probability of 95%, is provided. When a mean-square statistic is the outside the range of the confidence intervals, the item is considered not to fit with the construct and the model. The confidence intervals were used to report item fit in this study alongside the recommended range of 0.77 - 1.3 by Adams and Khoo (1996).

The values of item mean-square fit statistics are reflected on the item characteristics curves. When these statistics are higher than 1, the items are said to be less discriminative than the model predicts, and the item characteristic curves for those items with mean-square fit statistics higher than 1 is flatter than the expected score curves, which means the persons with higher ability do not have a higher probability of being successful with the items. Meanwhile, the item characteristic curves for those items with mean-square fit statistics lower than 1 are steeper. The farther from 1 that mean-square fit statistics are, the more different the slope of the observed score curve is from that of the expected score curve.

5.3.1.2 Reliability and its Contributions to Validity Arguments.

The reliability of a scale is a measure of how well it can establish the capability of persons on the construct being measured (Wu et al., 2016). Consistency of the scale's items is the most important consideration because problematic items cause most response variation. To determine reliability, two approaches were followed in this study: firstly, the internal consistency of the items was established via the coefficient alpha named Cronbach's alpha (Cronbach, 1951) as recommended by CTT, and secondly, the degree of person-separation reliability was established via the Rasch separation index (Wright & Masters, 1982). These reliability indices were obtained through CONQUEST 4.0 (Adams et al., 2015). Cronbach's alpha is defined as the interrelatedness of a set of items within a scale (Cortina, 1993; Schmitt, 1996). This internal consistency reliability value increases as the number of items in the scale increases (Schmidt & Embretson, 2013). Wright and Masters (1982) claim that a scale is said to be more internally reliable if its Cronbach's alpha coefficient is closer to 1. There are no standard benchmarks for this value. Bond and Fox (2001) propose the lowest acceptable Cronbach's alpha coefficient of 0.8 while Nunnally (1978) suggests the level of Cronbach's alpha 0.7 as acceptable for newly developed instruments. Others (see Deković, Janssens, & Gerris, 1991; Holden, Fekken, &

Cotton, 1991; Myers & Oetzel, 2003) characterise the level of Cronbach's alpha for all scales as varying from .69 to .90 indicating moderate to high internal consistency reliabilities.

While the Cronbach's alpha offers the interpretation for the entire scale score (Wu & Adams, 2007), Rasch model (1980) analyses derive separation reliability indices for both modelled parameters (i.e., item and person). Analogous to Cronbach's alpha, the Rasch reliability estimates vary from 0 to 1. A separation index close to 1 indicates that items or persons were well separated along the variable. It also means that the errors of measurement are so small that there is little noise in the estimation of item and person parameters and their estimates are not overlapping (Griffin, 1997; Wright & Masters, 1982). Conversely, a reliability index of 0 indicates that the items or persons are not well separated, or the estimates have overlapped because the measurement errors are too large (Wright & Masters, 1982; Griffin, 1997). Examination of person separation reliability indices provides information about the extent to which persons are distributed along the continuum defined by the items on a scale. Where persons are clustered at extremes of a continuum, the difficulty levels of the items that defined the continuum are a poor match to the abilities of the persons being measured. Where persons are clustered too closely together, the ability to draw inferences about persons' competence from the scale is compromised. In combination with evidence of good fit to the Rasch (1980) model, which supports that the items measure a single, dominant construct, high separation reliability and the conceptual framework for the construction of a scale provide weight to the construct validity argument of the scale (Wright & Masters, 1982; Wright & Stone, 1979). Besides these, a comparison of descriptive statistics for item and person estimates can tell if the items are compatible for this group of people. If their means are close, and they are spread over a similar range, it is highly likely that they are compatible.

5.3.1.3 Summary Goodness-of-Fit.

In summary, the items that are performing well in a scale should possess the following desirable goodness-of-fit measures:

- The CTT discrimination index is above 0.2 (Linacre, 2007; Wu & Adams, 2007).
- The FIT MNSQ range 0.77 – 1.3 by Adams and Khoo (1996) and/or within the confidence intervals
- T statistics are within -1.96 to +1.96 (Wu & Adams, 2007)
- The average measure values should be increasing with increasing scores.
- The observed item characteristic curve is close to the theoretical one.

5.3.2 Multiple Regression Analysis

This study was interested in accurately predicting the competence levels of Vietnamese teachers in multiple-choice test item writing for English reading and listening skill tests to inform policy makers and practitioners of the factors contributing to the teachers' item writing competence. The main purpose was to predict the teachers' self-perceived item writing competence in multiple-choice test item writing for English reading and listening skill tests from a set of hypothesised predictor variables. These variables were selected based on rigorous prior research and a sound theoretical grounding (see Chapter 4) as delineated by Field (2018). They were teacher engagement in both informal and formal professional learning and their professional background characteristics. By doing so, the researcher aspired to identify the relative importance of each independent variable in the prediction of the dependent variable, and also to generate a regression model that best predicts Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. That model would identify the most significant factors that lead to increased competence in multiple-choice test item writing, and would inform intervention activities targeted at developing EFL

teacher item writers at tertiary institutions in Vietnam. Multiple regression appeared to be one multivariate technique that best suited the proposed research objectives. The multiple regression analysis predicts values of a DV from several IVs, known as predictor variables by weighing each IV to maximise prediction about the dependent variable (Field, 2018; Hair et al., 2010). The weights denote the relative predictive or explanatory power of the independent variables in the overall prediction and measure the influence of each IV on the DV. The weighted independent variables are arranged in a linear combination that best predicts the dependent variable. Based on the concept of minimising the sum of squared errors of prediction, the "best" independent variable can be selected based on the correlation coefficient; when the correlation coefficient is higher, the relationship is stronger, and the predictive power/accuracy the variable is greater (Hair et al., 2010; Tabachnick & Fidell, 2013). To apply the regression model, the researcher selected teacher competence in multiple-choice test item writing for English reading and listening skill tests as the dependent variable (Y) to be predicted by the following nine independent variables (see Table 5.12).

Table 5.12 *The Nine Predictor Variables*

X_1	Engagement in self-regulated professional learning (PROLEARN)	X_6	Length of experience in teaching English examination preparation courses (EXAMEXP)
X_2	Institutional induction and training (INTRAIN)	X_7	Length of experience in constructing English tests (TESTEXP)
X_3	Coursework training in language testing and assessment (LTA)	X_8	Deliberate practice in multiple-choice test item writing (VSTEP)
X_4	Highest degree major (MAJ)	X_9	English language proficiency (ELP)
X_5	Length of experience in teaching English language skills (TEACHEXP)		

The regression analysis predicts a teacher's IWCOMP score by multiplying the available IV scores by their regression coefficients and adding to the intercept, which is the value of

IWCOMP when all IVs are zero. The hypothesised relationship between the DV and the IVs in this study can be summarised in the following regression equation.

$$Y = A + B_1X_1 + B_2X_2 + B_3X_3 + B_4X_4 + \cdots + B_9X_9 + \text{SE}$$

Y = Predicted MCQ item writing competence score

X = Independent variable score

A = The value of Y when all Xs are zero/the intercept

B_1 to B_9 = the regression coefficients of the independent variables

SE = Standard error

The multiple regression estimation conducted for the current study followed the six steps below, which were adapted from the six-stage model building framework suggested by Hair et al. (2010). Each step is elaborated in the next subsections.

5.3.2.1 Tests of assumptions for individual variables.

The assumptions fundamental to multiple regression apply to the individual variables as well as the relationship between the variables as a whole. Because the variate generated from multiple regression analyses functions collectively in predicting the dependent variable, the first step is to assess the assumptions for both individual variables and the variate itself (Hair et.al, 2010). Thus, tests of assumption must be performed both prior to and after the regression model has been estimated. This step describes the assessment of assumptions for the individual variables involved in the regression analysis, which took place in the initial phase of the regression analysis. The individual variables entering multiple regression analysis are required to meet four assumptions: normality, linearity, homoscedasticity, and independence of the errors. To assess normality of the variables involved, both the graphical plots and statistical tests were employed, as recommended by Hair et al. (2010).

5.3.2.2 Estimating the regression model.

The second step is the estimation procedure. IBM SPSS 25 solves the mathematical equation used to estimate regression parameters. Three major types of estimation procedures in multiple regression are standard multiple regression, sequential multiple regression, and statistical multiple regression. In order to fulfil the research objectives, this study implemented a two-stage combinational approach to the regression estimation. First, the standard regression model was applied as a confirmatory approach. In this estimation process, the dependent variable, IWCOMP, and all the independent variables specified by the researcher (see Section 5.1) were simultaneously entered into the regression equation, and each of the IVs was assessed as if it had entered the regression after all other IVs had entered. Each IV is assessed in terms of what it contributes to prediction of the DV that is different from the predictability contributed by all the other IVs (Tabachnick & Fidell, 2013). This process enables the researcher to employ a confirmatory perspective in identifying factors (predictor variables) affecting teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests and specifying the predictive power of these variables which informed the rejection and acceptance of the hypotheses established. Secondly, the multiple regression analyses were run with different models in which IVs were removed from the equation one at a time to observe the changes of the regression variates. The typical utility of stepwise regression is that it could draw out the most important/useful subset of IVs that provide the maximal prediction of the DV, which aligns with the current research aims.

5.3.2.3 Tests of assumptions for the regression variate.

Once the variate is derived, the assumptions regarding linearity, independence, variance, and normality of the errors are examined. Both Hair et al. (2010) and Tabachnick and Fidell (2013) contend that the most widely used methods of assessing assumptions are analyses of residual plots, partial regression plots and normal probability plots. It is critical in regression

analysis that the correlation is based on a linear relationship**.** The linearity of the relationship between the dependent and independent variable in the post-regression analysis is examined through residual plots involving the residual versus the predicted dependent value. When this assumption of the linearity of relationship between dependent and independent variables is met, the null plot of residuals is observed.

5.3.2.4 Assessing the overall model fit.

The fourth step is the assessment of the model fit. This step entails the interpretation of the results from standard multiple regression given in the model summary including *R*, R^2, adjusted R^2, and the standard error of the estimate and the ANOVA tables. R is the value of the multiple correlation coefficient between the IVs and the DV. R provides a good estimate of the overall fit of the regression model. R^2 which is the correlation coefficient squared provides the size of the model fit because it denotes the proportion of total variation of the DV explained by the IVs included in the regression model. Cohen's (1988, 1992) benchmark norms for R^2 were adopted to test the overall model significance. As such, $R^2 = .02$ indicates a small size effect, $R^2 = .13$ (medium), and $R^2 = .26$ (large). An adjusted coefficient of determination (adjusted R^2) is also given; this measure shows how much variance in the DV would be accounted for if the model had been derived from the population from which the sample was taken. In other words, the adjusted R^2 denotes "the loss of predictive power or shrinkage" (Field, 2018, p. 312). Ideally, the adjusted R^2 should be similar or very close to the observed value of R^2 since this demonstrates the generalisability of the model. The adjusted R^2 is useful for examining the cross-validity of the model and how well the model generalises. For example, if the difference is .005 or 0.5%, it indicates that if the model was derived from a population instead of a sample, it would account for roughly 0.5% less variance in the outcome (Field, 2018). Regarding the model's generalisability, the confidence interval can be calculated given the standard error reported in SPSS 25. Standard error is the expected variation of the estimated coefficient due to

sampling error. The significance level (alpha) at a value of .05 was adopted; it denotes that there is a 5% chance of being inaccurate about if the estimated coefficient is different from zero. The confidence interval for the regression coefficient established at the significance level of .05 was calculated as ±1.96 × Standard Error, denoting that the coefficients will be estimated from 95% of repeated samples.

5.3.2.5 Interpreting the regression variate.

The fifth step is to interpret the individual regression coefficients. Parameter estimates in regression analysis are the unstandardised regression coefficients (*B* weights) which represent the type and strength of positive or negative relationship between a DV and each IV in the regression variate. The value of a B weight for a particular IV presents the change in the DV which results from one unit of change in the IV while all other IVs are held constant. These regression coefficients are the indicators of the IVs in their relationship with the DV. Beta (β) coefficient is the regression coefficients B which have been standardised into a common scale and variability before the regression equation is estimated. β coefficients indicate how the DV is relatively impacted by a change in one standard deviation in either variable. As all the variables now have a common unit of measurement, it is possible to determine which variable has the most impact. In other words, the β values allow for comparison between the effect of one IV on the DV and the effect of other IVs on the DV at each stage. Logically, if an IV significantly predicts a DV, its B weight and β is different from zero.

5.3.2.6 Validating the results.

The validation of the regression model is the final step to ensure that the results are not specific to the sample used in the estimation, but they can be generalisable to the population. According to Hair et al. (2010), testing the regression model on a new sample from the population is considered a sound empirical validation method. In the absence of such additional samples, other alternatives could be used. The first one is to assess the difference between R^2 and

adjusted R^2. Little or no difference between these coefficients indicates a lack of overfitting in the estimated regression model. The second one is to calculate the predicted residual sum of squares (PRESS) statistics of the original sample. PRESS statistics are a measure similar to R^2 and is employed to assess the predictive accuracy of the estimated regression model (Hair et al., 2010). Thus, PRESS statistics are used as a form of cross-validation in regression analysis to provide a summary of the model fit to a sample that was not used in the estimation of the model. The validation process involving calculation of PRESS statistics is said to be the leave-one-out validation procedure which deletes one observation in the model estimation, and then predicts the deleted observation with the estimated model; therefore, the observation does not impact the coefficient of the model used to calculate its predicted value. The process is iterated, deleting another observation, estimating a new model, and making prediction. The residuals for the observations are summed, so an overall measure of predictive fit can be provided. The PRESS statistics are not automatically provided by SPSS 25. However, since the deleted residuals can be saved as a variable in the SPSS regression procedure, they are squared and summed using a descriptive statistics procedure.

Figure 5.2 depicts the process of fitting and interpreting the regression model within this study. First, descriptive statistics, histograms, scatterplots and expected normality probability plots are produced to test whether the assumptions for the individual variables are met. Secondly, the standard multiple regression is run to generate initial diagnostic statistics. Thirdly, the overall model fit is assessed, and the residuals are examined for normality, linearity, homoscedasticity, and independence of errors. Supposing that all the assumptions are met, and no bias is found, the regression variate is interpreted and cross validated using adjusted R^2 and the predicted residual sum of squares (PRESS) statistics.

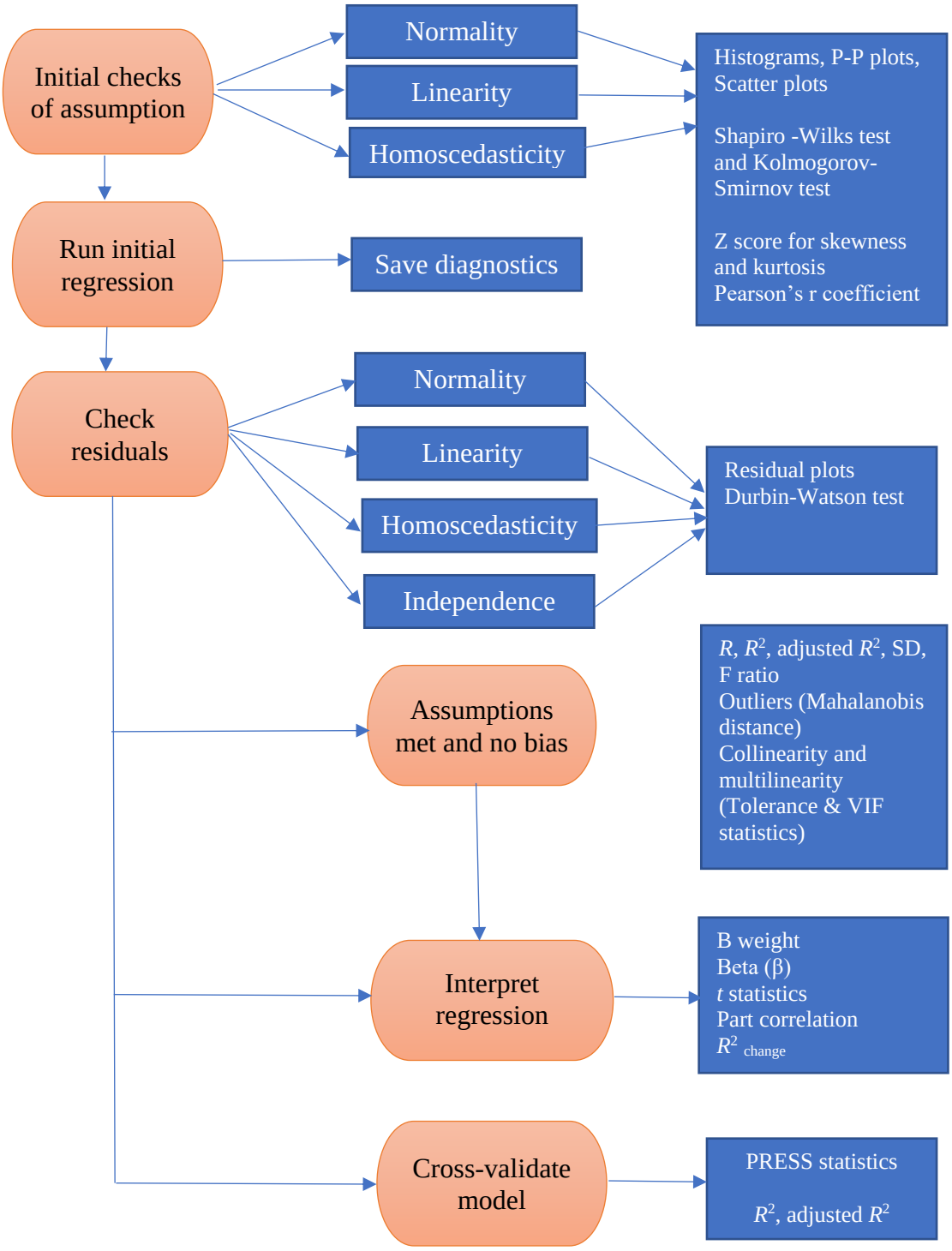


Figure 5.2 The process of fitting the regression model.

5.4 Chapter Summary

This chapter elaborated on the methodological rationales and operational processes involved in the conduct of this research. First, the scales to measure three latent variables, namely item writing competence, engagement in self-regulated professional learning, and access to institutional induction and training, were developed or adapted. The first scale to measure teacher test item writing competence was constructed based on extensive examination of the literature, focus group discussions with item writers, workshops with subject matter experts (SMEs) and expert review. The other two scales were adapted from prior research. The items from the scales and a set of background questions were later embedded into the teacher survey instrument. The data generated was used to empirically calibrate the scales designed using IRT modelling procedures. After that, multiple regression analyses were conducted to explore the best predictors for teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. In the next chapter, the calibration process of the three scales is presented in detail.

CHAPTER 6: SCALES CALIBRATION

This chapter provides an overview of the calibration of the three construct-measuring scales used in this study to explore the factors that were hypothesised to affect Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. The CONQUEST software version 4.0 (Adams et al., 2015) was employed to calibrate each of the scales. The analyses aimed to describe the measurement properties of each scale and levels of each of the constructs, to examine the performance of individual items of each scale, which sought to generate estimates of teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests, their engagement in self-regulated professional learning, and the access to induction and training. To start with, the calibration of each of the three latent variables is reported, followed by the discussion of the psychometric properties of each of the calibrated scales and how these inform scale item revision. The final section summarises the statistics of the scales.

6.1 Calibration of Variables

The three scales developed for the present study were "Multiple-choice item writing competence (IWCOMP)", "Engagement in self-regulated professional learning" (PROLEARN), and "Access to institutional induction and training for item writers (INTRAIN). The assumption underpinning the development and adaptation of these scales was that responses to the subsets of items were determined by the informants' positions on a series of latent continua. The following sections describe the calibration results of each item set for the three scales.

6.1.1 Calibration of the IWCOMP Scale

The IWCOMP scale, as presented in the previous chapter, was constructed based on the rubrics of the framework for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests developed through an extensive review of the scholarly literature on test development and item writing, focus groups with item writers,

workshops with SMEs, and expert reviews (see Chapter 5, Section 5.1.1). This 14 - item behaviourally anchored rating scale was developed for the participants to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests. The items used for Vietnamese EFL teachers to self-assess their test item writing competence and the levels of proficiency for these items are presented in Table 5.5.

Based on the results from the descriptive statistics of the scale, a total of 150 useable responses was received from 196 Vietnamese EFL teachers who participated in the self-reported assessment of their test item writing against the scale. Forty-six responses were rejected because these respondents completed only one or two questions in the survey, providing insufficient data for the analysis. A total of 150 cases was included in the IRT estimation procedure.

The results demonstrated that item 9 (i.e., *Write item distractors)* showed category disordering. According to Linacre (1999), category disordering occurs when the scores assigned to each item category do not correspond to the level of the latent ability. For example, persons with lower levels of item writing competence are likely to achieve higher category scores, and vice versa. Category disordering can be identified by examining the average measure value produced by CONQUEST 4.0. The average measure values should be increasing as item categories increase. The analysis showed that the item had disordered average mean ability at categories 0 and 1. Teachers who scored 1 with item 9 did not seem to have higher ability in test item writing than teachers who scored 0. It is notable that while only 10 per cent of teachers selected item category 1, the vast majority of the participants chose the other item categories *(*see Table 6.1 for the details of the item analysis). An examination of the content of the item categories 0 and 1 indicated that these categories described behaviours that require similar cognitive skills in accordance with the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986). The item category 0 "*Follow the format of the distractors given in the test specifications or sample tests*" demonstrates that the item writers mechanically make an exact

copy of the distractors in the sample given. The item category 1 *"Write distractors with reference to guidelines on Dos and Don'ts for writing test item options"* shows that the item writers follow the standardised guidelines and examples to produce distractors. It was indicated that these categories could be collapsed as they both referred to a similar behaviour of following the standardised guidelines and examples. The categories for this item were re-coded in accordance with the statistical indices. Table 6.1 shows item 9 originally had disordered mean ability at categories 0 and 1.

Table 6.1 *Details of Item 9's Analysis*

Item: 9 (9)								
Cases for this item 150			Item-Rest Cor. 0.32			Item-Total Cor. 0.50		
Item Threshold(s): -0.56 -0.27 0.85						Weighted MNSQ 1.05		
Item Delta(s): 0.71 -1.23 0.64								
Label	Score	Count	% of total	Pt Bis	t	sig	PV1Avg:1	PV1 SD:1
0	0	38	25.33	-0.16	-1.95	0.053	-0.128	0.430
1	1	16	10.67	-0.2	-2.48	0.014	-0.275	0.458
2	2	57	38	-0.05	-0.6	0.55	0.057	0.441
3	3	39	26	0.35	4.58	0	0.466	0.437

As illustrated in Table 6.2, the analysis for item 9 after collapsing category 0 and 1 showed that there was no more disordering in the average mean ability between the item categories.

Table 6.2 *Details of Item 9's Analysis after Recode*

Item: 9 (recoded)								
Cases for this item 150			Item-Rest Cor. 0.36			Item-Total Cor. 0.50		
Item Threshold(s): -0.41 0.91						Weighted MNSQ 0.96		
Item Delta(s): -0.09 0.60								
Label	Score	Count	% of tot	Pt Bis	t	sig	PV1Avg:1	PV1 SD:1
0	0	54	36	-0.27	-3.44	0.001	-0.129	0.538
1	1	57	38	-0.05	-0.6	0.550	0.037	0.462
2	2	39	26	0.35	4.58	0	0.311	0.526

The results of the calibration for the finalised 14-item IWCOMP scale generated from both Classical and Rasch analyses are presented in Table 6.3. For each item, the summary statistics are displayed as follows. The item Thurstonian thresholds (ɣ) which are reported in

Logit value followed by the standard error in relation to each item difficulty estimate (SE) are presented first. The Thurstonian thresholds for each item category in the scale illustrated in Table 6.3 show increasing levels of difficulty. Next, the item-rest correlation is reported, which refers to the relationship between the item score and the scale score after removing the item score from the total test score. Finally, the INFIT mean square (INFIT MNSQ), and t-statistics are presented, showing the extent to which the pattern of item responses fit the Rasch model.

Table 6.3 *Calibration Estimates for the Test Item Writing Competence Scale*

Item	Descriptions	γ_1	SE_1	γ_2	SE_2	γ_3	Item -rest corr	FIT	T
1	Review test specifications	0.82	0.28	1.18			0.26	1.01	0.1
2	Utilise test specifications	-0.24	0.22	0.22	0.28	0.71	0.40	0.98	-0.2
3	Search for input materials	-1.27	0.23	-0.19	0.27	0.21	0.24	1.11	1.2
4	Evaluate the input materials for text-fit-specifications	-0.67	0.18	0.42			0.30	0.99	-0.1
5	Make textual changes to the selected input materials for tests	-0.73	0.20	-0.08			0.46	0.89	-1.4
6	Adapt input materials for texts-fit-items and items-fit-texts	-1.00	0.17	0.81			0.41	0.93	-0.7
7	Determine the approach to item writing	-0.94	0.17	1.10			0.32	0.99	-0.1
8	Write item stems	-1.45	0.30	-0.87	0.28	1.35	0.35	0.96	-0.3
9	Write item distractors	-0.41	0.17	-0.91			0.36	0.96	-0.5
10	Provide answer keys	-0.40	0.18	0.62			0.12	1.15	1.9
11	Decide on item difficulty	-0.65	0.17	1.16			0.30	0.99	-0.1
12	Check item content fit	-0.52	0.17	1.25			0.25	1.02	0.2
13	Verify answer keys and distractors	-0.37					0.11	1.03	0.7
14	Respond to item feedback	-1.68	0.18	0.56			0.45	0.95	-0.4

The first aspect investigated was whether the performance indicators/items functioned together in measuring the construct of English test item writing. Table 6.3 demonstrates that the item threshold estimates increased as item categories increased. The threshold estimates of the item

categories varied from – 1.57 to + 2.02, with a range of 3.59 logits. The threshold estimates of the item categories are an indication of the cognitive demand of quality criteria. The range of these threshold estimates demonstrate that the levels of teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests were suitable to measure. The mean item difficulty was constrained to zero. The standard deviation of the item difficulty level was 0.40 with the relatively small standard errors of estimates for each of the items ranging from 0.08 to 0.16. The difference in item difficulty is revealed in the high item separation reliability (r = 0.92). As described in Chapter Five (see Section 5.3.1), this measure of IWCOMP scale reliability determines how sufficiently well separated items were in terms of their increasing levels of difficulty within the latent construct.

In regard to the weighted means-square fit statistics, all the items were within the confidence intervals ranging from 0.89 to 1.15, which showed that they were not significantly different from 1. The mean Item INFIT was 0.99, with a standard deviation of 0.06. The mean-square fit statistics of all the items indicated good discrimination. Item 10 (*Search for input test materials*) had the highest MNSQ fit of 1.15 but was within the confidence interval reported by CONQUEST. Adams and Khoo (1996) recommend a rule of thumb for acceptable ranges of INFIT and OUTFIT between 0.7 and 1.3. Similarly, de Ayala (2009) and Boone, as well as Staver, and Yale (2014) propose that a reasonable range for fit statistics is between 0.5 and 1.5. Wright and his colleagues (1994) suggest the reasonable ranges for INFIT and OUTFIT of multiple-choice high-stakes tests are between 0.8 and 1.2, while ranges of a rating scale (Likert/questionnaire) lie between 0.6 and 1.4. Thus, these recommendations confirm that all 14 items within the IWCOMP scale have an acceptable fit with the Rasch model and that there is evidence of a dominant underpinning construct in the variable being measured.

Besides providing item estimates and fit statistics, CONQUEST 4.0 (Adams et al., 2015) also reports classical statistics on item discrimination index, which furnish additional evidence

on the error in the responses to the items. This discrimination index is the correlation between a person's score on an item and their total score on the scale, which is regarded as a measure of the underlying ability. The correlation between the item score and the total score is high, indicating that a positive linear relationship exists between the person's item score and their ability. This study opts for the item-rest correlation, rather than item-total correlation. The item-rest correlation reflects the relationship between the item score and the scale score or the result reported after removing the item score from the total test score. This is a slightly more robust version of the item-total correlation because as explained by L. Le (2012) "total scores that have the item in question embedded within them will have a spuriously higher relationship than total scores made up of only the other items in the scale" (p. 1).The use of item-rest correlation is also advocated by Wu et al. (2016) as they posit that including the item itself in the total score could "inflate" the correlation between the two scores (p. 84). A rule of thumb is that item discrimination indices should be greater than 0.2, and those lower than 0.2 should be reconsidered (Wu & Adams, 2007). Poor item discrimination may result from poor formatting, and/or confusing wording, or that an item may be testing a different construct rather than the intended one (Wu et al., 2016). These factors should be checked when examining poorly discriminating items.

As can be seen from Table 6.3, the item-rest discrimination estimates for each item ranged from 0.11 to 0.46. According to Wu and Adams (2007) and Linacre (2007), as discussed above, item discrimination indices should be greater than 0.2. Those lower than 0.2 should be reconsidered. Accordingly, ten items in the test item writing competence generally had reasonably good discrimination within the acceptable range of 0.24 and 0.40 and a number of the items had very high discrimination indices greater than 0.4, such as item 6 (item-rest correlation = 0.41), and item 5 (0.46). Of the rest, item 13 (*Verify answer keys and distractors)* and item 10 (*Provide answer keys*) had low discrimination indices smaller than 0.2. The frequency statistics

for the categories of these partial credit items did not show why such a low statistic was acquired. This low discrimination may have resulted from the fact that these performance tasks were differently prescribed by different institutions. For example, for item 10 (*Provide answer keys*) some institutions may have required test item writers to provide a reference or justification for the answer keys, but others may not. As a result, test item writers would do as instructed; thus, their selection of the quality criteria may not have reflected their own cognitive capability. They may have tended to select the option that best described the way they were required to carry out this task. This was identified through the follow-up discussion with the participant who trialled out the scale as part of the questionnaire pilot process. Similarly, for item 13 (*Verify answer keys and distractors*), the matter of test security may be differently addressed across institutions. In some extreme cases, test item writers might have had no better alternatives than to try the test items on their own due to their commitment to test confidentiality. In some other scenarios, there might have been no specific restriction to what they can do to verify the test items they write. They may have conducted an unofficial pilot of the items as embedded items in a small-scale classroom-based test to verify their items. However, these items are essential components of the test item writing process, and the item categories/descriptors did reflect the increasing cognitive demand given their increasing item thresholds and mean abilities, so the items were retained.

The results for the item categories' fit come from the statistics for the interactions between items and categories (item*step) in the application of the unidimensional partial credit Rasch model. Similar to items, category mean-square fit statistics should be close to 1 and t-statistics should be close to 0. The results revealed that the item categories were significantly well fit to the model with all the categories' INFITs within the confidence interval, ranging from 0.90 to 1.15. This means the categories did seem to measure the same construct underlying the performance indicators/items.

The overall good fit statistics and the item-rest discrimination estimates of the items indicate that there is one dominant construct of multiple-choice test item writing for English reading and listening skill tests, and that the items work well together to measure the underlying construct. Overall, the items formed a cohesive scale representing "Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests".

The following item characteristics curves provide a visual aid to the fit and discrimination of items. Observed score curves are represented by dotted lines and the expected score curves are represented by straight lines. Item 6 *(Adapts texts for texts-fit-items and items-fit-texts)* is among the good items, judging from the fact that the observer score curve is close to the expected score curve (see *Figure 6.1*). The weighted mean-square fit statistic of this item is 0.93, so the item discriminates teachers as well as the model predicts.

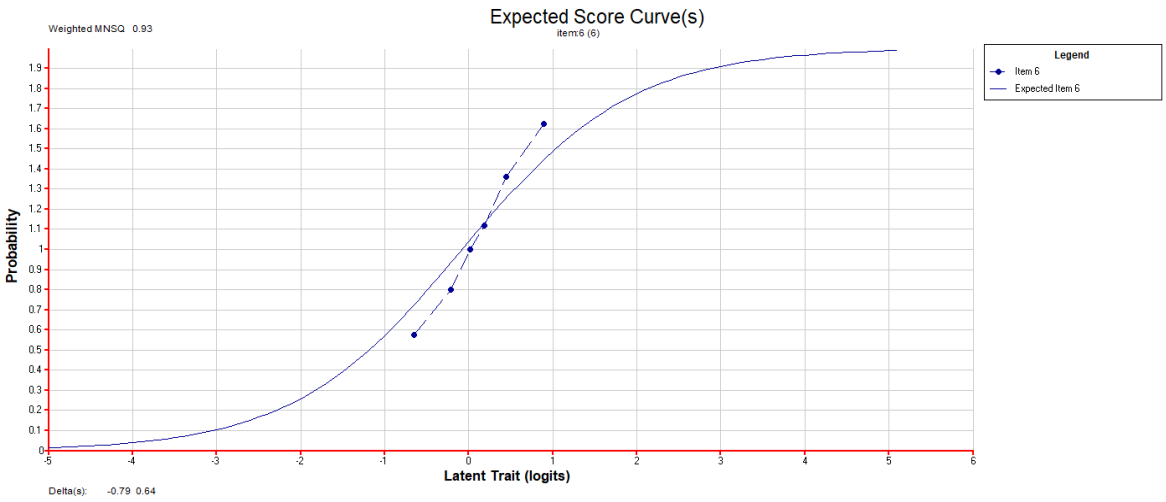


Figure 6.1 Item Characteristics Curve of Item 6 (Adapts texts for texts-fit-items and items-fit-texts).

Similarly, the observed score curve of item 12 (*Check item content fit*) (see *Figure 6.2*) has nearly the same slope as its expected score curve, which means the item discriminates teachers as well as the model predicts.

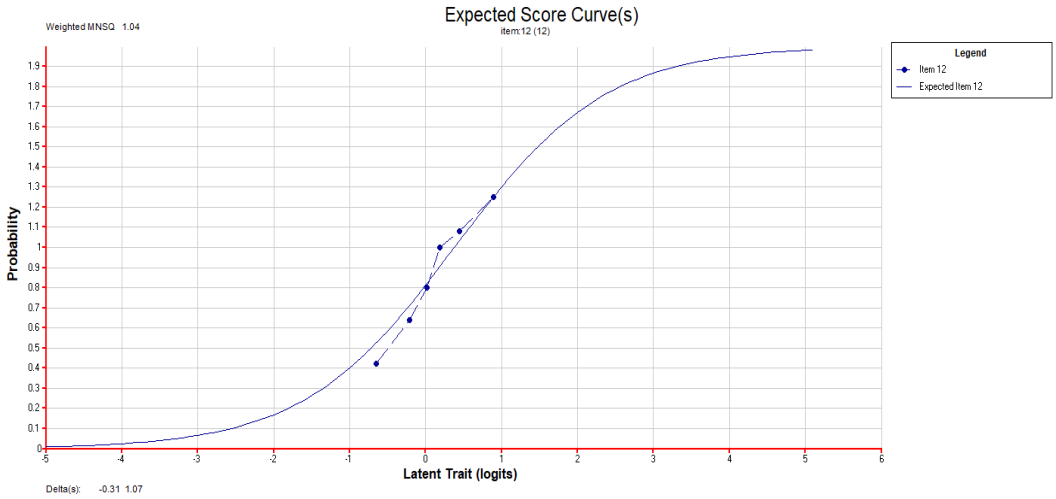


Figure 6.2 Item Characteristics Curve of Item 12 (Check item content fit).

Item 5 *(Make textual changes to the selected materials for tests)* is more discriminating than expected (see *Figure 6.3*). The expected curves of this item are steeper than those of the observed score curve.

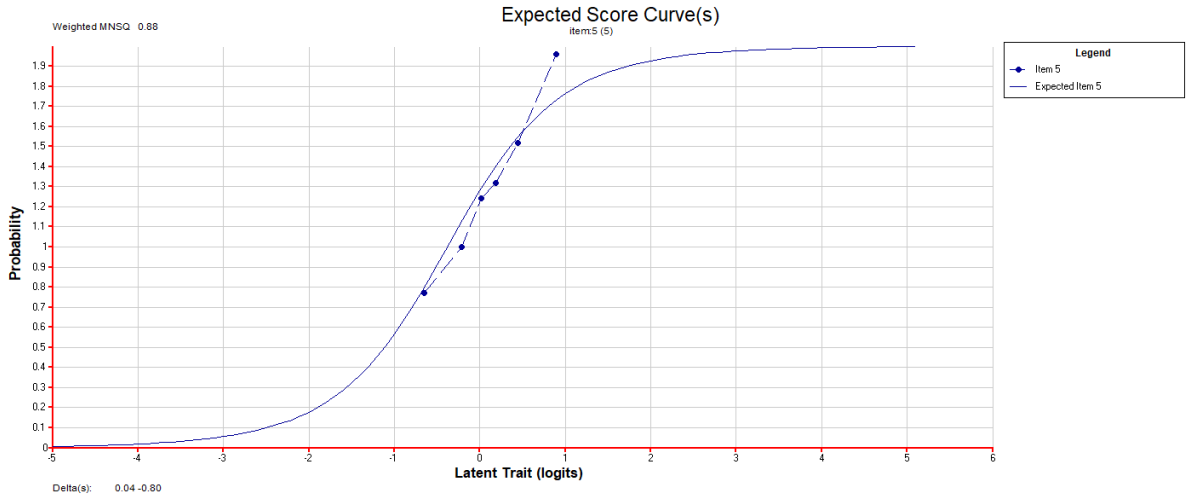


Figure 6.3 Item Characteristic Curve of Item 5 (Make textual changes to the selected materials for tests).

The item separation reliability coefficient (r=0.92) suggests that the items and categories were well separated in terms of difficulty, and the significant chi-square test of item parameter equality (p=0.000<0.05) indicated that the items were significantly different from each other in

difficulty. From these results, it can be concluded that the items in the IWCOMP scale can be divided into different levels of proficiency. Judging from the person separation reliability indices (r=0.68) and the expected a posteriori/plausible value (EAP/PV) reliability index of 0.70, the scale has moderately acceptable reliability for distinguishing teachers' underlying abilities. These modest indices are attributed to the fact that the participants in the sample were similar in terms of their educational backgrounds, work-related experience, and access to professional development.

The estimates of the person ability parameter and the item difficulty parameters were simultaneously plotted on this variable map, which illustrates the relative competence of the teachers and the difficulty of the item categories. The comparison between person ranges and items spread in the variable maps provides an illustration of whether the items match the teacher ability. As displayed in *Figure 6.4*, on the left is the *logit scale*, which provides the Rasch model metric of teacher competence and item difficulty. It enables teacher competence and item difficulty to be mapped on the same scale (Griffin, 1997). The next section, *persons* represents the frequency of distribution of the respondents, each represented by an 'X' (shown with X; each X represents 0.2 cases). The teacher competence estimates in *Figure 6.4* varied from -1.60 to +2.19, with a range of 3.79 logits. Of the 150 teachers surveyed, none had a zero or perfect score. This range suggests that variation was moderate in the selected-response test item writing competence of the teachers in the sample. On the right-hand side of the map is the item category difficulty distribution, which illustrates each item located on the scale according to its difficulty measure. As analysed above (see Section 6.1.1, Para. 6), the delta estimates of the item categories varied from – 1.57 to + 2.02, with a range of 3.59 logits. Furthermore, the data revealed that the item mean was 0 and the person mean was 0.066, indicating that these means are very close to each other. The person-item map and the means indicate there is a good match between the teacher's competence levels and the range of difficulty values for the behavioural

descriptors, suggesting that the scale has the appropriate level of difficulty to assess this group of teachers.

The Wright map of the items within the IWCOMP scale (see *Figure 6.4*) shows an ordering of typical performance behaviours exhibited by Vietnamese EFL teachers to demonstrate their possession of multiple-choice test item writing competencies. At the bottom end of the map are the behaviours that one would expect to see exhibited by novice item writers, and at the top are those one would expect to see exhibited by proficient item writers. Wilson, Scalise, and Gochyyev (2018) explain that:

> These construct maps are hierarchical in the sense that a person who would normally exhibit competencies at a higher level would also be expected to be able exhibit the competencies at lower levels of the hierarchy. The maps are also probabilistic in the sense that they represent different probabilities that a given competence would be expected to be exhibited in a particular context rather than certainties that the competence would always be exhibited. (p. 194 & p. 197).

The IWCOMP progression scale describes how novice item writers progress into proficient item writers and reflects their typical performance behaviours at each level. The scale is developmental and hierarchical so that those who are at the higher level can, based on probability, undertake the lower level and that the partial credit model has tested this in terms of the fits of the items. It is noted that how long one item writer stays in a level varies, and that not all the item writer teachers have to go through all the lower stages in order to be at the highest level.

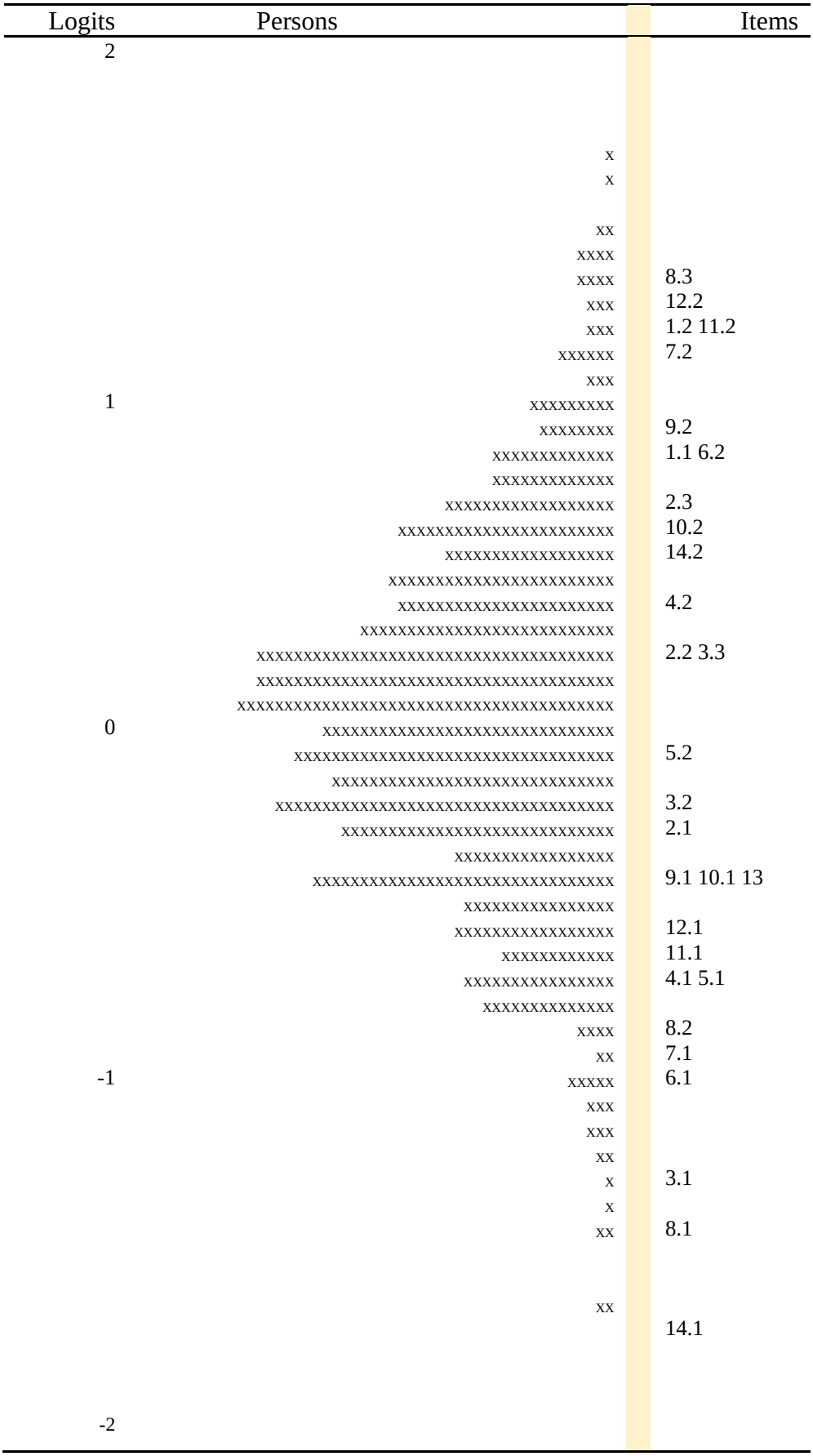


Figure 6.4 Interpretability of the variable map for the IWCOMP scale.

The item categories or quality criteria were grouped into clusters according to their estimated difficulty. A content analysis of the group of items was performed to explore the common underlying theme or meaning that emerged from its quality criteria and to summarise that meaning in a coherent description. Content analysis of criteria within each identified cluster resulted in development of summary descriptions of each level along a continuum of proficiency. These are shown in detail in Tables 6.4 to 6.8. The derivation of these levels is central to the notion of criterion- or standard-referenced assessment. As discussed in Chapter Two, the primary purpose of the scale was to help Vietnamese EFL teachers frame their judgements about their multiple-choice test item writing competence in ways that were functional and criterion-referenced, rather than being qualification- or length of experience-based or norm-referenced.

Table 6.4 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 4*

Item Category	Threshold	Item Content	Competence Level
8.3	1.35	Build on various resources to innovate items	
12.2	1.25	Attempt items on own to make judgements of item-fits-specifications for content complexity and difficulty levels	
1.2	1.18	Consider operational implications of the test specifications with target examinees in mind	
11.2	1.16	Draw on a range of factors affecting item difficulty from experience, expert colleagues and contemporary research to make judgements of item difficulty	Level 4: Proficient item writers
7.2	1.10	Multi-task between text/script adaptation and item construction	
9.2	.91	Use flexible repertoire of strategies to write distractors with consideration of typical errors or misconceptions of the subject content by test takers	
1.1	.82	Retrieve from prior experience important aspects of the test to be written	
6.2	.81	Call upon a range of strategies to transform source texts to fit specified test purposes and test items requirements	
2.3	.71	Develop own internal guideline for item writing in accordance with the given test specifications and informed by other training materials	
10.2	.62	Justify correct and incorrect answer options based on the evidence from text/script	
14.2	.56	Use information from experience and research to justify decision to implement or not to implement received feedback	
4.2	.42	Use multiple sources of evidence regarding text density, lexical difficulty, text potentiality for item development to judge the relevance of input materials	
2.2	.22	Use experience and available resources to internalise the test specifications	
3.3	.22	Source texts from a diverse range of resources with consideration of the target examinees	

Table 6.5 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 3*

Item category	Threshold	Item Content	Competence Level
5.2	-.08	Employ an extensive repertoire of strategies to adapt the original text to meet the required textual feature in terms of language use, style and their rhetorical functions	
3.2	-.19	Locate texts from previously used high-yield sources	
2.1	-.24	Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications	
13.1	-.37	Analyse responses from a self-conducted pilot with a sample of students to identify any miskeys and/or poor performing distractors	Level 3: Established item writers
10.1	-.40	Provide a reference for each answer key	
9.1	-.41	Utilise a range of paraphrasing strategies to construct distractors	
12.1	-.52	Consider abilities required to complete the items to compare what the items are thought to be testing with what is claimed in the test specifications	
11.1	-.65	Consider the text complexity and cognitive demand/sub-skills required to answer each item	
4.1	-.67	Assess texts based on defined characteristics of text contents and difficulty levels	
5.1	-.73	Make lexical and syntactic changes to achieve the target register	
8.2	-.87	Rely on different sources to create items accommodating test specifications and input materials	

Table 6.6 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 2*

Item category	Threshold	Item Content	Competence Level
7.1	-.93	Focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items	
6.1	-.99	Use research-grounded ruses to adjust texts to fit items	Level 2: Emergent item writers
3.1	-1.27	Identify texts of specified complexity and difficulty from newspaper, magazines and books written for non-native English readers	
8.1	-1.46	Construct items from a pool of item shells/item stems derived from established items in past test materials	
14.1	-1.68	Consider the test purposes to make the decision to implement feedback	

Table 6.7 *Item Criteria, Threshold Parameters and Derived Multiple-Choice Test Item Writing Competence Level 1*

Item category	Threshold	Item Content	Competence Level
1.0 -14.0	Below -1.68	1.0 Check aspects of the test specifications (e.g., the skills/subskills to be tested, text length, text type, item formats, quantity, and item difficulty levels) 2.0 Follow the instructions from the test specifications 3.0 Browse materials of required length on recommended topics 4.0 Use text analysis tools to determine text-fit-specifications in terms of difficulty and complexity 5.0 Restructure and cut texts to required length 6.0 Design items closely adhering to the generated text/script 7.0 Follow a standardised text/script-first item writing process 8.0 Replicate items from sample tests provided in the test specifications 9.0 Follow the format of the distractors given in the test specifications or sample tests + Do and don't guidelines 10.0 Assign an answer key for each item 11.0 Align items with the descriptors defined in the item blueprint and/or in the six-level English language proficiency framework for Vietnamese learners 12.0 Examine the conformity of the items to the test specifications regarding item content, wording of items and instructions 13.0 Try out items on own to check the original keys 14.0 Incorporate received feedback	Level 1: Novice item writers

A content analysis of the group of items was performed to determine whether there was a common interpretation of the underpinning construct (Griffin et al., 2004, 2007, 2018). An analysis of the relative positions and contents of the item clusters suggested four levels of competence among the population investigated, from the bottom end (Level 1: Novice item

writers) to the top end (Level 4: Proficient item writers). These levels in the empirical scale are relatively similar to those described in Dreyfus and Dreyfus model of skill acquisition (1986), which was the developmental learning model underpinning the development of the IWCOMP scale. These levels are related to what the teachers do, write, create and say in the level. Because the empirical continuum was formed with item thresholds, all item categories coded with 0, which formed the lowest level in the intended instrument (novice item writers), were not present on the variable map produced by CONQUEST 4.0. Thus, they were added to the bottom end of the map for easy interpretation. These item categories formed the first band level, which refers to the teachers who are identified as novice item writers.

Table 6.8 *The Developmental Progression for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests*

Level 4	**Teachers at this level are proficient item writers.** The item writers at this level have internalised the rules, guidelines or principles to the extent that they can perform the tasks independently. Proficient item writers intuitively organise, understand the task holistically and consciously analyse options to solve it. They draw on experience, research and various resources to internalise the test specifications. They develop their own internal guideline for item writing by synthesising the information from the test specifications and other training materials, reflecting on prior experiences in writing test items and teaching target examinees. They source input materials from diverse sources and take into consideration the characteristics of the target examinees. They use multiple sources of evidence regarding text density, lexical difficulty, text potentiality for item development (i.e., seeing items leap from the text) to judge the relevance of input materials. They employ flexible and extensive repertoires of strategies, build on various resources, and draw on successful outcomes in their past experiences and contemporary research to create items that accommodate both texts and test specifications. Proficient item writers demonstrate the ability to recognise the tasks and act upon this in accordance with prior successful experience. They tend to transform source texts and to write distractors with consideration of typical errors or misconceptions of the subject content held by examinees. They make connections between their wealth of experience and multiple sources of evidence to make intuitive and analytical judgements regarding text complexity, item content fit, and cognitive demands required for answering

	each of the items. They make justification for the correct and incorrect answer options based on the evidence from the text or scripts and their decision to implement or not to implement the item feedback received from reviewers. They multi-task between text/script adaptation and item construction.
Level 3	**Teachers at this level are established item writers.** Established item writers demonstrate deep understanding and analysis of rules, theories and alternatives and develop a sense of how to vary the uses of maxims for guidance in accordance with the situation in a flexible way based on previously successful item writing experiences. They source input materials for tests from previously used high-yield sources and evaluate the materials based on defined characteristics of text contents and difficulty levels. They employ an extensive repertoire of strategies to adapt the original text to meet the required textual features in terms of language use, style and their rhetorical functions. They rely on different sources to construct items that accommodate both the test specifications and the features of the input materials and utilise a range of paraphrasing strategies to construct item distractors. They provide references to support the answer keys by underlining part/parts of the text/script that justify the answers. They self-conduct an item pilot with a sample of students and analyse these responses to identify any miskeys and/or poor performing distractors. Apart from using well-established guidelines and strategies, these item writers begin to make intuitive judgements about the congruence between the items and the test blueprint and the cognitive demands required for successfully answering the items.
Level 2:	**Teachers at this level are emergent item writers.** Emergent item writers appear to consciously and deliberately plan and use standardised and routinised procedures with consideration of whether the rules are pertinent to a given context and show their adequate fund of experience in multiple-choice test item writing. They search for the input materials of specified complexity and difficulty from newspapers, magazines and books written for non-native English readers. They utilise research-grounded ruses to adjust texts to fit items and focus on skills to be tested to write items in accordance with the test requirements and to alter the text to accommodate the items. Thus, they tend to employ an item-first approach rather than the standardised text-first approach. They construct items from a pool of well-documented item shells/item stems derived from established items in past test

	materials. Emergent item writers begin to make connections between the test items and test purposes when making decisions regarding implementing feedback.
Level 1	**Teachers at this level are novice item writers.** Novice item writers closely follow learnt rules, clearly defined procedures, and general "recipes" and examples provided to carry out most of their tasks in the item writing process. They start with the review of the test specifications by checking mandatory aspects of the test specifications such as the skills/subskills to be tested, text length, text type, item formats, quantity, and item difficulty levels. They strictly follow the instructions from the test specifications. They browse the Internet for input materials on recommended topics that meet the required length and make use of available text analysis tools to help them decide if the materials fit the test specifications in terms of text difficulty and complexity. They tend to design items closely adhering to the generated text/script; thus, they make minimum changes to the text/script. They replicate exemplary item shells and item distractors from sample tests provided in the test specifications and past test materials to write item stems and distractors. They focus on conforming the items they write to the requirements in the test specifications with regard to item content, wording of items and instructions. They try out the items on their own to check the original keys and identify any issues related to miskeys. They align items with the descriptors defined in the item blueprint and/or in the six-level English language proficiency framework for Vietnamese learners. As they are novice in the task, they tend to attune to and incorporate all the feedback they receive from reviewers. Novice item writers are well aware of default procedures and rules to write selected-response test items and apply them rather rigidly, mechanically and intently focused on the examples provided.

In summary, the behaviourally anchored rating scale designed for Vietnamese teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests fitted the model well. These 14 items were also well separated along the continuum to differentiate between levels of item writing proficiencies. There were four levels of competence in multiple-choice test item writing for English reading and listening skill tests among the teachers in the sample: 'novice', 'emergent', 'established', and 'proficient'.

The development and validation of the IWCOMP scale have illustrated the unitary notion of validity proposed by Messick (1989, 1995) and adopted as general validity criteria for all assessment and measurement by the Standards for Educational and Psychological Testing (hereafter referred to as the *Standards*) (AERA et al., 2014). According to Messick (1989, 1995), all validity is subsumed under construct validity and is concerned with "overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions on the basis of test scores or other models of assessment" (Messick, 1989, p. 741). This unitary concept of validity is also advocated by the *Standards* (AERA et al., 1985, 1999, 2014). In the most recent *Standards*, validity is defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA et al., 2014, p. 11). Content validity considers the representativeness of the construct and the capacity of the items to represent the construct, and structural validity addresses the degree to which the outcomes reflect underpinning domain theory (Wolfe & Smith, 2007). This validity requirement for the IWCOMP scale was supported by grounding the conception of multiple-choice test item writing competence in domain theory (as depicted in Figure 3.4). The alignment of the items within the IWCOMP to the reviewed literature provided evidence of content validity (see Appendix E). This validity requirement was supported by engaging item writers and subject matter experts in the focus groups and workshops and utilising expert review during the development process of the scale. The detailed documentation of this process constituting a validity argument is elaborated in Chapter Five (see Section 5.1.1.1). The Rasch model supports a unified concept of validity grounded in construct validity in terms of separating the item writers and the items along a defined variable. The analysis of the psychometric property of the scale considered construct validity - establishing the degree to which the items accurately measured the underlying construct - while analysis of person fit established the extent to which the construct abilities of all teacher groups were fairly

represented by the items (Wright, 1995). In terms of the appropriateness of item difficulty and type for the target group, the scale was found to target the sample of teachers well in terms of matching difficulty levels of the full spectrum of teacher competence and provided a range of types of item difficulty at each level. Teachers of varying levels of competence were able to achieve across the scale.

The results of the calibration for the finalised 14-item progression scale generated from both classical and Rasch analyses have shown that the scale had satisfactory psychometric properties. The overall good fit statistics and the item-rest discrimination estimates of the items indicated that there was one dominant construct of multiple-choice test item writing for English reading and listening skill tests, and that the items worked well together to measure the underlying construct. Overall, the items formed a cohesive scale representing "Vietnamese teacher competence in multiple- choice test item writing for English reading and listening skill tests". The item separation reliability coefficient (r=0.92) suggested that the items and categories were well spread along the continuum, enabling it to be defined as a meaningful developmental progression. The chi-square test of item parameter equality was significant (p= <0.05), indicating that the items were significantly different from each other in difficulty. From these results, it can be concluded that the items in the IWCOMP scale can be divided into different levels of proficiency. Judging from the person separation reliability indices (r=0.68) and the expected a posteriori/plausible value (EAP/PV) reliability index of 0.70, the scale has moderately acceptable reliability for distinguishing the teachers' underlying abilities. The person-item threshold map indicates there was a good match between the teachers' self-perceived competence levels (3.79 logits) and the range of difficulty value (3.59 logits) among behavioural descriptors, suggesting that the scale has the appropriate level of difficulty to assess this group of teachers. Together, these figures gave support for the construct validity of the scale.

Furthermore, validity was supported by a unidimensional construct. From the scholarly literature, this study identified a unidimensional construct of multiple-choice test item writing competence. The fit statistics for the IWCOMP scale (see Table 6.3) indicated that the data fitted the Rasch model specifications, supporting that the multiple-choice test item writing competence was a unidimensional construct, which means that the responses to the items in the scale derive only from ability in the latent trait (Wu et al., 2016). The derived order of the quality criteria within the progression scale and their clustering together in larger bands accord with their estimated difficulty. Content analysis of criteria within each identified cluster resulted in development of summary descriptions of each level along a continuum of proficiency. The items within the IWCOMP scale fitted the model well and were also well separated along the continuum to differentiate between levels of item writing proficiencies as described in the hypothesised progression guided by the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) during the scale construction. This also contributed to a validity argument because it indicated the extent of alignment of outcomes with the original purposes as well as providing an ordered definition of the construct. The results visually depicted on the item-person map indicated the derived order of difficulty of the item steps, and established how well the spread of item difficulties matched the original conception and the spread of abilities (see *Figure 6.4*). Thus, the IWCOMP scale developed and validated in this study for Vietnamese teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests has acceptable reliability and has demonstrated the unified concept of validity as described in Messick (1989, 1995) and AERA et al., (2014).

6.1.2 Calibration of the PROLEARN Scale

The PROLEARN scale, as presented in the previous chapter, was adapted from the survey instrument to measure teachers' professional learning in the work-place developed by Kwakman (2003) and Evers et.al. (2016). Consisting of 15 items, the scale was designed to

measure the teachers' engagement in four categories of professional learning: Keeping up-to-date, experimenting, reflection, and collaboration (Kwakman, 2003). For this scale, the items were rated on a 4-point Likert scale, but the response patterns were recoded into a dichotomous score of one (Yes) and zero (No) for "Sometimes and Often" and "Never and Rarely" respectively. The items within the PROLEARN scale can be found in Table 5.7.

The results from both the classical and Rasch analyses for the 15 items within the scale are presented in Table 6.9. The summary of contents in the findings is as follows: the item average delta estimates (Logit, δ_i), the standard error of measurement associated with each delta estimate (SE), the item-rest correlation, and the INFIT MNSQ and t-statistics. The items are presented in terms of decreasing item difficulty in Table 6.9.

Table 6.9 *Calibration Estimates for the PROLEARN Scale*

Item	Descriptions	Avg Delta (δ_i)	SE	Item-rest corr	INFIT MNSQ	T
6	I conduct/co-conduct training sessions for item writers.	2.33	0.25	0.33	1.01	0.1
5	I attend professional learning opportunities outside my institutions.	1.20	0.20	0.34	0.97	-0.2
3	I read research publications on test item writing.	1.08	0.19	0.30	1.05	0.6
4	I participate in compulsory training workshops organised by my institutions for teacher item writers.	0.60	0.18	0.45	0.93	-0.9
15	I support or coach colleagues in test item writing.	0.60	0.18	0.38	1.01	0.1
12	I experiment with new strategies and approaches to test item writing.	0.50	0.18	0.38	0.99	-0.2
2	I read item writing manuals from online resources.	0.47	0.18	0.31	1.07	1
9	I record/document my own best practices in test item writing.	0.13	0.18	0.38	1.01	0.2
7	I tend to write items in a similar way to how I have written then in the past.	-0.79	0.19	0.38	0.97	-0.3
8	I reflect on/recall prior item writing experiences to suggest lessons learnt for future tasks.	-0.79	0.19	0.35	1.01	0.2
14	I share resources and ideas about test item writing with colleagues.	-0.87	0.19	0.37	0.98	-0.2
13	I talk about problems with writing test items with colleagues.	-1.02	0.19	0.38	0.98	-0.2
1	I study guidelines provided to item writers.	-1.05	0.19	0.30	1.04	0.4
10	I utilise my previously used strategies for test item writing.	-1.06	0.19	0.34	1.00	0
11	I adjust my test item writing strategies and approaches based on lessons learnt from past experiences.	-1.34	0.20	0.38	0.97	-0.3

According to Table 6.9, the INFIT MNSQ values ranged from 0.93 to 1.07, falling within the accepted range proposed by Adams and Khoo (1996). The mean Item INFIT was 1.0, with a standard deviation of 0.03. The INFIT MNSQ values demonstrated that the items had an

acceptable fit to the Rasch model, and that there was evidence of a dominant underlying construct in the variable being measured (Wright & Masters, 1982; Wright & Stone, 1979).

The mean teacher engagement level was -0.02, with a standard deviation of 1.30, indicating that the average difficulty of the scale was well matched to the average engagement of the teachers.

The item difficulty estimate varied from +2.33 to – 1.34, a range of 3.67 logits. The mean item difficulty was set to be zero. The standard deviation of item difficulty levels was 1.04. The measurement errors for each of the items were acceptable, with the largest value of 0.25 for item 6 item (*I conduct/co-conduct training sessions for item writers*) which was the item that represented the highest level of teacher engagement in self-regulated professional learning. The item that reflected the lowest level of teacher engagement was item 11 (*I adjust my test item writing strategies and approaches based on lessons learnt from past experiences*). The teacher engagement estimates ranged from +4.07 to -3.88, a range of 7.95 logits. None of the teachers received perfect scores on the scale (i.e., none sometimes or often engages in all of the professional learning activities mentioned in the scale) and none received a zero score (i.e., none never or rarely engages in any of the professional learning activities mentioned in the scale). This indicated that the logit range of the set of items was much narrower than the engagement level range of the teachers.

All items within the PROLEARN scale discriminated well, given the range of the item-rest correlation being within the range of between 0.30 and 0.45 (Wu & Adams, 2007; Linacre, 2007). The item separation reliability coefficient (r=0.97) suggests that the items were sufficiently well separated in terms of difficulty levels within the latent construct. The significant chi-square test of item parameter equality ($p=0.000<0.05$) indicates that the items were significantly different from each other in difficulty. The person separation reliability indices (r =

0.732), demonstrating the teacher engagement in self-regulated professional learning is moderately spread.

As a result of the item logit range, the good item-rest discrimination, the acceptable fit statistics and the relatively small measurement errors, it can be concluded that the scale permitted the reliable measurement of teacher engagement in self-regulated professional learning spanning varied levels of teacher engagement.

The estimates of the teacher engagement parameter θ_n and the item difficulty parameters δ_i were simultaneously plotted on the variable map (see *Figure 6.5*).

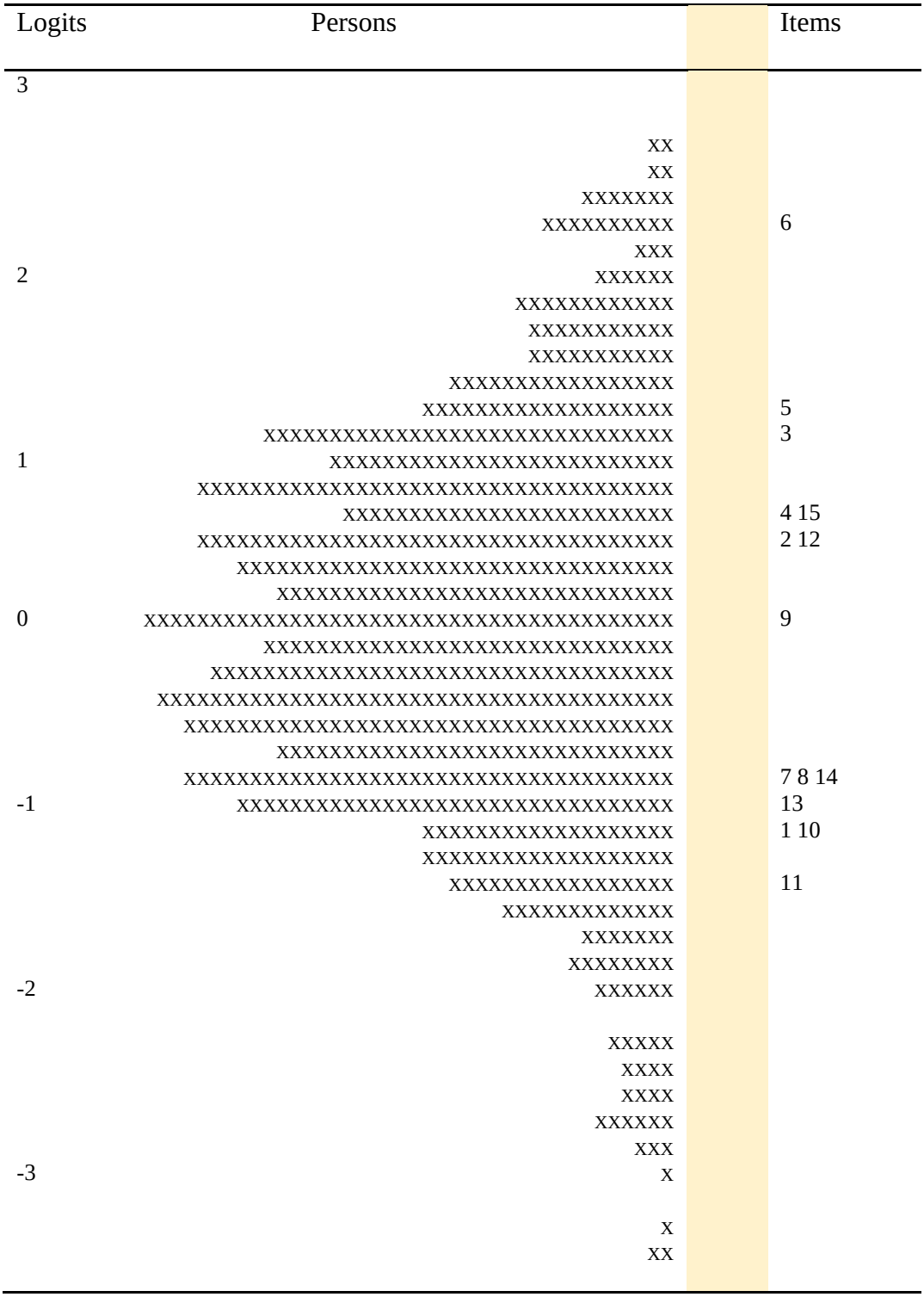


Figure 6.5 The variable map for the PROLEARN scale

A content analysis of the items that were clustering together in terms of difficulty estimates was conducted to seek to identify the common set of behaviours underpinning teachers' self-regulated professional learning within each cluster. The groups of item clusters

and the derived levels of engagement in teachers' individual professional learning produced as a result of the interpretation of the item clustering analysis is presented in Table 6.10 and the gist statement describing each professional learning level is presented in Table 6.11.

Table 6.10 *Item Criteria, Delta Parameters and Derived Levels of Vietnamese EFL Teacher Engagement in Self-Regulated Professional Learning*

Item	Delta	Item content	Professional Learning Level
6	2.33	I conduct/co-conduct training sessions for item writers.	Level 3: Proactive learning
5	1.20	I attend professional learning opportunities outside my institutions.	
3	1.08	I read research publications on test item writing.	
4	0.60	I participate in compulsory training workshops organised by my institutions for teacher item writers.	Level 2: Dynamic learning
15	0.60	I support or coach colleagues in test item writing.	
12	0.50	I experiment new strategies and approaches to test item writing.	
2	0.47	I read item writing manuals from online resources.	
9	0.13	I record/document my own best practices in test item writing.	
7	-0.79	I tend to write items in a similar way to how I have written them in the past.	Level 1: Collegial learning
8	-0.79	I reflect on/recall prior item writing experiences to suggest lessons learnt for future tasks.	
14	-0.87	I share resources and ideas about test item writing with colleagues.	
13	-1.02	I talk about problems with writing test items with colleagues.	
1	-1.05	I study guidelines provided to item writers.	
10	-1.06	I utilise my previously used strategies for test item writing.	
11	-1.34	I adjust my test item writing strategies and approaches based on lessons learnt from past experiences.	

Table 6.11 *Interpretability of the Variable Map for the PROLEARN Scale*

Items	Derived Descriptions of Professional Learning Levels
6,5&3	Level 3: Proactive learning The teachers plan for and self-direct their professional learning by accessing and utilising relevant research and participating in high quality targeted opportunities to improve practice. They lead /co-lead professional learning programs that address the needs of colleague item writers
4, 15,12,2 & 9	Level 2: Dynamic learning The teachers initiate and engage in professional discussions with colleagues in both internal and external forums to be exposed to peer-feedback, open discussions and new ideas and knowledge. They are dedicated to implementing newly acquired strategies directed at improving knowledge and practice.
7,8,14,13, 1,10 &11	Level 1: Collegial learning The teachers learn from the materials that are offered and shared with them by reflecting on them in light of their own prior experiences and by engaging in informal conversations with colleagues.

As shown in Table 6.11, it is indicated that at the top level, the item writers in the sample practised proactive professional learning. These teachers plan for and self-direct their professional learning by accessing and utilising relevant research and participating in high quality targeted opportunities to improve practice. They lead/co-lead professional learning programs that address the needs of colleague item writers. At the middle level, the teachers exercised dynamic professional learning practices. The teacher item writers initiated and engaged in professional discussions with colleagues in both internal and external forums to be exposed to peer-feedback, open discussions and new ideas/shared resources and to evaluate test item writing practices directed at improving knowledge and practice and to support colleagues to identify and achieve their professional goals. At the bottom level, the item writers conducted

their professional learning in a closed community of practice. They read the materials shared with them and learned to improve their item writing practice by recalling their prior experiences and to overcome their difficulties through engaging in informal professional conversations with colleagues. In conclusion, the results from the Rasch analysis of the PROLEARN scale showed that the teacher participants had varying levels of engagement in self-regulated professional learning. It is noted that an EFL teacher participant does not have to go through all lower levels in order to reach a higher level. The PROLEARN scale was adapted to locate where the teachers were in the scale to distinguish the teachers from one another.

6.1.3 Calibration of the INTRAIN Scale

The Induction and Training (INTRAIN) scale comprised 21 dichotomous items that indicated varying levels of access to and sophistication of formal induction and training activities offered to teacher item writers by their institutions in supporting teachers' professional learning in the domain of test item writing. The items covered the aspects of training and induction and the scale was built based on the Cambridge Assessment English training and development framework (RITCME) in light of Desimone's (2009) core features framework and the Observation Checklist for High-Quality Professional Development Training (HQPD Checklist) developed by Erickson et al. (2017) (See Chapter 4, Section 4.1.2) for measuring high quality professional training. Respondents were required to indicate if the training and induction activities are provided to them by their institutions targeting their development of expertise in multiple-choice item writing. The items within the INTRAIN scale can be found in Table 5.9. Table 6.12 displays the calibration estimates for the Induction and Training Scale.

Table 6.12 *Calibration Estimates for the INTRAIN Scale*

Item	Description	Avg delta (δ_i)	SE	Item-rest Correlation	INFIT MNSQ	T
17	I have opportunities to engage in feedback on live test item performance with colleagues.	1.85	0.21	0.52	0.97	-0.3
16	I have opportunities to analyse recent live test papers.	1.62	0.21	0.47	1.1	0.9
18	I have opportunities to learn about basic measurement characteristics of test items	1.51	0.21	0.57	0.95	-0.5
15	I have opportunities to engage in hands-on item writing practice tasks with colleague item writers of varying experience.	1.02	0.2	0.56	1.02	0.2
20	I am trained by relevant trainers/have opportunities to work with and learn from experienced/ well qualified trainers.	0.88	0.2	0.5	1.09	1
14	I have opportunities to work with colleagues to identify features of sound items.	0.76	0.2	0.65	0.88	-1.3
10	I am trained about techniques for writing different objective test item types	0.41	0.21	0.63	0.86	-1.5
13	I have opportunities to consider the strengths and weaknesses of the items with colleagues.	0.26	0.21	0.64	0.89	-1.1
3	I am provided with additional guidelines for item writers.	0.07	0.21	0.38	1.23	2.2
8	I am provided with readings, activities and/or questions to think about prior to the training.	0.04	0.21	0.47	1.12	1.2
9	I am trained about the principles of test item writing.	-0.1	0.22	0.69	0.76	-2.5
19	I am trained about using different tools and resources to support my item writing.	-0.29	0.22	0.42	1.09	0.9
6	I am provided with a training agenda	-0.3	0.22	0.46	1.07	0.7
7	I am provided with a description of the training with learning objectives prior to training.	-0.4	0.22	0.53	0.98	-0.1
12	I am provided with example test items to analyse.	-0.59	0.23	0.62	0.85	-1.3
21	I have opportunities to participate in training and development workshops for teacher item writers 1-2 working weeks a year.	-0.7	0.23	0.48	0.96	-0.3
11	I have opportunities to engage in discussions with colleagues on item writing techniques.	-0.72	0.23	0.55	0.92	-0.6
4	I am briefed about the test specifications.	-0.74	0.23	0.42	1.09	0.7
1	I am provided with a documentation of the task description.	-1.05	0.25	0.48	0.96	-0.3
5	I have opportunities to discuss the test specifications with colleagues for further clarification.	-1.62	0.28	0.49	0.92	-0.4
2	I am provided with sample tests.	-1.91	0.3	0.25	1.28	1.3

The 21-item scale has a range of approximately 3.76 logits, from a minimum of -1.91 to a maximum of +1.85. Item 17, *'I have opportunities to engage in feedback on live test item performance with colleagues'* was the highest-level item on the Induction and Training Scale with only 38.5 % of the sample indicating that they had opportunities to engage in the activity. In contrast, item 2, *'I am provided with sample tests'* was the lowest-level, with the majority of the sample (87.8%) reporting that they were provided with the materials. The mean item delta was set to zero, with a standard deviation of 0.99. The 21 items on the scale all had acceptable measurement errors, ranging from a minimum of 0.20 to a maximum of 0.30. As expected, the largest measurement error was for the lowest-level item on the scale (i.e., item 2). The item-rest correlation indices ranged from 0.25 to 0.69 within the acceptable range according to Adams and Khoo (1996). The item and person separation indices of the INTRAIN variable were relatively high, r= 0.95 and r= 0.80 respectively (Wright & Masters, 1982). These indicated sufficient separation of the 21 items on the variable and lent support to the dominant construct validity of the scale.

The scale had a mean item INFIT value of 1.01, with a standard deviation of 0.15. All items, with the exception of item 3 (INFIT MNSQ = 1.31) and item 9 (INFIT MNSQ = 0.76), had an acceptable fit to the model as proposed by Adams and Khoo (1996). Item 3 and item 9 were designed to measure whether the item writers were provided with additional guidelines for item writing and were trained about the principles of test item writing. The descriptive statistics for these items showed that the resources and training were accessible to approximately 70% of the sample. These items had INFIT values which were outside the accepted range proposed by Adams and Khoo (1996). Nevertheless, these fit statistics were within the acceptable range proposed by Wright and colleagues (1994) for analysing questionnaire items (i.e., 0.6 -1.4), and therefore were retained on both statistical and substantive grounds.

Overall, the fit statistics, in addition to the discrimination estimates for each item, provided evidence that the set of items were measuring a dominant construct and appeared to differentiate levels of item writers' access to induction and training.

Similar to the previous scales, the estimates of the access parameter θ_n and the delta parameter δ_n were simultaneously plotted on the variable map, illustrated in *Figure 6.6*. The variable map comprising three different sections is displayed in *Figure 6.6*. The person-item variable map displays the range of induction and training levels and item difficulties on the same scale. On the left side of the map is the induction and training distribution (shown with 'X'). In this example, the 'X' within the frequency distribution represents 0.3 cases. The induction and training estimates in Table 6.12 varied from -4.23 to +4.26, a range of 8.49 logits. This broad range indicates a large variability in the amount and quality of induction and training offered to the teachers in the sample. On the right side of the map is the item difficulty distribution, which presents each item located on the scale according to its difficulty measure. The item difficulty varied from -1.91 to +1.85, a range of 3.76 logits. There were two teachers with zero scores (i.e., they responded 'No' to all of the 21 items) and 21 with perfect scores (i.e., they endorsed all the items within the scale).

The variable map demonstrates the induction and training levels were broader than the item difficulty range, indicating the instrument might have a moderate level of difficulty to assess the induction and training offered to this group of teachers.

Logits	Persons	Items
	x	
	xxx	
6	x	
	xx	
	xxxx	
	xxxxxx	
5	xxxx	
	xxxxxx	
	xxxxxxxx	
	xxxxxxxxx	
4	xxxxxxxxxx	
	xxxxxxxxxxxx	
	xxxxxxxxxxxx	
	xxxxxxxxxxx	
3	xxxxxxxxxxxxxxxxxx	
	xxxxxxxxxxxxxxx	
	xxxxxxxxxxxxxxxxxxxxxxx	
	xxxxxxxxxxxxxxx	
2	xxxxxxxxxxxxxxxxx	17
	xxxxxxxxxxxxxxxxxxxxxxxxx	16
	xxxxxxxxxxxxxxxxxxxxxxxxx	18
	xxxxxxxxxxxxxxxxxxxxxxxx	
1	xxxxxxxxxxxxxxxxxxx	15 20
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxx	14
	xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx	10
	xxxxxxxxxxxxxxxxxxxxxxxxxxxx	3 8 13
0	xxxxxxxxxxxxxxxxxxxxxxxxxxxx	9
	xxxxxxxxxx	6 7 19
	xxxxxxxxxxxxxxxx	4 11 12 21
-1	xxxxxxxxxxxx	
	xxxxxxxxxxxxxx	1
	xxxxxx	
	xxxx	5
-2	xxxx	2
	xxxx	
	xxx	
	xx	
-3	xx	
	xx	
	x	
	x	
-4		
	x	
-5	x	

Figure 6.6 The variable map for the INTRAIN scale.

Figure 6.6 shows that the item difficulty estimates for the INTRAIN scale were plotted in decreasing order at varying levels on the measurement. These clustered items were further examined for their common themes of behaviours that underpinned the latent construct. Through the interpretation of these themes, the four band levels were identified. The results of this content analysis are presented in Table 6.12 above. Table 6.13 indicates that the 21 items formed several clusters that exhibit different levels of accessibility and quality of induction and training along the continuum.

Table 6.13 *Item Criteria, Delta Parameters and Derived Description for Vietnamese EFL Teacher Access to Institutional Induction and Training*

Item	Induction and Training Levels
17 16 18 15 20	**Level 4: Advanced training** Teachers have access to professional training that allows for hands-on collegial and collaborative learning through application, peer feedback and reflection on outcomes. They also have opportunities to analyse items' robustness based on their psychometric properties.
14 10 13 3 8	**Level 3: Participatory training** Teachers are offered opportunities to acquire knowledge and skills through active participation in sample test item analysis with colleagues.
9 19 6 7 12 21 11 4	**Level 2: Content-focused training** Teachers have access to organised, coherent and content-focused induction and training. They are equipped with theoretical principles of test item writing, essential tools and resources, and established additional guidelines to support their item writing tasks.
1 5 2	**Level 1: Essential provision of documentation** Teachers are provided with test specifications and sample tests to start with. They have opportunities for clarification and discussion of the test specifications through formal briefings and/or informal collegial talks.

Analysis of the relative positions and contents of the item clusters suggests four levels of induction and training among the population investigated, from the bottom end (Basic training or Essential provision of documentation) to the top end (Advanced training). The first band level

refers to the teacher item writers who barely received any formal induction and training. They are provided with essential documentation such as an invitation letter, the test specifications and some sample tests to start with. They often seek clarification of the test specifications through informal collegial talk. This level of induction and training was termed "*Essential provision of documentation*". The second band is identified as "*Content-focused training*". The teachers in this band have access to organised, coherent and content-focused induction and training in which they are briefed about the test specifications and are equipped with theoretical principles of test item writing, essential tools and resources to support their item writing task, and example test items. The teachers in the third band, which is termed "*Participatory training*", are given opportunities to acquire knowledge and skills through active participation in sample test item analysis with colleagues. The fourth band is described as "*Advanced training*". The teachers in this band have access to professional training which allows for hands-on collaborative learning through application, peer feedback and reflection on outcomes with colleagues. They also have opportunities to analyse item robustness based on their psychometric properties.

The four bands logically provide a direction along the continuum of INTRAIN. This method of interpretation offers a more formal description of the underlying construct of induction and training for item writers originating from the item and score point description. The results indicated that that the teacher participants had varying levels of access to institutional induction and training. The calibration of the INTRAIN scale helped to measure where the teachers were in the scale to distinguish the teachers from one another.

In summary, the 21 items designed to construct the latent variable INTRAIN to measure the accessibility and quality of institutional induction and training for teacher item writers involved in the study fitted the model well. These items were also well separated along the continuum to differentiate between induction and training levels. There are four levels of induction and training among the teachers in the sample: *'Essential provision of documentation,*

Content-focused training, Participatory training and Advanced training'. The scale to measure institutional induction and training, thus, had logic, face validity, content validity and construct validity.

6.2 Summary of Statistics

The estimates of the classical and Rasch analyses for each scale are summarised in Table 6.14.

Table 6.14 *Summary Estimates of the Classical and Rasch Analyses for Each Scale.*

Variable name	No. of items	Scale logit			Person logit			Item INFIT			Item separation reliability	Person separation reliability	Cronbach's Alpha
		Min	Max	Range	Min	Max	Range	Mean	Min	Max			
IWCOMP	14	-1.57	2.02	3.59	-1.60	2.19	3.79	0.99	0.89	1.15	0.92	0.68	0.69
PROLEARN	15	-1.34	2.33	3.67	-3.88	4.07	7.95	1.00	0.93	1.07	0.97	0.73	0.76
INTRAIN	21	-1.91	1.85	3.76	-4.23	4.26	8.49	1.01	0.76	1.31	0.95	0.80	0.90

As shown in Table 6.14, the range of the teacher item writers' parameter estimates were adequately matched with the range of the item parameter estimates within the multiple-choice item writing competence (IWCOMP). With regard to the Professional Learning (PROLEARN) and Induction and Training (INTRAIN) scales, it appears that the range of the teachers' parameter estimates was much broader than the range of the item parameter estimates. This suggests that these two scales would need to include more items at the scale ends to better match the range of the teachers' parameters in future revisions.

In examining the internal consistency reliability (i.e., Cronbach's alpha) of the scales, classical item analysis procedure was employed to measure the intercorrelation of the individual items on each scale. As acknowledged by Clark and Watson (1995) and Schmitt (1996), there is no standardised acceptable and unacceptable levels of Cronbach's alpha. Previous literature highlights a variation in the values that are considered to be the acceptable Cronbach's alpha levels. For example, Nunnally (1978) found .70 the acceptable level of Cronbach's alpha. Other

researchers such as Dekovic, Janssens, and Gerris (1991), Holden, Fekken, and Cotton, 1991, and Myers and Oetzel (2003) endorse the level of Cronbach's alpha for all scales as varying .69 to .90 demonstrating moderate to high internal consistency reliabilities. Thus, all the scales used in the present study had acceptable internal consistency reliabilities.

To further assess the reliability of each of the scales of measurement, the reliability of item separation on each scale was computed (Wright & Masters, 1982). The Item Separation Reliability Estimates ranged from a minimum of 0.92 for the IWCOMP variable to a maximum of 0.95 and 0.97 for the INTRAIN and the PROLEARN variables respectively. The high item separation reliability index for each of the scales supported the scale's ability to be used in defining the latent construct where the items acted as pointers to location on the variable. It was also used as a measure of construct validity (Wright & Masters, 1982).

In addition to the classical analyses of internal consistency reliabilities for all the scales, calculation of the Rasch reliabilities on person separation (Wright & Masters, 1982) for these scales were undertaken to further examine the reliability of each scale. Table 6.14 reports that the person separation reliability estimate ranged from 0.68 for the IWCOMP scale to 0.80 for the INTRAIN scale, demonstrating moderate to high internal consistency reliabilities. Hence, the results obtained from both Classical and Rasch analyses show that all the scales had satisfactory measurement properties. The next chapter reports the results from the regression analysis of the study.

CHAPTER 7: MULTIPLE REGRESSION ANALYSES

This chapter reports the results obtained from the multiple regression analyses of the data collected in the study. The results are organised into four parts. Part One presents the univariate results. Part Two reports the results for univariate, bivariate, and multivariate normality. Part Three and Four present the bivariate results and the multivariate results from the multiple regression analyses.

7.1 Survey Sample Characteristics

The sample consisted of 150 EFL teachers who worked at universities across 12 cities and provinces of Vietnam, though the majority of the sample (75.3%) taught English at colleges and universities in urban areas in the North and the Central Region of Vietnam. In general, they represented EFL teachers from state-owned tertiary institutions who had been selected to be involved in the government-funded Strategic National Foreign Language Project 2020 to develop test items for the Vietnamese Standardised Tests of English Proficiency (VSTEP®)

Of the sample, only 20 were male, and the rest were female. 148 of the 150 surveyed teachers had obtained their post-graduate qualifications, with 131 teachers having Master's degrees and 17 doctoral degrees. Only two teachers had a Bachelor degree as their highest level of qualification. Slightly more than half of the sample had their highest degree in TESOL/TEFL, followed by Applied Linguistics and Education which accounted for 27.3% and 16% respectively. Only two respondents majored in language testing and assessment, and 6 in other unclassified fields.

Approximately three quarters (72.7%) of the surveyed teachers had undertaken a course in language testing and assessment as part of their highest degree study, whereas 27.3% had not received such pre-service training. Of those who undertook pre-service training in language testing and assessment, nearly half reported that the course prepared them either fairly or very

well for writing EFL test items while a third claimed the course "somewhat" prepared them for the task, and only 2 said that it did not help at all.

The teachers in the sample had varied years of experience in teaching English language skills, teaching EFL examination preparation courses and constructing EFL tests from published test books, with a mean period of 12.31 years (SD=6.65), 6.81 (SD=4.51), and 4.97 (SD=5.52) correspondingly.

With regard to the number of VSTEP© tests, this number refers to the number of VSTEP© listening and/or reading skill tests the participants had written to date, which reflected their experience in multiple-choice English test item writing. According to the study's descriptive statistics, on average, the respondents had written 20.45 tests (SD=30.33) and 5 teachers reported that they had written 100-150 tests for VSTEP© examinations. Given the frequency in which VSTEP was delivered; for example, in one tertiary institution, the VSTEP© examination was organised twice a week, it meant the examination boards were in a great need of new tests. Moreover, as the operational VSTEP© was newly introduced, it could be the intention of the examination boards to develop the maximum test materials possible, offering them flexibility in assembling operational test forms to maximise test security. It was understandable that the participants had a high workload in terms of test item development for test item banking and operational tests in the first three years of VSTEP© development and operation. According to the VSTEP© specifications for reading and listening skill test provided in Table 1.2, each VSTEP© listening and reading test consists of 35 and 40 multiple-choice items respectively.

The teachers' self-reported levels of English proficiency ranged between the IELTS band scores from 5.5 to 8.5, with a mean score of 7 (SD= 0.5). Thus, the majority of the sample had their English language proficiency levels equivalent to proficient users of English according to the Common European Framework of Reference for English.

In summary, the surveyed teachers were highly qualified and experienced EFL teachers who had received both pre-service and in-service training in language testing and assessment and test item writing. They were proficient English users in the context of Vietnam where English is a foreign language. These teachers also had hands-on practice in multiple-choice test item writing given that they had been involved in the VSTEP project in the previous three years. The professional background characteristics of the sample are summarised in Table 7.1 and Table 7.2. Within the table, 'SD' indicates the standard deviation.

Table 7.1 *Professional Background Characteristics of the Sample.*

Background variables	Value	Frequency
Working location	Ha Noi	28.7% (43)
	Da Nang	14.7% (22)
	Thai Nguyen	24% (36)
	Other provinces	24.6% (37)
Gender	Male	13.3% (20)
	Female	86% (129)
Qualification	Bachelor's degree	1.3% (2)
	Master's degree	87.3% (131)
	PhD	11.3% (17)
Discipline	TESOL/TEFL	51.3% (77)
	Applied Linguistics	27.3% (41)
	Education	16% (24)
	Language testing and assessment	1.3 % (2)
With uptake of language testing and assessment course		72.7% (109)
Preparedness of the course for test item writing	Not at all	1.3% (2)
	Somewhat	22.7% (34)
	Fairly well	38% (57)
	Very well	8.7% (13)
Without uptake of language testing and assessment course		27.3% (41)

Table 7.2 *Professional Background Characteristics of the Sample.*

Background variables	Value	Range	Mean and SD
Years of EFL teaching experience		1-31	12.31 (SD = 6.65)
Years of EFL exam preparation experience	Number of years	1-19	6.81 (SD= 4.51)
Years of EFL test compiling experience		1-31	4.97 (SD = 5.52)
Number of VSTEP© tests written		0-150	20.45 (SD = 30.33)
English language proficiency	IELTS scores	5.5-8.5	7.0 (SD= 0.5)

7.2 Univariate Descriptive Statistics for Assumption Checks

In order to answer the main research questions regarding the effects of the hypothesised factors on teachers' item writing competence, stepwise regression analyses were conducted. Preceding and succeeding the statistical procedure for multiple regression, assumptions underlying linear regression models were tested. These tests included assumption checks for normality, linearity, homoscedasticity, and independence of errors within the data for both individual variables and the regression variate. The following section first presents the data screening for the dependent variable, IWCOMP and the continuous independent variables, PROLEARN and INTRAIN, which were measured by the scales constructed and calibrated under this study (see Chapter 6 for the calibration of the scales). After that, the screening of the variate derived from multiple regression analyses will be conducted through an analysis of the residual plots. As suggested by Tabachnick and Fidell (2013, p. 82), it is not required to screen all the individual variables for normality once the residuals plot shows a normal distribution.

7.2.1 Data Screening for Normality, Linearity and Homoscedasticity of Variables

For univariate normality, the estimates of means, standard deviations, and the descriptive statistics for both skewness and kurtosis were computed employing the utilities in IBM SPSS 25. The results are displayed in Table 7.3 in terms of each scale measuring the latent constructs,

mean, standard deviation (SD), skewness, kurtosis estimates, and the significance tests for both skewness and kurtosis that test the obtained value against the null hypotheses of zero, known as the *z* distribution (Tabachnick &Fidell, 2013, p. 80). The scores of the IWCOMP, PROLEARN, and INTRAIN scales were reported in plausible values in logits, which were derived from the Rasch analysis using CONQUEST 4.0 (refer to Chapter 6)

Table 7.3 *Mean, Standard Deviation, Skewness and Kurtosis Estimates.*

Variable name	Mean	SD	Min	Max	Skewness		Kurtosis	
					Est.	Z-score	Est.	Z-score
IWCOMP	-.16	.65	-1.78	1.20	-.24	-1.25	-.27	-.70
PROLEARN	-.43	1.2	-3.90	2.38	.013	.065	.017	.04
INTRAIN	.84	1.9	-5.03	5.12	-.35	-1.75	.46	1.15

The normality of each scale was examined through assessment of the kurtosis and skewness estimates illustrated in Table 7.3, together with their standardised equivalents. It has been recommended that the values of skewness and kurtosis that exceed the range between +2.58 and -2.58 (at the significance level of .01) or +1.96 and -1.96 (at the significance level of .05) would be considered to be significantly different from zero, indicating that a transformation of the scale might be necessary (Hair et al., 2010). As indicated in Table 7.3, all scales had the value of skewness and kurtosis within the recommended ranges, denoting that the scales satisfied the assumption of univariate normality.

The actual shape of the distribution for the scores of each variable can be seen in the histograms below (see *Figure 7.1*, *Figure 7.2* and *Figure 7.3*). The scores appeared to be reasonably distributed.

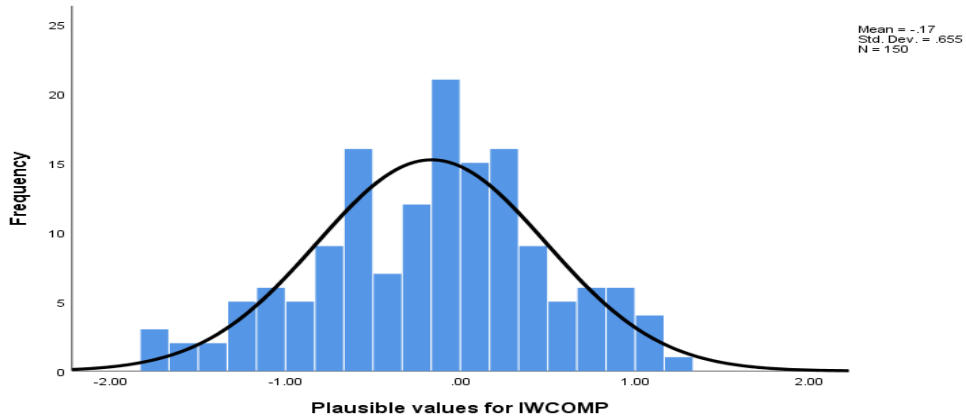


Figure 7.1 Scores distribution of the Multiple-Choice Test Item Writing Competence scale.

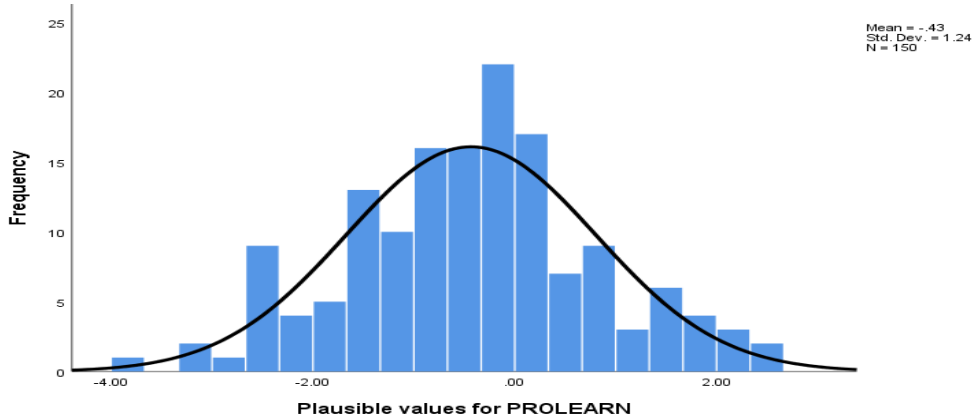


Figure 7.2 Scores distribution of the Professional Learning scale.

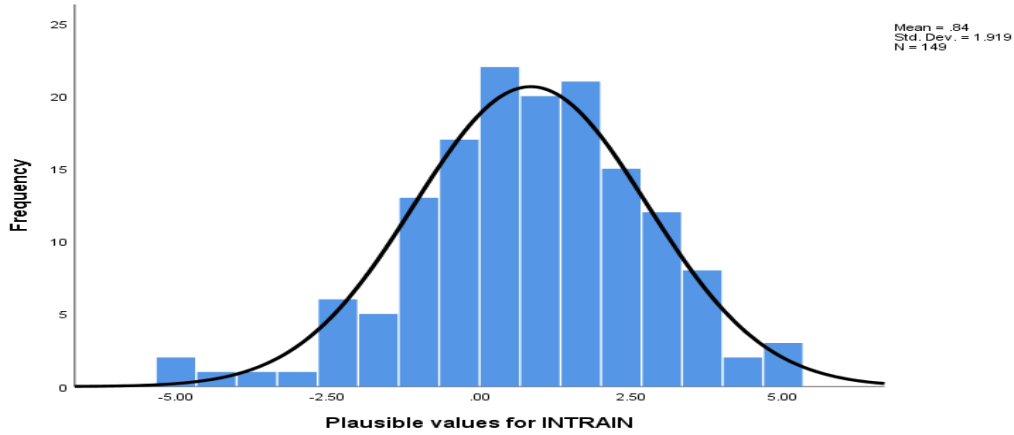


Figure 7.3 Scores distribution of the Induction and Training scale.

The normal distribution of the scores in all scales was also supported by an inspection of the expected normal probability plots (Q-Q plots) (see *Figure 7.4*, *Figure 7.5*, and *Figure 7.6*).

In these plots, the observed values for each score were plotted against the expected values from the normal distribution. As seen in the normal Q-Q plots, the scores for the cases in each of the scales fall along a reasonably straight diagonal line running from the lower left to the upper right. In the detrenched normal Q-Q plots where the linear trends from the lower left to the upper right are removed, the cases distribute themselves reasonably evenly above and below the horizontal line of zero deviation from expected normal values. This suggests that the data were normal due to the absence of major deviations from the normal (Tabachnick & Fidell, 2013).

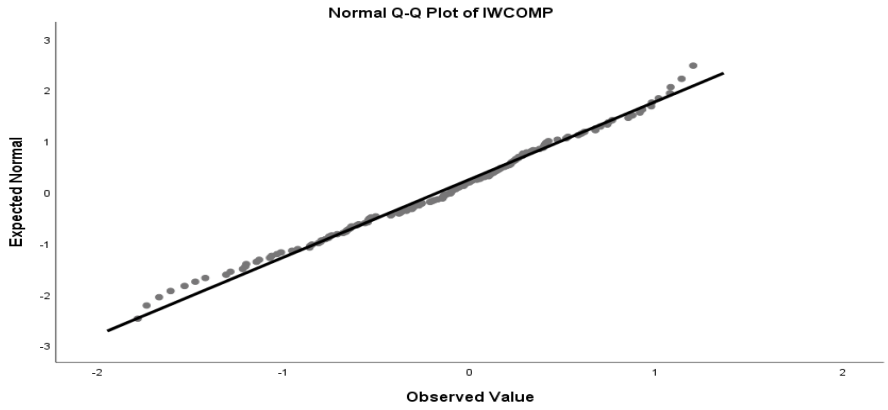


Figure 7.4 Normal Q-Q plot for the Multiple-Choice Test Item Writing Competence scale.

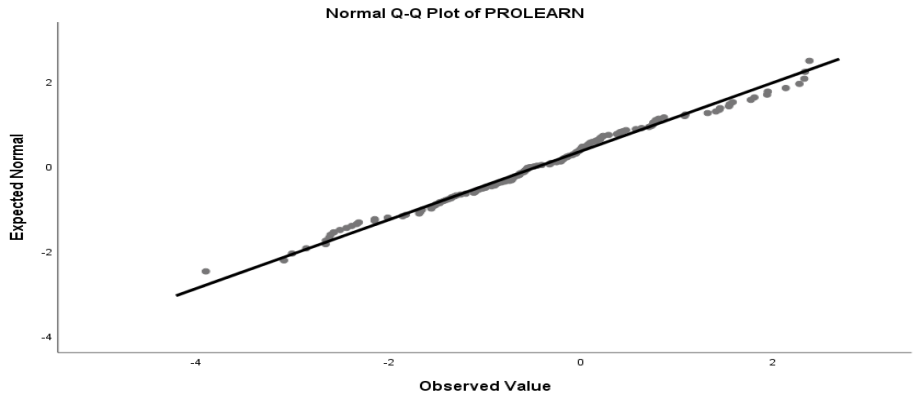


Figure 7.5 Normal Q-Q plot for the Professional Learning scale.

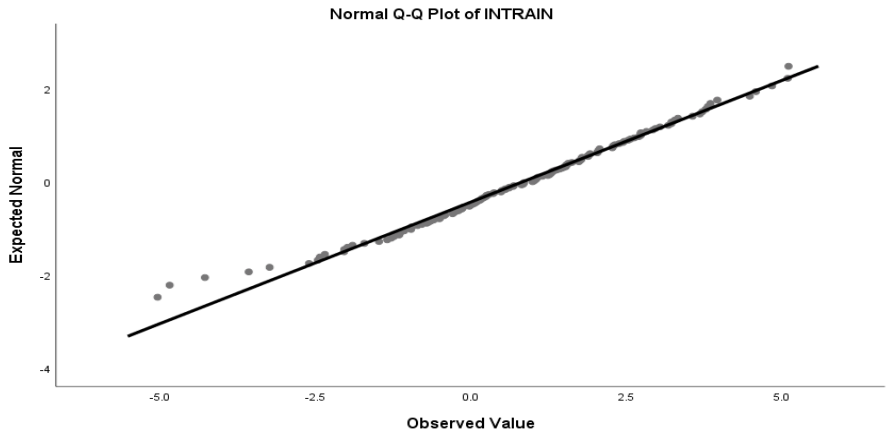


Figure 7.6 Normal Q-Q plot for the Induction and Training scale.

Linearity between the dependent variable, IWCOMP, and each of the independent variables, namely PROLEARN and INTRAIN and the linearity between PROLEARN and INTRAIN themselves were assessed roughly by inspecting their bivariate scatterplots. *Figure 7.7*, *Figure 7.8*, and *Figure 7.9* are the bivariate scatterplots as produced by IBM SPSS 25 Graph illustrating the linear relationships between IWCOMP and PROLEARN, IWCOMP and INTRAIN, and PROLEARN and INTRAIN respectively. The overall shapes of the scatterplots are oval; thus, the variables are linearly related. Moreover, the bivariate scatterplots between each pair of variables are of mostly the same width all over with some bulging toward the middle, demonstrating an elliptical distribution of scores (Hair et al., 2010, p. 74). It can also be concluded that the relationship between the variables are homoscedastic.

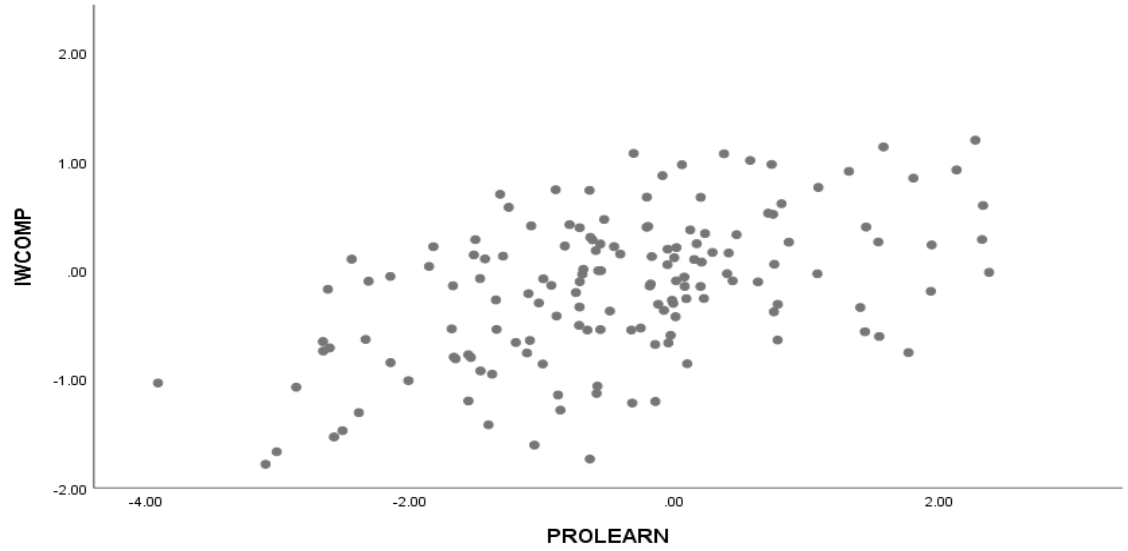


Figure 7.7 Linearity through bivariate scatterplot for IWCOMP and PROLEARN.

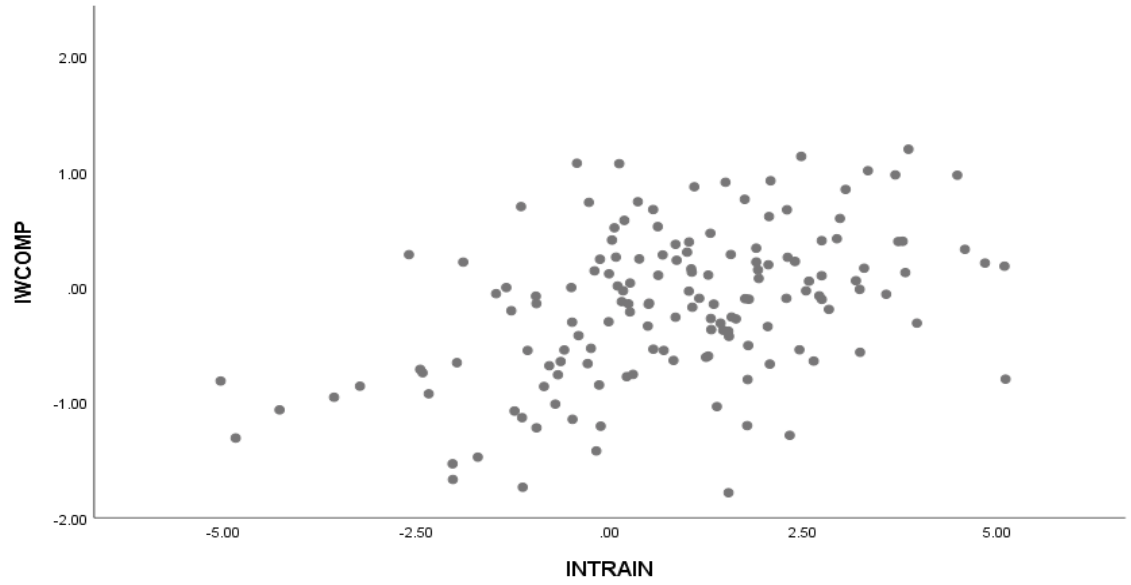


Figure 7.8 Linearity through bivariate scatterplot for IWCOMP and INTRAIN.

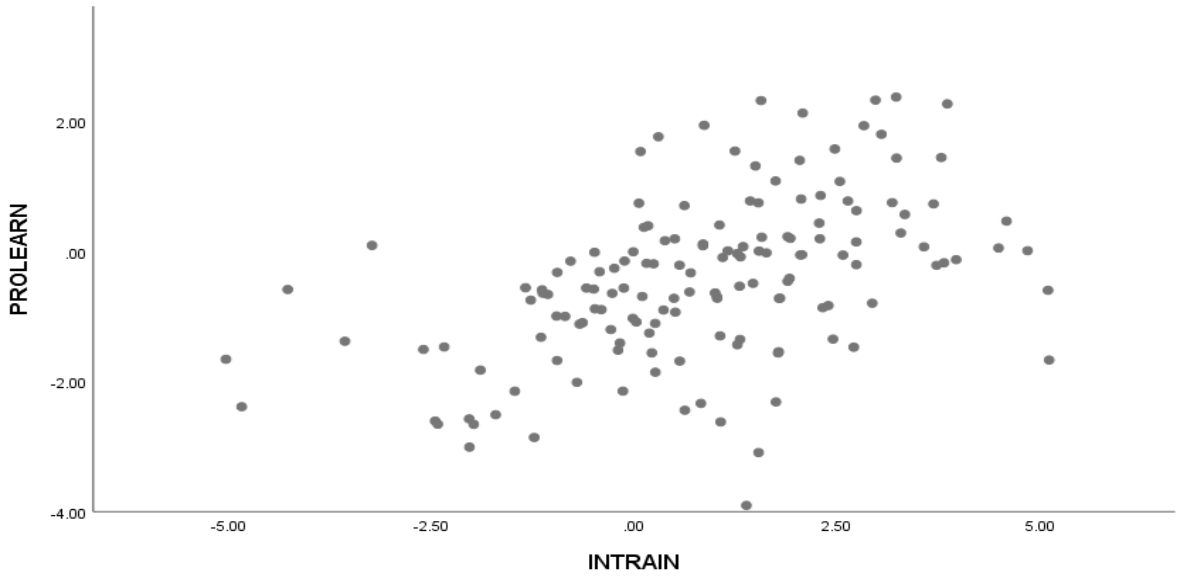


Figure 7.9 Linearity through bivariate scatterplot for INTRAIN and PROLEARN.

7.2.2 Normality, Linearity and Homoscedasticity of Residuals

The assumptions underlying multiple regression apply to both the individual variables and to the relationship as a whole. Because the variate derived from multiple regression analyses acts collectively in predicting the dependent variable, it is essential to assess the assumptions for the variate itself (Hair et.al, 2010). The normal P-P plot of regression standardised residual as illustrated in *Figure 7.10* indicates that the assumptions of normality of the error terms were met. As can be seen, the normal probability plot of the regression standardised residual points approaches a positive straight diagonal line from the left to the right validating the assumption of normality for the errors in the model (Tabachnick & Fidell, 2013). The spread of dots did not form a curve shape, denoting the relationship between the variables was linear. They did not show a funnel shape either, i.e., narrow at one end and wide at the other end. Therefore, the homoscedasticity – homogeneity of variances – assumption was also met.

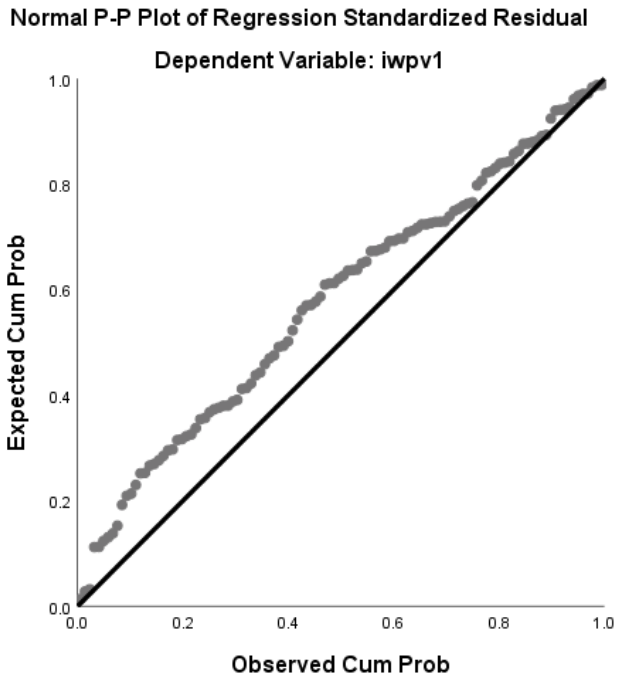


Figure 7.10 Normal P-P Plot of regression standardised residual.

An analysis of the scatter plot of the regression standardised residuals (see *Figure 7.11*) shows that most of the scores concentrated in the centre of the scatter plot (along the 0 point). This also indicates that the data were normally distributed (Tabachnick & Fidell, 2013), as seen in *Figure 7.11*. The presence of outliers can also be detected from the scatterplot. Tabachnick and Fidell (2013) defined that outliers are cases that have a standardised residual outside the +3.3 and -3.3 range. As illustrated in the scatterplot, there were not outlying residuals in this sample. This has led to the conclusion that the assumption of homoscedasticity was met, and heteroscedasticity was not an issue of concern in the regression model.

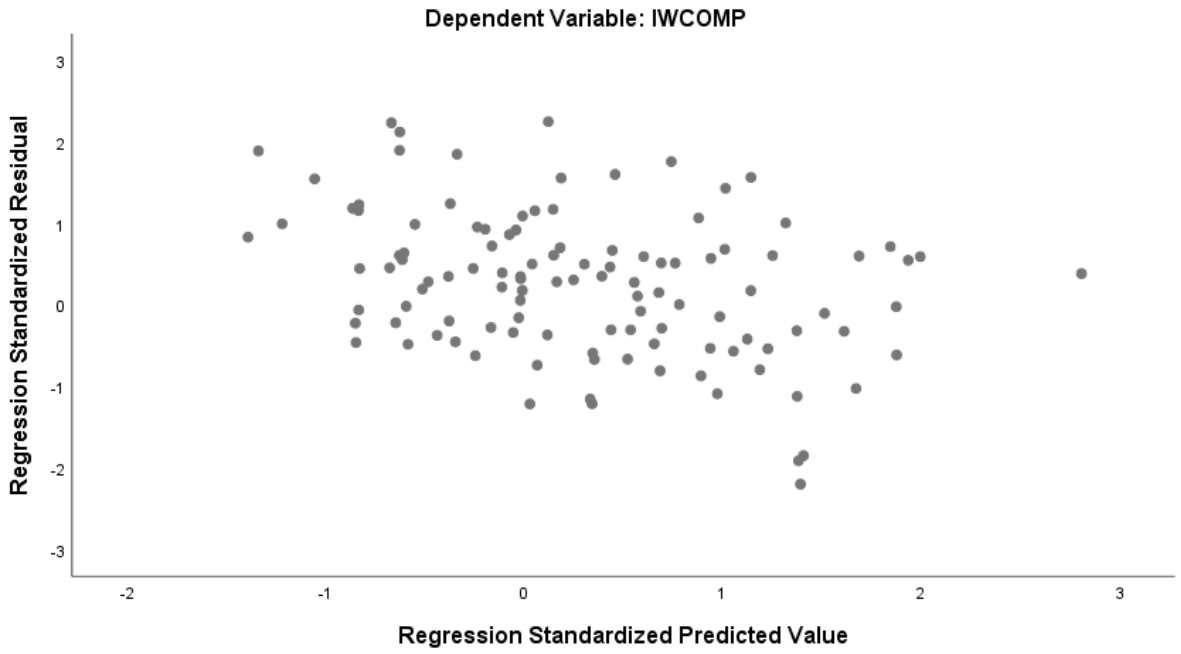


Figure 7.11 Scatterplot of the regression standardised residual.

To further test for multivariate normality of all scales, Mahalonobis distances were calculated using the Regression menu in SPSS 25 as suggested by Tabachnick and Fidell (2013, p. 74). The results showed that the maximum Mahalonobis distance obtained was 33.75. This value was larger than the recommended critical value of 27.87 (Tabachnick & Fidell, 2013, p. 952, Appendix C.4), p = .001 for nine independent variables. Further investigation to find out how many cases were involved revealed that only one person had the Mahalonobis distance that exceeded the critical value. This was the person with ID = 108. Because there was only one case, and this outlier's score was not too high, the outlying case was retained in the data file on the grounds that Kolmogorov-Smirnov and Shapiro-Wilk statistics, which assess the normality of the distribution of scores, showed a non-significant result (Sig. value of more than .05), indicating normality.

Given each variable is itself normally distributed and the relationships between pairs of variables are linear and homoscedastic, and the residuals of analysis are also normally distributed and independent, it is concluded that the assumptions of normality, homoscedasticity and linearity of the residuals in the collected data were met.

7.2.3 Independence of Errors

The assumption of independence of errors was examined by testing for serial correlations between errors by the Durbin-Watson test. According to Durbin and Watson (1951), the Durbin-Watson statistics in accordance with the sample size of 150 and the number of predictors of 9 should be closer to 2. The results show that the Durbin-Watson statistics for the regression model for prediction of EFL teacher competence in multiple-choice test item writing, which includes 9 hypothesised predictors, is 1.804, falling within the Durbin-Watson significance range of 1.608 to 1.862 (p=.05). This confirms that the residuals are uncorrelated for any two observations. Thus, the variables met the assumptions of independence of errors.

In summary, all the assumptions for both individual variables and the regression variates were statistically and graphically assessed, and it can be concluded that the tests of assumptions for both individual variables and the regression variate underlying multiple regression of all scales were met.

7.3 Bivariate Results

The second purpose of the study was to identify factors that best predict Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests. The research questions addressed in this investigation included:

What teacher-related factors predict Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

A number of propositions hypothesising the linear relationships between the dependent variable (IWCOMP) and independent variables were put forward:

1. Teachers' engagement in self-directed professional learning (PROLEARN) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.
2. Teachers' access to institutional induction and training (INTRAIN) for item writers predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.
3. Teachers' length of experience in teaching English language skills (TEACHEXP) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.
4. Teachers' length of experience in teaching English examination preparation courses (EXAMEXP) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests
5. Teachers' length of experience in constructing English tests from published resources (TESTEXP) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests
6. Teachers' repeated practice in multiple-choice test item writing (VSTEP) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.
7. Teachers' English language proficiency (ELP) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.
8. Teachers' highest degree major (MAJ) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.

9. Teachers' coursework training in language testing and assessment (LTA) predicted their judgements of their competence in multiple-choice test item writing for English reading and listening skill tests.

Of all the independent variables, IWCOMP, PROLEARN, INTRAIN, ELP, TEACHEXP, EXAMEXP, TESTEXP, VSTEP, and ELP are continuous variable and MAJ and LTA are dichotomous categorical variables. As a pre-analysis, the descriptive statistics for all predictor variables used in the study were measured and reported in Table 7.4.

Table 7.4 *Descriptive Statistics for the Nine Predictors and the Dependent Variable*

		IWCOMP	**PRO-LEARN**	**IN-TRAIN**	**LTA**	**MAJ**	**TEACH EXP**	**EXAM EXP**	**TEST EXP**	**VSTEP**	**ELP**
N	Valid	150	150	149	150	150	149	143	142	136	124
	Missing	0	0	1	0	0	1	7	8	14	26
Mean		-.16	-.43	.84	.73	.51	12.31	6.81	4.97	20.45	7.0
Std.Deviation		.65	1.23	1.91	.48	.50	6.65	4.51	5.52	30.33	0.5

At the preliminary stage of the data analysis, Pearson product-moment correlation coefficients were computed to identify the continuous independent variables that contributed significantly to the dependent variable, IWCOMP. Spearman Rho correlation coefficients were computed to explore the contribution made by the categorical variables, MAJ and LTA. The results of these analyses are displayed in Table 7.5.

Table 7.5 *Correlation Matrix for the Variables involved in the Model.*

Variables	PROLEARN	INTRAIN	MAJ	LTA	TEACHEXP	EXAMEXP	TESTEXP	VSTEP	ELP
IWCOMP	.**506**[*]	**.442**[*]	**.132**[*]	**.195**[*]	-.025	-.039	.030	**.220**[*]	**.289**[*]
PROLEARN	1	**.485**[*]	.004	.086	-.081	.003	-.029	**.184**[*]	.**206**[*]
INTRAIN		1	**.174**[*]	**.186**[*]	.069	-.041	.108	.159	.038
MAJ			1	.031	-.026	.084	.005	.048	.129
LTA				1	.104	.024	.105	.110	.114
TEACHEXP					1	.**562**[*]	**.586**[*]	.021	.029
EXAMEXP						1	**.489**[*]	.118	.**215**[*]
TESTEXP							1	.099	.048
VSTEP								1	.098
ELP									1

[*] *Correlation is significant at the 0.05 level (2-tailed).*

Statistically, the linear relationships among variables are diagnosed through numerical data denoted by *Pearson's r*., where a *Pearson's r* coefficient of +1 indicates that the correlation between two variables is perfectly positive, and -1 indicates that the correlation is perfectly negative (Tabachnick & Fidell, 2013). To provide an indication of correlation strength, Cohen (1962, 1988) suggested benchmarks for classifying correlation magnitude of r = .10, .30, and .50 as small, moderate, and large respectively, which are often referred to as effect size. Cohen's benchmarks have enjoyed an immense popularity since they have become the norm and been widely adopted in behavioural and social science research (Aguinis & Harden, 2009; Hill,

Bloom, Black, & Lipsey, 2008). However, the benchmarks have recently been challenged for their lack of empirical validity, lack of generalisability and unrealistically high minimum cutoff values (Aguinis & Harden, 2009; Hemphill, 2003). A number of updated context-specific correlation benchmarks have been developed in different disciplines. For example, in psychological research, Hemphill (2003) defined a medium correlation effect size between r = .18 and .30; that means the author set r <.18 as a small effect, .18 ≤ r ≤ .30 as medium, and r >.30 as large. These benchmarks demonstrate substantial shifts from Cohen's. Based on an empirical study of approximately 150,000 correlation effect size estimates published between 1980 and 2010, Bosco, Aguinis, Singh, Field, and Pierce (2015) (2015) found that Cohen (1988)'s benchmarks did not depict findings in behavioural research. The authors provided the new empirically derived benchmarks that define an omnibus medium distribution as lying between r = .09 and .26. That is, the minimum cut-off for a medium and large effect size is .09 and .26 respectively and r < .09 indicates a small size effect. These benchmarks were adopted in this study based on their empirical and practical significance. For this investigation, Bosco et al.'s (2015) omnibus effect size range for correlation coefficient between r = .09 and .26 was adopted. As such, a correlation coefficient r <.09 indicates a small effect, .09 ≤ r ≤ .26 as medium, and r >.26 as large. In terms of percentage of variance explained, a medium range of Pearson's coefficient between r = .09 and .26 explains 0.81% to 6.76% of the variance in the dependent variable, IWCOMP, by an IV.

Table 7.5 depicts 14 significant relationship among the variables. With regard to the relationship between the dependent variable IWCOMP and the nine hypothesised independent variables, the results reported six statistically significant relationships between the IWCOMP variable and PROLEARN (r = .506, p≤.05), INTRAIN (r = .442, p≤.05), ELP (r = .289, p≤.05), VSTEP (r =.220, p≤.05), LTA (r = .195, p≤.05), and MAJ (r =.132, p≤.05). These significant correlations implied that the more competent EFL teacher item writers were more likely to

engage in autonomous and formally offered professional learning activities. They also tended to be more proficient in English skills and have experience in writing multiple-choice test items for high-stakes tests. In addition, they tended to have had pre-service training in TESOL and language testing and assessment. The correlation coefficients of the relationships between IWCOMP and the variables TEACHEXP, EXAMEXP, and TESTEXP were relatively small, $r < .09$. This means each of these variables contributed less than 0.81% to the variance in teacher competence.

According to Table 7.5, there were 9 significant correlations among the predictor variables. First, the PROLEARN, and INTRAIN variables, which were to be measures of the teachers' professional learning were positively correlated ($r = .485$, $p \leq .05$), indicating that higher scores on the Professional Learning scale, in general, were significantly associated with higher scores on the Induction and Training scale. Thus, those teachers who tended to engage more frequently in self-regulated professional learning activities were also likely to have extensive exposure to quality, formal professional learning opportunities, that is, institutional induction and training for item writers. Secondly, the INTRAIN, MAJ and LTA variables that were assumed to measure the in-service and pre-service formal professional training were significantly related (INTRAIN – LTA $r = .186$, $p \leq .05$ and INTRAIN-MAJ, $r = .174$, $p \leq .05$). This suggests the teachers who were trained in TESOL and undertook a course in language testing and assessment as part of their highest degree study also tended to have better access to formal on-the-job professional training. Third, in relation to the variables that were thought to be measures of the teacher item writers' work-related experience, there was a positive relationship between the TEACHEXP, EXAMEXP, and TESTEXP variables. The lengths of experience in teaching English language skills (i.e., TEACHEXP variable) was found to be significantly related to those in teaching English examination preparation courses (TEACHEXP-EXAMEXP, $r = .562$, $p \leq .01$), as well as those in constructing English tests from published test books and

other resources (TEACHEXP-TESTEXP, r = .586, p≤.01). This implied that the teachers who were trusted to be competent in multiple-choice test item writing seemed to be those with extensive experience in both English language teaching and testing and assessment. Fourthly, the correlation coefficient results also revealed a significant relationship between the teachers' experience in multiple-choice test item writing measured by the number of VSTEPs they had written (i.e., VSTEP variable) and their engagement in self-regulated professional learning (PROLEARN-VSTEP, r =.184, p≤.05); so too was their English language proficiency level (PROLEARN- ELP, r =.206, p≤.05). The examination of the correlation matrix shows that there was no correlation coefficient of any two IVs of between .70 and .90, indicating no substantial collinearity or multilinearity (Field, 2018; Hair et al., 2010; Tabachnick & Fidell, 2013).

7.4 Multivariate Results

As the nine independent variables hypothesised to be involved in the prediction of the dependent variable, IWCOMP, were proved to meet the requirements of multiple regression, a standard multiple regression was run. Table 7.6 displays the standardised and unstandardised Beta (β) coefficients of each of the nine independent variables, their significance level, and their standard error.

Table 7.6 *Regression Coefficients with all 9 Independent Variables.*

	Regression Coefficients			
Predictor variables	Unstandardised B weights	Standard Error	Standardised Beta (β)	Sig.
PROLEARN	.165	.047	**.313**	**.001**
INTRAIN	.091	.031	**.267**	**.004**
LTA	.131	.113	.089	.249
TEACHEXP	.002	.010	.022	.830
EXAMEXP	-.018	.014	-.124	.206
TESTEXP	.003	.011	.029	.764
VSTEP	.002	.002	.094	.224
ELP	.232	.092	**.200**	**.013**
MAJ	.195	.100	**.149**	**.054**

The standardised Beta (β) coefficients provided the comparative contributions of each independent variable. The largest beta coefficients were .313 for PROLEARN (p =.001), and .267 for INTRAIN (p= .004). The β values for ELP and MAJ were slightly lower, ELP (β=.200, p =.013), and MAJ (β=.149, p =.05). The significant positive β weights for the predictor variables, PROLEARN and INTRAIN indicate that teachers with higher scores in the PROLEARN and INTRAIN scales showed higher scores in the IWCOMP scale. Specifically, if there was a one standard deviation unit change which was 1.23 in the predictor variable, PROLEARN, the IWCOMP score would be likely to increase by a .313 standard deviation unit. Similarly, one standard deviation unit change of .192 in the INTRAIN would equate to a rise of a .267 standard deviation unit in the item writing competence score. This means that if the teachers were more engaged in self-regulated professional learning and/or had access to high-quality induction and training for item writers, the competencies in test item writing was more likely to be enhanced. The β values for the other variables being low with the significance value greater than .05 meaning that these variables were not making a significant contribution to the

prediction of the teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests.

Multiple regression analysis conducted by SPSS 25 produced the overall model fit results when the dependent variable, IWCOMP, and all the nine hypothesised independent variables were simultaneously entered into the regression equation. In this regression process, each of the IVs is assessed as if it entered the regression after all other IVs had entered. Table 7.7 summarises the overall model fit indices which include the multiple correlation coefficient between the IVs and the DVs, *R*, the correlation coefficient squared, R^2, often referred to as the coefficient of determination, adjusted R^2, and the standard error of the estimates.

Table 7.7 *The Overall Model Fit Results from Standard Multiple Regression Analysis*

Multiple R	Coefficient of Determination (R^2)	Adjusted R^2	Standard error of the estimate
.630	.397	.348	.529

Table 7.7 shows that the overall regression models predicting test item writing competence from the independent variables were statistically significant, $R^2 = 0.397$, $F(9,109) = 7.98$, and $p = .000$. The R^2 value of .397 indicates that approximately 40% of the variability in the dependent variable, IWCOMP, was predicted by the hypothesised independent variables. In other words, teachers' self-assessment of their competence in multiple-choice test item writing for English reading and listening skill tests was predicted by the amount of formal and informal professional learning they were engaged in, their experience in test item writing, their highest degree major in TESOL, their English language proficiency, and their uptake of language testing and assessment course. The F ratio is different from zero and greater than 1 ($F = 7.985$), indicating that the results of this regression model can be applied in other similar samples from the population and the results found can be applied in other similar samples.

In order to identify the unique contribution of variance of each independent variable, a standard regression analysis was run. In this analysis, one independent variable was removed from the equation at a time, and the R^2 change was observed. Table 7.8 summarises the results of this regression analysis. The first column presents ten different regression models being tested. In Model 1, the DV and all the nine IVs were simultaneously entered into the equation. In each of Models 2, 3, 4, 5, 6, 7, 8, 9 and 10, one of the IVs was taken out of the equation. The second column shows the variable being removed in the corresponding model. The next four columns display the model fit indices, namely *R*, R^2, adjusted R^2, and the standard error of the estimate. The last column reports the amount of change in the R^2 value of the target model in comparison with the R^2 of Model 1.

Table 7.8 *The Fit Indices of 10 Regression Models*

Model	Variable removed	*R*	R^2	Adjusted R^2	Std.Error of the estimate	R^2 Change
1	None (all 9 IVs entered)	.630	.397	.348	.529	
2	TEACHEXP	.630	.397	.353	.526	.000
3	TESTEXP	.630	.397	.353	.527	.000
4	LTA	.624	.390	.346	.530	.007
5	VSTEP	.624	.389	.345	.530	.008
6	EXAMEXP	.623	.388	.344	.530	.009
7	MAJ	.614	.376	.331	.535	.021
8	ELP	.602	.362	.320	.540	.035
9	INTRAIN	.591	.349	.302	.547	.048
10	PROLEARN	.574	.329	.280	.555	.068

According to Table 7.8, in Model 2 when TEACHEXP was removed from the equation, there was no R^2 change recorded. The amount of change in R^2 when the other variables related to lengths of experience, namely TESTEXP, EXAMEXP, VSTEP were removed, was minimal,

with R^2 change ranging from .000 to .009. The most significant R^2 change was detected when PROLEARN or INTRAIN was taken out of the equation, R^2 change = .068 and .048 respectively. This demonstrates that PROLEARN uniquely contributed approximately 7% to the variability of the dependent variable, IWCOMP, and INTRAIN made a unique contribution of roughly 5% to this predictive power. MAJ and ELP also made marked contributions to the variance of the dependent variable with substantive R^2 changes of .021 and .035 correspondingly.

It could be concluded from the multiple regression analyses that the best set of variables for predicting Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests were teacher engagement in self-regulated professional learning (PROLEARN), teachers' access to institutional induction and training (INTRAIN), their English language proficiency (ELP) and pre-service training in TESOL (MAJ).

The validity of the results was assessed by an assessment of the adjusted R^2. Examining the adjusted R^2 value revealed that when compared to the R^2 value (.348 versus .397, see Table 7.7), there was little loss in predictive power. With four variables in the model, the R squared value maintains a sufficient ratio of observations to variables in the variate. Besides, it was calculated that the predicted residual sum of squares (PRESS) statistics was 33.78, demonstrating a relative similarity to the predictive fit of the original regression model (R^2 = .397, see Table 7.7). This predictive fit indicates that the regression results can be generalised to the population of Vietnamese EFL teacher item writers under the assumption that the sample is randomly selected from the population.

CHAPTER 8: CONCLUSIONS AND IMPLICATIONS

This chapter is divided into three sections. The first section reviews the rationale for the research and the methodology. The second section discusses the results and addresses the research questions and propositions. The third section provides the implications of the research for theory, methodology, practice and the training of item writers.

8.1 An Overview of the Study

8.1.1 Revisiting the Rationale and Aims of the Study

The primary purposes of the study were to operationalise Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests as a developmental progression and to determine the factors that predicted the development of this competence. The investigation was driven by the nationwide demands for a workforce of English test item writers. The Vietnamese tertiary sectors had a pressing need for research-based evidence to inform their identification, recognition and development of competent English test item writer teachers who are responsible for developing items for large-scale, high-stakes English proficiency tests (i.e., VSTEP®) that were locally created in accordance with the test specifications and administration requirements of some English international tests such as IELTS or TOEFL iBT. This pressing demand from the local context was the main inspiration for this investigation into Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests.

The research was also motivated by the recognition of the need for scholarly work in two main aspects of test item writing competence: the identification and definition of competency development in test item writing, particularly for EFL teachers; and the identification and measurement of factors affecting this competence. Understanding these aspects is important for the effective professional development of English test item writer teachers.

A thorough review of the literature revealed the scarcity, possibly non-existence, of prior research pertaining to defining the progression for competence in multiple-choice test item writing for English reading and listening skill tests. A few case studies of item writers such as those by Fulkerson and colleagues (2009, 2010, 2011), Salisbury (2005), and Kim et al. (2010) have contributed a solid understanding of test item writing competence. Nevertheless, these studies did not enhance our understanding of how test item writing competencies were typically acquired and developed. The literature on test item writing competencies was scattered across several case studies focusing on a few individual item writing competencies (see Green & Hawkey, 2011; Kim et al., 2010; Lynch & Davidson, 1994; Sydorenko, 2011). These studies aimed at distinguishing between novice and expert practice, limiting the competence into dichotomous levels, and were not able to capture the progression of these competencies as item writers transitioned from novice to expert.

Knowledge about item writing skill development is helpful for educative and professional development purposes. A large part of the literature was replete with an indefinite number of manuals, guidelines and handbooks for item writers and best practices in designing, developing and conducting tests and in constructing test items. These products for test item development are expressed more in terms of guidelines, principles and procedures to be followed as opposed to developmental tools that promote acquisition and development of test writing expertise. Hence, prior to this study, there was an absence of empirical research that offered a holistic insight into test item writing competencies, and there was a lack of research into developmental pathways for acquiring these competencies. There was therefore a need to operationalise these competencies on a continuum in light of skill acquisition and developmental expertise models so that item writers' acquisition of and development in these competencies could be better understood.

It seems that a common understanding of the acquisition and development of EFL teachers' item writing capabilities and a shared voice to foster professional discussions and reflections on EFL test item writing practices and improvement was much needed. The study aspired to contribute to the scholarly literature a more thorough understanding of how test item writing competencies could be acquired and developed. It filled this literature gap by formulating and validating a developmental progression that was made up of a set of comprehensive and coherent descriptors articulating in observable terms the core constituents of teacher competence in multiple-choice test item writing for English reading and listening skill tests and illustrating exemplars of the competence in practice along a developmental progression from novice to expert.

The other rationale for this study arose from a lack of research that has adequately and empirically measured and examined factors influencing teacher competence in multiple-choice test item writing for English reading and listening skill tests. This inadequacy and scarcity have imposed constraints on the professional development of item writers and understanding of factors that individually and in concert contribute to English test item writing competence. Drawing on previous research on expertise development and teacher competence growth, this study proposed three groups of factors that predicted teachers'self-perceived competence in multiple-choice test item writing for English reading and listening skill tests, namely, professional learning, and educational and professional background characteristics. It was hypothesised that teacher engagement in self-regulated professional learning and institutional induction and training were two primary factors affecting their competence in EFL test item writing. The self-regulated professional learning activities that teacher item writers were supposed to effortfully and mindfully engage in are those that meet the requirements prominently found in theories and studies of deliberate practice (Ericsson et al., 1993, 1996), self-regulated learning (Pintrich, 2000; Zimmerman, 2000), reflective learning and experiential learning (Kolb, 1984; Schon, 1987). The description in the contemporary related literature of

induction and training offered to the item writers by their institution bear the critical features of high-quality professional development. These factors chosen for this current study emerged from the social theories of learning, which help explain the development of competence in terms of acquirable or malleable factors. These factors might well facilitate and/or limit the role of innate capacity, individual giftedness and affective characteristics. The study also investigated other educational and professional background factors as they played a decisive role in the recruitment of item writers in Vietnam universities and they were helpful in understanding the practice of English test item writers in the context where English is a foreign language.

The primary and subsidiary research questions were:

1. What is the typical developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests?

2. What teacher-related factors predict Vietnamese teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.1 To what extent does professional learning *(i.e., self-regulated professional learning and access to institutional induction and training)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.2 To what extent do teachers' professional background characteristics *(i.e., length of experience in teaching English language skills, teaching English exam preparation courses, constructing English tests from published resources, and writing multiple-choice test items)* predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

2.3 To what extent do teachers' educational background characteristics **(***i.e., their highest degree's major, uptake of coursework in language testing and assessment and their*

English language proficiency) predict the teachers' judgements of their competence in multiple-choice test item writing for English reading and listening skill tests?

8.1.2 A Review of Methodology.

A developmental progression scale for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests was developed based on extensive examination of the literature, focus group discussions with item writers, workshops with subject matter experts (SMEs) and expert review. Two scales to measure teacher engagement in self-regulated professional learning, and access to institutional induction and training were adapted from prior research. The items from the scales and a set of background questions were later embedded into the teacher survey instrument. The data generated were used to empirically calibrate the scales designed using IRT modelling procedures. After that, multiple regression analyses were conducted to explore the best predictors for Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests.

8.1.2.1 Scale Development Procedures

Detailed explanation of the development process of all the scales involved in this study is provided in Chapter Five (see Sections 5.1.1, 5.1.2). The multiple-choice test item writing competence developmental progression (the IWCOMP scale) was designed to elicit self-reported statements of Vietnamese EFL teachers' test item writing competence, which functioned as the dependent variable in the study. The other two scales, the self-regulated professional learning scale (the (PROLEARN scale) and institutional induction and training scale (the INTRAIN scale), were adapted from available measures of professional learning used in prior research (see Chapter 4, Section 4.1) and were contextualised to be relevant to the profession of EFL teacher item writers. These scales were used to elicit teachers' frequencies of engagement in self-

regulated professional learning and teachers' indication of the presence and accessibility of the induction and training opportunities provided to them by their institutions.

The basic approach to the development of the IWCOMP scale and subsequent calibration of this scale was based on a combination of both theoretical and psychometric approaches to scale development (Gillis, 2003) (see Chapter 5, Section 5.1.1.1 for the development of the scale and Chapter 6, section 6.1.1 for the empirical calibration of the scale). Adopting the developmental, multiple methods approach to competency framework development, multiple types of evidence such as theoretical models, standards, and guidelines on test item writing, research results, expert opinions and psychometric measures were employed to create the scale, which conforms to the unified notion of validity put forward by Messick (1985, 1995) and adopted by the Standards for Educational and Psychological Testing (AERA et al., 2014). The items within the progression were drafted by the researcher based on an extensive examination of the literature in test development and test item writing. The items were subsequently scrutinised at focus groups and workshops with Vietnamese English teacher item writers. The items were further refined through expert reviews before they were used to collect self-reported responses from English teacher item writers in Vietnam and calibrated through Item Response Modelling procedures.

The PROLEARN scale was adapted from the measure of teacher engagement in professional learning in the workplace developed by Kwakman (2003) and Evers et.al. (2016). The Cambridge Assessment English's RITCME framework for training and development of item writers was employed to construct the items for the INTRAIN scale in reference to the critical features model of professional training synthesised by Desimone (2009) and Darling-Hammond and Richardson (2009), and the Observation Checklist for High-Quality Professional Development Training Checklist developed by Erickson et al. (2017) (see Chapter 5, Section 5.1.2).

8.1.2.2 Data Collection Procedures

The developmental progression developed under this study was used for Vietnamese teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests. The two scales, teacher engagement in self-regulated professional learning and institutional induction and training, were adapted and designed to measure two major latent factors contributing to this competence. A set of professional and educational background questions was also developed. The items from the three scales and the background questions were subsequently collated into a questionnaire.

Emails with the link to the online questionnaire were sent to 239 teachers from tertiary institutions across Vietnam who were identified as having been involved in writing multiple-choice items for VSTEP® reading and listening skill tests.

One hundred and fifty completed responses were received, yielding a return rate of 63%. The respondents were highly qualified and experienced EFL teachers who had received pre-service and/or in-service training in language testing and assessment and test item writing. They were proficient English users in the context of Vietnam, where English is a foreign language. These teachers also had hands-on practice in multiple-choice test item writing given that they had been involved in the VSTEP® project since 2015. Details of the sample were described in Chapter 7.

The participants completed four parts of the questionnaire. For the first part, which was a 14-item scale, the participants were asked to self-report their multiple-choice test item writing competence by selecting the options that best described their typical behaviours in their item writing practice. In the second part, the participants rated their engagement in four categories of self-regulated professional learning by completing 15 items using a 4-point rating scale ‘never, rarely, sometimes, and often’. In the third part, the participants reported on the induction and training provided to them by their institutions, using a 2-point Likert scale “Yes and No” on 21

items. In the final part of the questionnaire, the participants were requested to supply professional and educational background information regarding their highest academic qualification, pre-service training, length of work-related experience and English language proficiency.

8.1.2.3 Data Analysis Procedures

The scales were calibrated using Item Response Theory modelling and proved to have acceptable reliability and construct validity as described by AERA et al., (2014), Messick (1989, 1995), Wright and Masters (1982), and Wu and Adams (2007) (see Chapter 6).

Multiple regression analysis procedures were employed using SPSS 25 to analyse the correlation between the dependent variable, *Multiple-Choice Test Item Writing Competence* (IWCOMP), and nine independent variables and to identify the best predictor variables for IWCOMP (see Chapter 7). The nine independent variables were:

- *Professional learning factors*
    - *Teacher Engagement in Self-regulated Professional Learning* (PROLEARN);
    - *Teacher Access to Institutional Induction and Training* (INTRAIN);
- *Professional background factors:*
    - *Length of Experience in Teaching English Language Skills* (TEACHEXP);
    - *Length of Experience in Teaching English Examination Preparation Courses* (EXAMEXP);
    - *Length of Experience in Constructing English Tests from published resources* (TESTEXP);
    - *Length of Experience in Writing Multiple-Choice Test Items (VSTEP)*
- *Educational background factors:*

- *Highest Degree Major in TESOL* (MAJ);
- *Uptake of Coursework Training in Language Testing and Assessment (*LTA);
- *English Language Proficiency* (ELP)

8.2 Key Findings and Discussion

8.2.1 The Developmental Progression for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests

In general, the study has met the proposed aim to develop and validate a developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. In terms of a developmental sequence for the test item writing competence, the use of IRT enabled the construction of quality criteria for each of the 14-test item writing skills, consisting of three to four observable descriptions of increasingly competent behaviours. The 14 skills represent what item writers do, say, create and make to indicate their possession of the five multiple-choice test item writing competencies, namely, analysing test specifications, generating input materials for tests, writing test item stems and distractors and reviewing and revising test items. The corresponding 45 quality criteria depict varying levels of performance quality that are observable and ordered in increasing proficiency. Teacher item writers were able to use the scale to self-assess their multiple-choice test item writing competence by selecting the options that best describe their typical behaviours in item writing practice. The modelling of item responses on an underpinning developmental framework for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests ensured results could be employed to support next steps in professional development. This is because teachers' competencies were located along the developmental progression. Thus, the developmental progression is of practical usefulness for teachers' own professional development purposes. For example, if an item writer self-assesses his/her

competence in multiple-choice test item writing for English reading and listening skill tests and emerged as a novice item writer, s/he could refer to the description of the next level of competence development to identify typical item writing behaviours demonstrated at this level. It is suggested that future research could work with item writer trainers to develop scaffold statements that inform the design, adaptation and/or adjustment of professional learning experiences, opportunities and choices to meet the needs of individual item writers and to support them in their professional practice and their development to the next level of competence.

In terms of the appropriateness of item difficulty and type for the target group, the scale was found to target the sample of teachers well in terms of matching difficulty levels of the full spectrum of teacher competence and provided a range of types of item difficulty at each level. Teachers of varying levels of competence were able to achieve across the scale. The results of the calibration for the finalised 14-item progression scale generated from both Classical and Rasch analyses have shown that the scale had satisfactory psychometric properties. The overall good fit statistics and the item-rest discrimination estimates of the items indicated that there was one dominant construct of EFL multiple choice test item writing, and that the items worked well together to measure the underlying construct. Overall, the items formed a cohesive scale representing "Vietnamese EFL teacher competence in multiple choice test item writing". The item separation reliability coefficient (r=0.92) suggested that the items and categories were well separated in terms of difficulty, and the chi-square test of item parameter equality was significant (p= <0.05), indicating that the items were significantly different from each other in difficulty. From these results, it can be concluded that the items in the item writing competence progression scale can be divided into different levels of proficiency. Judging from the person separation reliability indices (r=0.68) and the expected a posteriori/plausible value (EAP/PV) reliability index of 0.70, the scale has moderately acceptable reliability for distinguishing

teachers' underlying abilities. The person-item threshold map indicates there was a good match between the teacher's competence levels (3.79 logits) and the range of behavioural descriptors' difficulty values (3.59 logits), suggesting that the scale has the appropriate level of difficulty to assess this group of teachers.

The performance quality criteria within the developmental progression were grouped into clusters according to their estimated difficulty. A content analysis of the group of items was performed to explore the common underlying theme or meaning that emerged from its quality criteria and to summarise that meaning into a coherent description. Content analysis of criteria within each identified cluster resulted in development of summary descriptions of each level along a continuum of proficiency (See Chapter 6, Section 6.1.1). In summary, the items within the progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests fitted the model well and were also well separated along the continuum to differentiate between levels of item writing proficiencies.

The development and validation of the progression scale have demonstrated the unitary notion of validity proposed by Messick (1989, 1995) and in accordance with the Standards for Educational and Psychological Testing (AERA et al., 2014). The content and structural validity requirement for the progression scale was supported by grounding the conception of multiple-choice test item writing competence in domain theory (as depicted in Figure 3.4), aligning the items within the scale to the reviewed literature (see Appendix E), and engaging item writers and subject matter experts in the focus groups, workshops and utilising expert review during the development process of the scale. The development and calibration process of the scale which were thoroughly documented in Chapter Five (see Section 5.1.1.1) and in Chapter Six (see Section 6.1) also constitute a validity argument. The Rasch model supports a unified concept of validity grounded in construct validity in terms of separating the item writers and the items along a defined variable. The results of the calibration for the progression scale generated from both

classical and Rasch analyses showed that the scale had satisfactory psychometric properties. Specifically, the scale has the good fit statistics, high item separation reliability coefficient (r=0.92), moderate person separation reliability index (r=0.68) and the expected a posteriori/plausible value (EAP/PV) reliability index of 0.70. Furthermore, there was a good match between the teacher's competence levels (3.79 logits) and the range of behavioural descriptors' difficulty values (3.59 logits). These figures gave support for the construct validity of the scale. The unidimensional construct of multiple-choice test item writing competence, which was initially derived from the scholarly literature, was sequentially validated through the Rasch analysis. The fit statistics for the progression scale (see Table 6.3) indicated that the data fitted the Rasch model specifications, supporting that the multiple-choice test item writing competence was a unidimensional construct. The hypothesised developmental progression for Vietnamese EFL teacher competence in multiple -choice test item writing developed based on the Dreyfus and Dreyfus model of skill acquisition was further validated through the data from the Rasch analysis. The data indicated that the items within the progression scale were well separated along the continuum to differentiate between levels of item writing proficiencies as described in the hypothesised progression guided by the Dreyfus and Dreyfus model of skill acquisition (Dreyfus & Dreyfus, 1986) (see *Figure 6.4*). This made an additional contribution to the validity argument. Thus, the progression scale developed and validated in this study for Vietnamese teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests has acceptable reliability and has demonstrated the unified concept of validity as described in Messick (1989, 1995) and AERA et al., (2014).

The developmental progression calibrated in the study depicts a path to competence in multiple-choice test item writing for English reading and listening skill tests by the teachers in the sample as described in the Dreyfus and Dreyfus (1986) construct of a continuum of skill acquisition and expertise development. It is a progression moving through four distinct stages

from being novice and emergent item writers who follow rules and instructional maxims, to established and proficient test item writers, who are able to execute skilled behaviours based on holistic recognition of tasks and an accumulation of concrete experiences (Dreyfus & Dreyfus, 1986, p. 35). The holistic levels of Vietnamese teachers' competence in multiple- choice test item writing for English reading and listening skill tests is presented in a single dimension 'rocket' chart (Griffin, 1995, p. 108) in *Figure 8.1*. The 'rocket' chart illustrates four levels of Vietnamese EFL teacher competence in multiple-choice test item writing and a gist description of the overall skill of each development stage. *Figure 8.2* presents the level statements of the five multiple-choice test item writing competencies demonstrated by Vietnamese EFL teachers.

Figure 8.1 A single-dimension 'rocket' chart of Vietnamese EFL teacher competence in multiple-choice test item writing.

1. Analysing test specifications	2. Generating materials for tests
Level 4: The item writers develop their own internal guideline for item writing by synthesising the information from the test specifications, and other training materials, reflecting on prior test item writing experiences, and knowledge of target examinees.	**Level 4:** The item writers source input materials from diverse sources with the characteristics of the target examinees as English language learners and their cultural background in mind. They judge the relevance of the input materials by using multiple sources of evidence such as text analysis tools, established lists of defined characteristics of texts/scripts contents and difficulty levels and their common sense of text potentiality for item development. They transform input materials to fit the test purposes and test item requirements by implementing a range of strategies such as amending and editing texts/scripts by deleting irrelevant elements, adding supplementary text/scripts, cutting redundancies and repetitions, using more or fewer synonyms/antonyms to adjust text/script complexity and lexical density.
Level 3: The item writers tend to refer to externally provided resources such as training materials and sample tests to clarify the test specifications.	**Level 3**: The item writers source input materials from previously used high-yield sources. They evaluate the input materials based on am established list of defined characteristics of text contents and difficulty levels, text analysis tools. They make lexical and syntactic changes to achieve the target register required of the text or script. They make further changes to the input texts/scripts by using a range of strategies such as cutting, restructuring, removing or glossing part/parts of the texts/scripts, deleting irrelevant elements, and editing part/parts of the texts/scripts to reduce or increase level of difficulty or/and complexity.
Level 2:	**Level 2:** The item writers identify texts of specified complexity and difficulty from newspaper, magazines and books written for non-native English readers. They use research-grounded factors such as use of paraphrase, antonyms and synonyms to avoid dictation effect, add distraction to texts/scripts to reduce guessability, word-form shift, and supplement second texts to adjust texts/scripts to fit items.
Level 1: The item writers pay attention to mandatory aspects of the test such as text length, text type, item formats and quantity.	**Level 1:** The item writers look for input materials that meet the requirements in terms of length and topic, and make minimum changes to the generated materials. They employ test analysis tools such as LexTutor, English Profile, Natural Language Processing, British National Corpus, Academic Word List to determine if the materials fit the test specifications in terms of level of difficulty and complexity.

3. Writing test items	4. Reviewing test items	5. Revising test items
Level 4: The item writers multi-task between text/script adaptation and item construction. They build on various resources to innovate items and utilise strategies flexibly to write distractors with consideration of typical errors or misconceptions of the subject content held by examinees. They draw on a range of factors affecting item difficulty from experience, expert colleagues and contemporary research to make judgements of item difficulty. They provide justifications for both correct and incorrect answer options.	**Level 4:** The item writers make judgements of item-fits-specifications for content complexity and difficulty levels by attempting items on their own.	**Level 4:** The item writers use information from experience and research to justify decision to implement or not to implement received feedback.
Level 3: The item writers make use of multiple resources such as the test specifications, past exam papers, example of good items and established guidelines for multiple-choice test item writing to create items accommodating the test specifications and input materials. They utilise paraphrasing strategies (e.g., use of antonyms/synonyms, word-form shift, word order, and referents) to construct item distractors. They justify the answer key by providing references for each answer (e.g., underlining chunks of texts/scripts that indicate the answer).	**Level 3:** The item writers make subjective judgements of the required level of item difficulty by considering the text complexity and cognitive demand/subskills required to answer each item. They analyse responses from a self-conducted pilot with a sample of students to identify any miskeys and/or poor performing distractors as a way of self-validating their test items.	**Level 3:**
Level 2: The item writers follow the item-first approach to item writing; they focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items. They construct items from a pool of item shells/item stems derived from established items in past test materials.	**Level 2:**	**Level 2:** The item writers begin to make connections between the test items and test purposes when making decisions to implement feedback.
Level 1: They typically follow the standard text/script-first approach; they adhere to the input materials to write items and distractors which resemble example item stems and distractors from supplied sample tests following the Dos and Don'ts guideline for writing multiple-choice test items.	**Level 1:** The item writers examine the conformity of the item content and format to the test specifications and check item difficulty against the lexical and grammatical descriptors in the six-level English language proficiency framework for Vietnamese learners. The item writers try out items on own to check the original keys.	**Level 1**: The item writers tend to incorporate all the item feedback they receive from their colleagues and/or mentors.

Figure 8.2 Level statements of five multiple-choice test item writing competencies demonstrated by Vietnamese EFL teachers.

8.2.2 Factors Predicting Vietnamese Teachers' Judgements of Their Competence in Multiple- Choice Test Item Writing for English Reading and Listening Skill Tests.

8.2.2.1 The pre-eminence of professional learning.

The results indicated that teacher engagement in self-regulated professional learning and institutional induction and training about item writing were positively and significantly associated with Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests. That was possibly that the Vietnamese EFL teacher item writers who engaged more frequently and purposefully in professional learning about test item writing at both individual and community level achieved higher scores in the test item writing competence. It could also mean that these item writers were better aware of what item writers are supposed to do, so they were better able to reflect that in their responses to the items in the item writing competence scale. It may have also appeared to these item writers that they perceived themselves as better item writers because they had done more professional training. Further research would be necessary to disentangle the proposition. It could include uses of objective measures such as an evidence-based portfolio of practice, third party reports, interview procedures and direct observation and record keeping in assessing the item writers' level of item writing competence and their engagement in professional learning. This finding is especially pertinent to the findings of this study, where professional learning factors were significantly predictive of the teachers' item writing competence. The finding from the current study were aligned with previous research that explored factors contributing to expertise development and the influence of deliberate professional learning on teacher professional competence. For example, Hatano and Inagaki (1986), Ericsson et al. (1993), and Bransford and Schwartz (2009) consistently reported that professionals evolved and became more competent when they constantly endeavoured and were intrinsically motivated to enhance their knowledge and skills by engaging in deliberate and self-regulated professional learning. The results from

both Hatano and Inagaki's (1986) and Ericsson et al.'s (1993) study emphasised that individual professionals improved by exploiting available resources, such as reading books, articles, and other materials from the related literature, seeking support and consultation from colleagues, experimenting with new ideas and reflecting on their own practice. Desimone (2009) added that these activities are embedded in teachers' everyday work lives and can be engaged with independently at their own pace and will. Furthermore, Bransford and Schwartz (2009) also asserted that expertise development was shown to be most successful in a community of practice where professionals were encouraged and offered opportunities to learn from each other, and to develop continually. In such a community, teachers upgrade their skills and knowledge by participating in discreet and formal training workshops in which professional knowledge and skills are transferred to them by experts (Feiman-Nemser, 2001). The findings from the current study supported the notion that "it takes expertise to make expertise" (Bransford & Schwartz, 2009, p. 436). The notion emphasises that professional expertise development is a social process driven by factors such as professionals' own learning expertise or their motivation to learn for their professional growth, and access to teaching expertise, and high-quality training.

The results from the current research also echo the findings from the expertise study conducted by van de Wiel et al. (2004) in that more competent EFL teacher item writers tended to more regularly perform learning activities, namely keeping up-to-date with professional literature, consulting and talking with colleagues and reflection which were strongly recommended in theories of deliberate practice (Ericsson et al., 1993) and self-regulated learning (Pintrich, 2000; Zimmerman, 2000). The results from this study further confirmed that engagement in professional learning is inherently associated with competence growth and is an indicator of making attempts and efforts to develop expertise as it is argued by van de Wiel et al., (2004).

Notably, the data analysis from this current research showed that the variable for engagement in self-regulated professional learning obtained the highest standardised weighted β coefficients, meaning that this factor made the strongest unique contribution to predicting the increased self-perceived competence in multiple-choice test item writing for English reading and listening skill tests by Vietnamese EFL teachers. This significant result resonates with the compelling evidence from teacher expertise studies that learning at the individual level is the most powerful mechanism for teacher growth and development (see Desimone, 2009; Dunn & Shriner,1999; Ericsson et al., 1993, 1996; Putnam & Borko, 2000; Sonnentag, 1998; van de Wiel et al., 2004).

In a similar vein, the data analysis revealed that the variable for institutional induction and training in multiple-choice test item writing had the second highest standardised weighted β coefficients, being another significant factor that exerts positive impacts on Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests. The positive relationship could be explained by the fact that institutional induction and training activities were explicitly mentioned in high-level items in the item writing competence scale. This finding was consistent with a number of studies in the field of item writing competence and expertise development. For example, Salisbury's (2005) case study revealed that testing bodies' provision of professional training and support through cooperative feedback, consultation, correction, and detailed and careful examination of test items contributes to the continual development of IELTS listening test item writing expertise. The finding from Wallin, Nokelainen, and Mikkonen (2019), who studied the expertise development of highly experienced professionals suggested that expertise development is a continuing, context-specific and personalised process. The finding also echoes the model of "adaptive expertise" by Hytönen, Palonen, Lehtinen, and Hakkarainen (2016), which explains how experts learn and develop their competence. The authors conceptualise "adaptive expertise" as "a professional's personal efforts aimed at deliberately improving his professional competence, seeking alternative solutions for

existing professional practices and becoming an active knowledge-building and networking actor in his professional field in order to reach the highest levels of professional competence" (p. 336). Adaptive experts possess excellent self-regulative skills (Zimmerman, 2000) since the highest level of expert performance demands self-regulation and self-reflection on improving professional competence (Nokelainen, Kaisvuo, & Pylvas, 2017).

8.2.2.2 The power of English language proficiency.

In investigating an association between teachers' English language proficiency and their perceived item writing competence, the current study revealed that English language competence was a significantly positive factor associated with item writing competence. This means that teachers who are more proficient in English language skills were very likely to have self-assessed themselves at a higher level of item writing competence than their counterparts. Although no research to date has reported the correlation between teachers' English competence and their EFL test item writing competence, to some extent, this finding reflected the result from Kim et al. (2010) who found that the item writers' native and non-native status seemed to affect their item writing practice. This finding was particularly meaningful for the Vietnamese participant item writers in this study who write English test items given that English is a foreign language for them.

8.2.2.3 The pertinence of pre-service training.

Interestingly, the results presented in this study suggested that there was a moderate relationship between teachers' perceived item writing competence and pre-service training, i.e., having a degree in TESOL/TEFL and taking a course in language testing and assessment as part of their highest degree study. Aspects of this finding offered a decent contributory explanation to the variance in the item writing competence; the teachers with a degree in TESOL and coursework training in language testing and assessment self-assessed themselves at a higher level on the test item writing competence scale than those with no such background training. The findings

coincided with the language testing and assessment standard of practice specified in EALTA (2006) and ILTA (2007) which indicated that language teaching experience and knowledge of language testing theory and practice were desirable requirements of item writers. The result may also explain the fact that a degree in TESOL/TEFL and pre-service training in language testing and assessment are core recruitment criteria for item writers (Kim et al., 2010; Buck, 2009). Furthermore, it complemented Kim et al.'s (2010) finding that the item writers in their study claimed that their growth as item writers was attributable to their knowledge of test specifications and the specifications-driven item writing process that they had gained in their degree program.

8.2.2.4 The paradox of experience.

Of all the four types of work-related experience investigated, the study found that there was no evidence of a relationship between a broad range of teachers' length of experience (i.e., Length of experience in teaching English language skills, in teaching EFL exam preparation courses, and in constructing English tests from published resources) and their perceived level of competence in multiple-choice test item writing for English reading and listening skill tests. Consequently, these types of experience did not explain much of the variability in the teachers' competence. This was consistent with the results from Sonnentag (1998), and Dunn and Shriner (1999) that it was not length of experience that influenced performance, but the quality and amount of effort devoted to continuing and deliberate professional activities, that predicted levels of proficiency. Vollmer, Spada, Caspar, and Burri (2013) also found that expertise requires a considerable amount of deliberate practice and that expertise is not automatically transformed from experience. However, this finding is not clearly or ostensibly consistent with Alderson et al.'s (1995) observation that experienced subject specific teachers with relevant qualifications are likely to produce better test items than those without such experience as their wealth of teaching and preparing students for English exams seemed to be a potential resource

for writing items. This inconsistency could be caused by the fact that all participants in the research sample were relatively equally qualified and experienced, so the distinction between their perceived levels of competence in multiple-choice test item writing for English reading and listening skill tests in terms of lengths of experience might not have been captured.

Notably, with regard to specific experience in multiple-choice test item writing, the findings from the current study demonstrated that there was a strong moderate relationship between EFL teachers' experience in multiple-choice test item writing for English reading and listening skill tests and their perceived competence levels. It seemed that experience in writing multiple-choice test items, specifically for VSTEP© reading and listening skill tests, did contribute to a high moderate extent, to the growth in the teachers' perceived multiple-choice item writing competence. The result supports Dreyfus and Dreyfus's (1960) argument that only experience in specific tasks could contribute to expertise development, not a broad notion of related experience. This bivariate result provides a potential explanation for why this study did not find an expert level in multiple-choice test item writing of the teachers in the study sample. When the survey was conducted, the teachers surveyed had been practising multiple-choice test item writing for at most four years. They had not yet accumulated enough test item writing experience through repeated exposures to reach the expert level. As stipulated by Dreyfus and Dreyfus (1980), "competence comes only after considerable experience" (p. 8). In other words, one cannot move to an expert level until one has accumulated a wealth of specific experience.

In summary, this current study found that engagement in self-regulated professional learning, institutional induction and training, pre-service training in TESOL and English language proficiency were determined as the most significant predictors of EFL teachers' item writing competence as measured through self-assessments. Regarding experience in test item writing, however, the length of work-related experience, namely experience in teaching English skills, English exams preparation courses and compiling tests from published resources was not

significant. Of all the factors hypothesised and tested, engagement in professional learning self-directed by teachers themselves and provided and supported by relevant communities of practice were the most powerful factors explaining Vietnamese EFL teachers' judgements of their item writing competence. The findings also confirmed the role of task-specific experience in expertise development as proposed by Dreyfus and Dreyfus (1960).

The data collected suggests that there were four important factors predicting Vietnamese EFL teachers' judgements of their item writing competence: teacher engagement in self-regulated professional learning, institutional induction and training, English language proficiency and pre-service degree training in TESOL. It can be inferred from the four important factors that Vietnamese EFL teachers with the following characteristics were more likely to assess themselves as more competent in multiple-choice test item writing. These are teachers who:

- engage with and self-direct their professional learning to develop their knowledge and skills for their own improved item writing practice. In other words, they are learning experts who independently learn in an endeavour to develop their expertise and evolve professionally;
- have access to test item writing expertise through institutional provision of quality; induction and training programs within their communities of practice;
- are highly proficient in English language skills;
- have taken coursework in language testing and assessment as part of their degree program;
- are continually and deliberately involved in the practice of multiple-choice test item writing;
- have been trained in TESOL in their degree program.

On the other hand, lengths of the other work-related experience were not found to have significant predictive power regarding Vietnamese EFL teachers' judgements of their item writing competence. That is, there was no evidence that the following teachers were likely to self-assess themselves as more competent at test item writing. They are the teachers who:

- have been teaching English language skills for many years;
- have been teaching English exam preparation courses such as IELTS, TOEFL iBT or TOEIC for many years.
- have been constructing English tests by compiling exercises and texts from published test materials for many years

From the evidence provided by this study, it is possible to suggest a range of implications for educational policies and practices to contribute to EFL teachers' development of expertise in test item writing in tertiary institutions in Vietnam. However, this study was conducted in a number of universities in Vietnam and conducted as self-reported assessments only; there might be a limitation in applying the findings of this study to other settings in Vietnam and other countries where English is a foreign language. Therefore, the extent to which the findings of this present study can be applied will be discussed in detail in the next section.

8.3 Contributions to Practice and Research

8.3.1 Contributions to the Field of Multiple-Choice Test Item Writing for English Reading and Writing Skill Tests Research

This study has developed a research-driven understanding of the development of expertise of EFL test item writing from the perspective of the item writers. Findings from the cross-validated psychometric and statistical results have defined and enhanced understanding of core competencies exhibited by EFL test item writers at Vietnamese universities, and the typical developmental pathways of these competencies, which were not made available in prior research in the field of test item writing expertise (see Chapter 3). This has been achieved by building on

the contemporary literature and drawing on educational measurement theory to construct a developmental progression for EFL multiple-choice test item writing competence. The current literature has provided professionals with abundant resources in the forms of manuals, guidelines, codes of practice and handbooks for test item development (e.g., ALTE, 2004; Downing & Haladyna, 2006; ETS, 2003; Haladyna, 2004; Haladyna & Rodriguez, 2013; Lane et al., 2015; Young, So, & Ockey, 2013). A number of case studies also focus on test item writing expertise, such as those by Fulkerson and colleagues (2009, 2010, 2011), Salisbury (2005), Kim et al. (2010) and Green and Hawkey (2007). However, in a review of the literature, the development of this area of expertise with underpinning progression was not found. The position of this research has been that such a clear account of the progression in the development of test item writing competence would provide a holistic understanding of what this competence constitutes as well as what the developmental pathways of the competence constituents are. Chapter Three's literature review noted that without such understanding, what is known about test item writing competence is limited to an indefinite number of guidelines and a few case studies on some particular test item writing competencies.

The major way in which this study has contributed to multiple-choice test item writing research is through the development and validation of a developmental progression for Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests. Thus, the study has contributed to advancing a theoretically driven and more thorough understanding of test item writing competence, demystifying the knowledge, skills and attributes exhibited by item writers at varying levels of proficiency by applying expertise development models to describe the competencies in terms of observable and measurable behaviours. As theory should, the progression contributes in a range of ways to an explicit understanding of EFL test item writing. It exposes the underpinning dimensions which map out a comprehensive and interpretative space to capture critical constituents of Vietnamese teacher

competence in multiple-choice test item writing for English reading and listening skill tests. Most importantly, the progression provides a hierarchical structure for organising past and current research findings and thinking about test item writing competence and offers a developmental map of how expertise in test item writing can be developed.

The research also advances the understanding of factors contributing to Vietnamese EFL teachers' judgements of their competence growth in test item writing. The study identifies a set of factors that predicted EFL teacher self-assessments of their competence in multiple-choice test item writing for English reading and listening skill tests. It provides a set of the most significant factors that strongly contributed to the prediction of test item writing competence. As it was the first study of its kind in Vietnam, neither the factors, nor the directions, nor the magnitudes of the relationships among such factors were known. The findings are consistent with those reported in expertise development studies and teacher competence growth research. The findings from this study, thus, contributes as a theoretical ground for intervention activities to build and develop teachers' capacity in test item writing.

8.3.2 Contribution to the Practice and Training of Item Writers.

The three scales which were either developed or adapted under this study to provide opportunities for Vietnamese teachers to self-assess their competence in multiple-choice test item writing for English reading and listening skill tests and to self-report their engagement in self-regulated professional learning, and access to institutional induction and training could be used as diagnostic and/or assessment or self-evaluation instruments by individual item writers and/or by language testing and assessment centres or departments - in - charge at Vietnamese universities.

The IWCOMP scale could be used as a developmental tool for teacher item writers to guide their item writing practice and monitor and scaffold their expertise development. The potential of the progression resides in its capability to demonstrate where an individual EFL

teacher item writer's competence is located on a continuum and what they need to know and be able to do to improve and monitor their progress over time (Abbott-Chapman et al., 1993; Ingvarson & Chadbourne, 1994; Gillis & McPherson, 1995; Griffin et al., 2007; Louden, 1994; Masters, 2003). Hence, it may serve to generate valuable data for the design of professional development programs or learning experiences that develop multiple-choice test item writing competence for in-service English teachers in Vietnam. For institutions or examination boards, the progression could be utilised as a diagnostic tool to identify different groups of item writers, and the identification may assist them to make informed decisions in allocating item writers who are suitable for writing English tests of different proficiency levels. It could be also a useful self-assessment measure to be utilised at the beginning of the induction, item writing training courses, or professional development workshops (hereafter referred to as training). For example, at the beginning of the training, item writers might be asked to conduct the self-assessment of their competence in multiple-choice test item writing to identify their current item writing practice. The results from this self-assessment could be analysed to specify where each item writer is located along the developmental progression. The detailed description of the item writer's expertise level could be generated and reported to the item writer using the 'rocket report' and the level descriptions for each item writing competency presented in *Figure 8.1* and *Figure 8.2* for their reference. In this report, their current level would be highlighted and the next levels' descriptions would also be presented for their information. This activity could provide the institutions with an overview of their human resource for test item writing and inform them of areas of professional learning needs that should be focused on or addressed. Identifying item writers of different expertise levels may also be helpful for organising collaborative activities taken place in the training sessions where proficient and established item writers could coach and mentor emergent and novice ones. For item writers, this self-assessment activity could offer them a glimpse into their practice, into what they have been able to do, and what they are

expected to do to move to the next level of expertise. For novice and emergent item writers, for example, they may refer to the report and reflect on their practice and look forward to the next levels to consider what their next steps in improving their item writing competence could be. Item writers may also wish to revisit the progression scale and use it to re-assess and reflect on their practice after a period of time to keep track and inform themselves of their progress. The reports on their expertise level at each point of self-assessment could be included in their profile of practice which could be regarded as a means to monitor and evaluate the progress of item writers for educative and self-development purposes.

The PROLEARN scale could be embedded in the content of the training for item writers to "feel the pulse" (the dynamics, the vibration) of their own professional learning, or to raise their awareness of the significance of independent learning for professional growth. As the scale specifies different learning activities in four categories – keeping up to date with related skills of knowledge, experimenting, reflection, and collaboration, teacher item writers may use and/or adapt them to create their own independent learning plan. This scale could also be utilised as an assessment tool to measure teachers' engagement in their own deliberate practice throughout a year or before, during, and after a training program. It would be best utilised if item writers are required to compile evidence of their deliberate learning in a profile of practice which could be displayed to showcase and share within their community of item writing practitioners at the closing ceremony of the training as their significant take-away from the training course. This could nurture their incentive and enthusiasm for professional learning and professional growth. Teacher item writers should be awarded the certificate in fulfilment of their engagement in deliberate practice and stimulated to continue to pursue and maintain such deliberate practice throughout their career.

The INTRAIN scale could be a tool for institutions to identify their strengths and areas for improvement in their formal professional training. Alternatively, it could be used to collect

item writers' response and feedback on the training they have been through and capture their training areas of interest for future improvement or for the revision of the training agenda.

One of the important targets of the foreign language education sector in Vietnam is to build teacher capability in test development in response to the national demands for a standardised measurement of English learning outcomes for high-stakes decision making. It is crucial to know which factors should be focused on in order to enhance teacher competence in multiple-choice test item writing for English reading and listening skill tests. As self-regulated professional learning, institutional induction and training, English language proficiency and pre-service degree training in TESOL/TEFL were the most significant factors associated with Vietnamese EFL teachers' judgements of their item writing competence, it may be inferred that policies designed to encourage and support teachers' development of expertise should be applied with reference to these factors.

Fortunately, these factors are interventional. The findings from the investigation suggest that provoking teachers to self-direct and deliberately engage in their own professional learning through staying up to date with the related professional literature, experimenting with new skills and knowledge, engaging in reflection and in collaboration with colleagues are imperative to their own judgements of their development of expertise in test item writing. Thus, it is implied that policies should be implemented to give incentives for teachers to be involved in such professional learning. It is suggested that their evidence - based engagement in independent professional learning should earn teachers credits toward their annual professional development and possibly lead to promotion and reward. Opportunities could be created for expert teacher item writers to share their profile of independent learning within their communities of practice to set exemplary models of self-regulating individuals' learning for professional growth.

The study has also revealed that provision of induction and training activities which are coherent, content-focused, and stimulate hands-on active learning, collaborative coaching and

co-operative feedback (Desimone, 2009) would lead to teachers' development of expertise as item writers. This presents the strong implication that such formal institutional induction and training for item writers should be continued with a constantly revised agenda to capture the professional needs of the item writers. Furthermore, enhancing teachers' English language proficiency could also be targeted to complement their item writing practice as it was found that English language competence was a significantly positive factor associated with item writing competence.

As it was found in this study that teacher engagement in self-regulated professional learning made the strongest and indeed a unique contribution to Vietnamese teachers' self-perceived competence in multiple-choice test item writing for English reading and listening skill tests. The study echoes the implications from recent research that teachers should be assisted as well as be empowered to manage their own learning to develop and strengthen their agency and self-regulative skills (Wallin et al., 2019). Tynjälä et al (2003) and Stephenson et al. (2006) also recommended that teachers be provided support to develop self-regulative and metacognitive knowledge which, in turn, may significantly support their development of professional expertise. As argued by Stephenson et al. (2006), professionals benefit greatly from examining and managing their learning independently; these activities involves designing 'personalised programs around their distinctive previous professional experience, current professional issues and longer-term professional plans' (p. 19) and perceiving themselves as 'the principal agent of control of a programme situated within critical and demanding academic and professional contexts' (p. 26). The results demonstrate that self-directed learning and autonomy are significant to developing the expertise of competent professionals. This resonates with Grenier's (2009) view that "learners need to take charge of their own development through continuing education, self-directed learning and reflection on prior experiences in order to move from competence to expertise" (pp. 154-155).

Additionally, the reviewed research emphasised the significance of peer and mentor support. Fellow students, tutors and lecturers assist learners to reflect on and see their work from new lenses, challenging them to become critical of their behaviours and prior perception (see Dalrymple et al., 2014). The findings from the current study acknowledge the shift from compliance to self-agency in teachers' learning for their professional growth. It also supports that understanding that expertise development is a constant and personalised process, indicating that it is crucial to provide multiple opportunities for teacher participation in formal training as well as autonomous learning and collaborative in accordance with individual needs.

8.4 Study Limitations and Recommendations for Future Research

Limitations of the research pertain to characteristics of the data sample and unexplored aspects of reliability and validity. While these limitations are not deemed to have affected the research findings, they imply possible areas for future inquiry.

Data sample considerations. The present study faced several limitations with regard to the sampling. One limitation was associated with the recruitment of the participants for the focus groups and quality criteria writing workshops. The item writers and subject matter experts invited to participate in the focus groups interviews and workshops were from one tertiary institution in Ha Noi. Access to those from other institutions in the other regions in Vietnam was unavailable to the researcher. This suggests the value of validating the materials with new groups of item writers for further extension and/or refinement of the framework.

Another limitation was linked to the representativeness of the item writers participating in the online survey. The total number of teacher item writers in the study (n = 150) was sufficient for the conclusions drawn about the scale; however, these teachers substantively resemble each other with respect to their educational and professional background characteristics and access to professional training for item writers. These sample characteristics may have contributed to the lack of power to find highly significant differences. For example, the small

variation in the academic qualifications of the sample, with 2 Bachelors, 131 Masters and 17 Doctoral degrees, resulted in the inability to investigate whether academic qualifications contributed to differing levels of self-perceived competence in multiple-choice test item writing for English reading and listening skill tests. Future research focussing on a greater number of participants with undergraduate and doctoral level qualifications should be considered. In addition, the sample was not representatively even in terms of work locations. More than half of the sample (79 teachers) came from two institutions in Ha Noi and Thai Nguyen. 22 teachers were from Da Nang and the rest (59) were from the remaining nine cities and provinces with some tertiary institutions where only one teacher item writer consented to participate in the study. One potential implication of this is that minimal data was collected about the teacher item writers from the Central and Southern parts of Vietnam. Despite such limitations, the findings from the current study are potentially important enough to invite replication and reconfirmation of the model of multiple-choice test item writing competence with a broader population and other disciplines beyond English in various settings.

Scales considerations. It is noteworthy there is a need to revisit the scales adapted from prior research for use in this current study to measure teacher engagement in self-regulated professional learning (the PROLEARN scale) and teacher access to institutional induction and training for item writers (the INTRAIN scale). Psychometric modelling confirmed the reliability and validity of the measurement. However, as these scales had not been previously calibrated using Item Response Theory, only when used for this study was it found that the instruments were too narrow to capture the developmental progression of the latent constructs. It is recommended that these adapted scales be further expanded for use in future research. It is essential to the interpretation of this study that all the three scales used in the study were used as self-assessment tools, so any factors claimed to be impacting the teacher expertise development are limited to their judgements of their own level of competence, which might possibly not be a

true indication of their real ability. It is proposed that these scales could be used in future research to inform the design of assessment materials such as an evidence-based portfolio of practice, third party reports, interview procedures and direct observation and record keeping so that more extensive data could be gathered and analysed.

Recommendations for Future Research.

It is suggested that future research could validate the developmental progression of EFL test item writing competence with a more diverse group of teacher item writers to gain a fuller comprehension of their item writing expertise. At the time the progression was identified in this study, the target Vietnamese English test item writers had only been recently exposed to item writing practice. It is possible that the progression could be further extended to cover the full spectrum of item writing competence as the teachers develop their expertise.

There are also fruitful avenues for further research to explore any other significant differences that might exist among item writers with regard to their educational and professional background characteristics and their workplace locations, and to determine whether different professional learning opportunities relate to differences in performance by teacher item writers in Vietnam. A follow-up study with item writer trainers and item writers could be undertaken to determine the standards-referenced scales' capability to guide item writers' practice, identify their professional learning needs and target professional learning interventions. This could include a repeat assessment of item writers involved in the targeted professional learning activities to indicate the outcome of the interventions, establishing the contribution of professional learning to improved item writing expertise. Future research could also be undertaken to explore the impact of affective factors such as intrinsic and extrinsic motivation, self-efficacy and interest on item writers' professional development and their expertise development. Additionally, future research could investigate the collaborative aspects of

multiple-choice test item writing and explore the impacts of such collaboration on item writers' development of expertise.

In conclusion, the need for English skills in the global Vietnamese citizens of the future entails the extensiveness and the power of the English tests in determining the futures of these young citizens. It is thus extremely important to develop competent test item writers who will influence the development opportunity of the country's young work force. The study has advanced our understanding of the developmental pathways of test item writing expertise and offered an insight into teacher-related factors that predict this expertise development. The findings from the study have highlighted the importance of policy and investment in the pre-service and in-service professional development conducted at both individual and institutional levels to facilitate and support English teachers' expertise development in test item writing.

References

Abbott-Chapman, J., Radford, R., & Hughes, P. (1993). Teacher competencies: A developmenal model. *Unicorn, 19*(3), 37-48.

Adams, R. J., & Khoo, S. T. (1996). *Quest: The interactive test analysis system*. Melbourne: Australian Council for Educational Research.

Adams, R. J., Wu, M. L., & Wilson, M. R. (2015). ACER ConQuest: Generalised Item Response Modelling Software [Computer software]. Version 4. Camberwell, Victoria: Australian Council for Educational Research.

Aguinis, H., & Harden, E. E. (2009). Sample size rules of thumb: Evaluating three common practice. In C. E. Lance & R. J. Vandenberg (Eds.), *Statistical and methodological myths and urban legends: Received doctrine, verity, and fable in the organisational and social sciences* (pp. 269-288). New York, NY: Routledge.

Alderson, C. J., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Alexander, P. A. (2003). The development of expertise: The journey from acclimation to proficiency. *Educational researcher, 32*(8), 10-14.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1985). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

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

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Anderson, P., & Morgan, G. (2008). *Developing tests and questionnaires for a national assessment of educational achievement* (Vol. 2). Washington DC: The World Bank.

Ash, S., O'Connor, J., Anderson, S., Ridgewell, E., & Clarke, L. (2015). A mixed-methods research approach to the review of competency standards for orthotist/prosthetists in Australia. *International Journal of Evidence-based Healthcare, 13*(2), 93-103.

Association of Languages Testers in Europe (ALTE). (2005). *Materials for the guidance of test item writers*. Retrieved from https://www.alte.org/resources/Documents/IWG%20July2005.pdf

Attri, R. K. (2019). *The models of skill acquisition and expertise development: A quick reference of summaries*. Singapore: Speed to Proficiency Research

Avalos, B. (2011). Teacher professional development in teaching and teacher education over ten years. *Teaching and Teacher Education, 27*, 10-20.

Axley, L. (2008). Competency: A concept analysis. *Nursing Forum, 43*(4), 214-222.

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

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

Baranowski, R. A. (2006). Item editing and editorial review. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 349-358). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.

Barbour, R. (2008). Introducing qualitative research: Focus groups. In *SAGE research methods* (pp. 132-148). doi:http://dx.doi.org/10.4135/9780857029034
Battel-Kirk, B., Barry, M. M., Taub, A., & Lysoby, L. (2009). A review of the international literature on health promotion competencies: Identifying frameworks and core competencies. *Global Health Promotion, 16*(2), 12-20.
Beddow, P. A. (2012). Accessibility theory for enhancing the validity of test results for students with special needs. *International Journal of Disability, Development and Education, 59*(1), 97-111.
Benner, P. (2001). *From novice to expert: Excellence and power in clinical nursing practice*. Upper Saddle River, New Jersey: Prentice Hall.
Benner, P. (2004). Using the Dreyfus model of skill acquisition to describe and interpret skill acquisition and clinical judgment in nursing practice and education. *Bulletin of Science, Technology & Society, 23*(3), 188-199.
Beretta, R. (1996). A critical view of the Delphi technique. *Nurse Researcher, 3*(4), 79-89.
Berliner, D. C. (2004). Describing he behaviour and documenting the accomplishments of expert teachers. *Bulletin of Science, Technology & Society, 24*(3), 200-212. doi:10.1177/0270467604265535
Blommaert, J. (2010). *The sociolinguistics of globalisation*. Cambridge: Cambridge University Press.
Bok, H., Jaarsma, D., Teunissen, P., Van der Vleuten, C., & Van Beukelen, P. (2011). Development and validation of a competency framework for veterinarians. *Journal of Veterinary Medical Edcuation, 38*(3), 262-269.
Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). New Jersey: Lawrence Erlbaum.
Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York: Springer.
Borko, H. (2004). Professional development and teacher learning: Mapping the terrain. *American Educational Research Association, 33*(8), 3-15.
Bosco, F. A., Aguinis, H., Singh, K., Field, J. G., & Pierce, C. A. (2015). Correlational effect size benchmarks. *Journal of Applied Psychology, 100*(2), 431-449.
Boshuizen, H. P. A., Bromme, R., & Gruber, H. (Eds.). (2004). *Professional learning: Gaps and transitions on the way from novice to expert* (Vol. 2). New York: Kluwer Academics Publishers.
Boud, D., & Garrick, J. (Eds.). (1999). *Understanding learning at work*. Bristol: The Society for Research into Higher Education and Open University Press.
Bransford, J. D., Brown, A. L., Cocking, R. R., Donovan, M. S., & Pellegrino, J. W. (Eds.). (2004). *How people learn: Brain, mind, experience, and school*. Washington, D.C: National Academy Press.
Buck, G. (2009). Challenges and constraints in language test development. In C. Alderson (Ed.), *Politics of language education* (pp. 166-184). Bristol, UK ; Buffalo, NY: Multilingual Matters.
Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1-47.
Carlson, J. E., & Davier, M. (2013). Item Response Theory. *ETS R&D Scientific and Policy Contribution Series*, https://onlinelibrary.wiley.com/doi/epdf/10.1002/j.2333-8504.2013.tb02335.x.
Carr, N. T., Nguyen, T. N. Q., Nguyen, T. M. H., Nguyen, T. Q. Y., & Nguyen, T. P. T. (2016). *Systematic support for a communicative standardised test of proficiency in Vietnam*. Paper presented at the 4th New Directions in English Language Assessment Conference, Hanoi, Vietnam.

Carroll, G., & Boutall, T. (2011). *Guide to developing National Occupational Standards*. Retrieved from Wath-upon-Dearne: http://www.ukces.org.uk/assets/ukces/docs/supporting-docs/nos/guide-todeveloping-nos.pdf

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

Chapman, M. L. (2008). *Assessment literacy and efficacy: Making valid educational decisions*. Unpublished PhD Thesis. University of Massachusetts Amherst.

Cheetham, G., & Chivers, G. (2005). *Professions, competence and informal learning*. Cheltenham: Edward Elgar.

Choy, S. P., Chen, X., & Bugarin, R. (2006). *Teacher professional development in 1999-2000: What teachers, principals and district staff report (NCES 2006-305)*. Washington, DC: National Center for Education Statistics.

Coates, H. B. (2005). *Measuring and modelling student engagement in online and campus-based higher education*. Unpublished PhD Thesis. Centre for the Study of Higher Education, The University of Melbourne.

Cobb, P. (1994). Where is the mind? Constructivist and sociocultural perspectives on mathematical development. *Educational researcher, 23*(7), 13-19.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104.

Cowan, D., Norman, I., & Coopamah, V. (2007). Competence in nursing practice: A controversial concept - a focused review of literature. *Accident and Emergency Nursing, 15*(1), 20-26.

Crisp, J., Pelletier, D., Duffield, C., Adams, A., & Nagy, S. (1997). The Delphi method. *Nursing Research, 46*(2), 116-118.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438-481.

Davidson, F., & Lynch, B. K. (2002). *Testcraft: A teacher's guide to writing and using language test specifications*. New Haven, CT: Yale University Press.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. London: Guilford.

Deković, M., Janssens, J. M. A. M., & Gerris, J. R. M. (1991). Factor structure and construct validity of the block child rearing practices report (CRPR). *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(2), 182-187.

Denham, P. (1992). English in Vietnam. *World Englishes, 11*(1), 61-69.

Department of Education Victoria. (2007). *The developmental learning framework for school leaders*.

Desimone, L. M. (2009). Improving impact studies of teachers' professional development: Toward better conceptualisations and measures. *Educational researcher, 38*(3), 181-199. doi:10.3102/0013189X08331140

Do, H. T. (1999). *Foreign language education policy in Vietnam: The emergence of English and its impact on higher education.* Paper presented at the The Fourth International Conference on Language and Development, Hanoi, Vietnam.

Do, H. T. (2006). *The role of English in Vietnam's foreign language policy: A brief history*. Paper presented at the 19th Annual EA Education Conference, Australia.

Doan, N., Pham, T., Pham, M., & Tran, K. (2018). English as an international language in Vietnam: History and development. *Asian Englishes, 20*(2), 106-121.

Downing, S. M. (2002). Construct-irrelevant variance and flawed test questions: Do multiple-choice item writing principles make any difference? *Academic Medicine, 77*, 103-104.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education, 10*(2), 133-143.
Downing, S. M. (2006). Selected-response item formats in test development. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 287-302). Mahwah, New Jersey: Lawrence Erlbaum Associates Publishers.
Downing, S. M., & Haladyna, T. M. (2006). *Handbook of test development*. Mahwah, NJ: Lawrence Erlbaum.
Dreyfus, H. L., & Dreyfus, S. E. (1986). *Mind over machine: human intuition and expertise in the era of the computer (1st Ed)*. New York: The Free Press.
Dreyfus, S. E. (2004). The five-stage model of adult skill acquisition. *Bulletin of Science, Technology & Society, 24*(3), 177-181.
Dunlea, J., Nguyen, T. N. Q., Nguyen, T. M. H., Nguyen, T. Q. Y., Thai, H. L. T., & Nguyen, T. P. T. (2016). *Reporting on the pilot phase of a test comparability project*. Paper presented at the The 4th New Directions in English Language Assessment Conference, Hanoi, Vietnam.
Dunlea, J., Spiby, R., Nguyen, T. N. Q., Nguyen, T. Q. Y., Nguyen, T. M. H., Nguyen, T. P. T., . . . Bui, T. S. (2018). *APTIS-VSTEP comparability study: Investigating the usage of two EFL tests in the context of higher education in Vietnam*. Retrieved from https://www.britishcouncil.org/sites/default/files/aptis-vstep_study.pdf:
Dunn, T. G., & Shriner, C. (1999). Deliberate practice in teaching: What teachers do for self-improvement. *Teaching and Teacher Education, 15*, 631-651.
Dunphy, B. C., & Williamson, S. L. (2004). In pursuit of expertise: Toward an educational model of expertise development. *Advances in Health Sciences Education, 9*(2), 107-127.
Duong, T. M. (2011). *The development and validation of instruments to measure process-oriented writing competence of Vietnamese EFL students*. Unpublished PhD Thesis. Melbourne Graduate School of Education, The University of Melbourne.
Duong, T. M., Nguyen, C., & Griffin, P. (2011). Developing a framework to measure process-oriented writing competence: A case of Vietnamese EFL students' formal portfolio assessment. *RELC Journal, 42*(2), 167-185.
Duong, T. M., Nguyen, C., & Griffin, P. (2019). Using Rasch analyses to inform the revision of a scale measuring students' process-oriented writing competence in portfolios. *Journal of Applied Measurement, 20*(4), 450-475.
Durksen, T. L., Klassen, R. M., & Daniels, L. M. (2017). Motivation and collaboration: The keys to a developmental framework for teachers' professional learning. *Teaching and Teacher Education, 67*, 53-66.
Educational Services Australia. (2011). *Australian professional standards for teachers*. Carlton South, VIC, Australia: Ministerial Council for Education, Early Childhood Development and Youth Affairs.
Ellsworth, R. A., Dunnell, P., & Duell, O. K. (1990). Multiple-choice test items: What are textbook authors telling teachers? *Journal of Educational Research, 83*(5), 289-293.
Enow, L., & Goodwyn, A. (2018). The invisible plan: How English teachers develop their expertise and the special place of adapting the skills of lesson planning. *Research Journal of the National Association for the Teaching of English, 52*(2), 120-134. doi:10.1080/04250494.2018.1438119
Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *JAMA: Journal of the American Medical Association, 287*(2), 226.
Eraut, M. (1998). Concepts of competence. *Journal of Interprofessional Care, 12*(2), 127-139.
Eraut, M. (2000). Non-formal learning and tacit knowledge in professional work. *British Journal of Educational Psychology, 70*, 113-136.

Erickson, A. S. G., Noonan, P. M., Brussow, J., & Carter, K. S. (2017). Measuring the quality of professional development training. *Professional Development in Education, 43*(4), 685-688. doi:10.1080/19415257.2016.1179665

Ericsson, K. A. (1996). *The road to excellence: The acquisition of expert performance in the arts and sciences, sports and games*. Mahwah: Erlbaum.

Ericsson, K. A. (2009). Enhancing the development of professional performance: Implications from the study of deliberate practice. In A. K. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 405-431). New York, USA: Cambridge University Press.

Ericsson, K. A., & Charness, N. (1994). Expert performance: Its structure and acquisition. *American Psychologist, 49*, 725-747.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review, 100*, 363-406.

Evers, A. T., Kreijn, K., & Van der Heijden, B. I. (2016). The design and validation of an instrument to measure teachers' professional development at work. *Studies in Continuing Education 38*(2), 162-178. doi:10.1080/0158037X.2015.1055465

Feiman-Nemser, S. (2001). From preparation to practice: Designing a continuum to strengthen and sustain teaching. *Teachers College Record, 103*(6), 1013-1055.

Fernandez, N., Dory, V., Ste-Marie, L., Chaput, M., Charlin, B., & Boucher, A. (2012). Varying conceptions of competence: An analysis of how health sciences educators define competence. *Medical Education, 46*(4), 257-365.

Field, A. (2018). *Discovering statistics using IBM SPSS statistics and sex and drugs and rock'n'roll* (5th ed.). Los Angeles, London, New Delhi, Singapore, Washington DC: SAGE Publication Ltd.

Finch, H., Lewis, J., & Turley, C. (2014). Focus group. In J. Ritchie, J. Lewis, C. McNaughton Nicholls, & R. Ormston (Eds.), *Qualitative research practice: A guide for social science students and researchers* (Vol. 2nd, pp. 211-242). Los Angeles: SAGE.

Fish, D., & de Cossart, L. (2006). Thinking outside the (tick) box: Rescuing professionalism and professional judgement. *Medical Education, 40*, 403-404.

Flyvbjerg, B. (2001). *Making social science matter: Why social inquiry fails and how it can succeed again*. Cambridge, UK: Cambridge University Press.

Fortus, R., Corriat, R., & Fund, S. (1998). Prediction of item difficulty in the English subtest of Israel's inter-university psychometric entrance test. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 61-87). Mahwah, NJ: Lawrence Erlbaum.

Frank, J. R., Mungroo, R., Ahmad, Y., Wang, M., De Rossi, S., & Horsley, T. (2010). Toward a definition of competency-based education in medecine: A systematic review of published definitions. *Medical Teacher, 32*(8), 631-637.

Fretwell, D. H., Lewis, M. V., & Deij, A. (2001). A framework for defining and assessing occupational and training standards in developing countries
Retrieved from https://files.eric.ed.gov/fulltext/ED458419.pdf. https://files.eric.ed.gov/fulltext/ED458419.pdf

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. London: Routledge.

Fulkerson, D., Mittelholtz, D. J., & Nichols, P. D. (2009). *The psychology of writing items: Improving figural response item writing*. Paper presented at the The annual meeting of the American Educational Research Association, San Diego, CA.

Fulkerson, D., Nichols, P., & Mittelholtz, D. J. (2010). *What item writers think when writing items: Toward a theory of item writing expertise*. Paper presented at the the annual meeting of the American Educational Research Association, Denver, CO.

Fulkerson, D., Nichols, P., & Snow, E. (2011). *Expanding the model of item writing expertise: Cognitive processes and requisite knowledge structures*. Paper presented at the Annual Meeting of the American Educational Research Association, New Orleans, Louisiana.
Gardner, A., Hase, S., Gardner, G., Dunn, S., & Carryer, J. (2007). From competence to capability: A study of nurse practitioners in clinical practice. *Australian Journal of Advanced Nursing, 24*(1), 8-14.
Ge, X., & Hardré, P. L. (2010). Self-processes and learning environment as influences in the development of expertise in instructional design. *Learning Environments Research, 13*(1), 23-41.
Gillis, S. (2003). *The domains of vocational assessment decision making*. PhD Thesis. The University of Melbourne, Australia.
Gillis, S., & Macpherson, C. (1995). *Examination of the links between pre-employment qualifications and on the job competency based assessment*. Paper presented at the The Annual Conference of the Australian Association for Research in Education, Hobart, Australia.
Glaser, R. (1981). The future of testing: A research agenda for cognitive psychology and psychometrics. *American Psychologist, 36*, 923-936.
Gonczi, A. (1994). Competency based assessment in the professions. *Assessment in Education: Principles, Policy & Practice, 1*(1), 27-44.
Gonczi, A. (2013). Competency-based approaches: Linking theory and practice in professional education with particular reference to health education. *Educational Philosophy and Theory, 45*(12), 1290-1306.
Gonczi, A., & Hager, P. (2010). The competency model. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International Encyclopedia of Education* (3th ed., pp. 403-410). Amsterdam: Elsevier.
Gonczi, A., Hager, P., & Oliver, L. (1990). *Establishing competency-based standards in the professions*. Canberra: Australian Government Publishing Service. Retrieved from http://www.voced.edu.au/content/ngv%3A29478
Goodwyn, A. (2011). *The expert teacher of English*. Milton Park, Abington, Oxon: Routledge.
Goodwyn, A. (2017). *Expert teachers: An international perspective*. Abington: Routledge.
Government of Vietnam. (1968). *The Prime Minister's Directive No 43 – TTG/VG on strengthening foreign language teaching and learning at high schools and for scientists, economists and technicians* Hanoi, Vietnam: Prime Minister Office,
Government of Vietnam. (1994). *The Prime Minister's Directive No 422-TTg on foreign language intensive training for government personnel*. Hanoi, Vietnam: Prime Minister Office
Government of Vietnam. (2008). *Decision No. 1400/QD-TTg on "Teaching and learning foreign languages in the national education system, period 2008-2020"*. Hanoi, Vietnam: Prime Minister Office
Government of Vietnam. (2017). *Decision No. 2080 on "Amendments and supplements tot he foreign language teaching and learning project for the 2017-2025 period"*. Hanoi, Vietnam: Prime Minister Office
Grant, J. (1999). The incapacitating effect of competence. *Advances in Health Sciences Education, 4*(271-277).
Green, A., & Hawkey, R. (2007). *An empirical investigation of the process of writing Academic Reading test items for the International English Language Testing System*. Retrieved from https://www.ielts.org/-/media/research-reports/ielts_rr_volume11_report5.ashx
Green, A., & Hawkey, R. (2011). Re-lifting for a different purpose: A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing, 29*(1), 109-129.

Greeno, J. G. (1997). On claims that answer the wrong questions. *Educational researcher, 26*(1), 5-17.
Greeno, J. G., Collins, A. M., & Resnick, L. B. (1996). Cognition and learning. In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 15-46). New York: Simon & Schuster Macmillan.
Griffin, P. (1997). *Developing assessment in schools and workplaces*. Faculty of Education: Deans Lecture Series 1997: An Inaugural Professorial Lecture delivered by Prof. Patric Griffin, Foundation Chair of Educational Assessment, The University of Melbourne, 18 September, 1997.
Griffin, P. (2001). Profiles and reporting in a developmental assessment framework. In V. Zbar & T. Mackay (Eds.), *Leading the education debate*. Jolimont: IARTV.
Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Education Evaluation, 33*, 87-99.
Griffin, P. (Ed.) (2014). *Assessment for teaching*. Cambridge: Cambridge University Press.
Griffin, P. (Ed.) (2018). *Assessment for teaching*. Cambridge: Cambridge University Press.
Griffin, P., & Care, E. (2014). Test construction. In P. Griffin (Ed.), *Assessment for teaching* (pp. 156-173). Cambridge: Cambridge University Press.
Griffin, P., Care, E., Roberton, P., Crigan, J., Awwal, N., & Pavlovic, M. (2013). Assessment and learning partnerships in an online environment. In E. McKay (Ed.), *ePedagogy in online learning: New developments in web medicated human computer interaction* (pp. 39-54). Harshey, PA: IGI Global.
Griffin, P., & Francis, M. (2018). Building a development progression. In P. Griffin (Ed.), *Assessment for teaching* (pp. 161-197). Cambridge: Cambridge University Press.
Griffin, P., & Francis, M. (2018). Writing rubrics. In P. Griffin (Ed.), *Assessment for teaching* (2nd ed., pp. 113-140). Cambridge: Cambridge University Press.
Griffin, P., Francis, M., & Robertson, P. (2018). Judgement-based assessment. In P. Griffin (Ed.), *Assessment for teaching*. Cambridge: Cambridge University Press.
Griffin, P., & Gillis, S. (1999). *Assessing higher order competencies: A multi source approach to competency-based assessment*. Paper presented at the The Annual Conference of the Australian Association for Research in Education - New Zealand Association for Research in Education (AARE-NZARE), Melbourne.
Griffin, P., & Gillis, S. (2001). *Using indicators of quality to infer competence.* Paper presented at the National ACACA Conference.
Griffin, P., & Gillis, S. (2014). Connecting competence and quality: Scored assessment in Year 12 VET. Retrieved from https://www.researchgate.net/profile/Patrick_Griffin3/publication/238777724_Connecting_Competence_and_Quality_Scored_Assessment_in_Year_12_VET/links/0deec52dec4c70d19f000000/Connecting-Competence-and-Quality-Scored-Assessment-in-Year-12-VET.pdf
Griffin, P., Gillis, S., & Calvitto, L. (2004). *Connecting competence and quality: Scored assessment in Year 12 VET. A report to the New South Wales Department of Education*. Melbourne, Australia: Assessment Research Centre, University of Melbourne
Griffin, P., Gillis, S., & Calvitto, L. (2007). Standard-referenced assessment for vocational education and training in schools. *Australian Journal of Education, 51*(1), 19-38.
Griffin, P., Gillis, S., & Taylor, M. (2002). *Scored assessment for senior secondary certificates: A national project*. Paper presented at the National Conference of the Australian Association for Research in Education, Brisbane, Australia.
Griffin, P., Nguyen, C., & Gillis, S. (2004). *Developing and validating primary school teacher standards in Vietnam.* Paper presented at the Annual Conference of the Australian Association for Research in Education, the University of Melbourne.

Griffin, P., Nguyen, C., Gillis, S., & Mai, T. T. (2006). An empirical analysis of primary teacher standards in Vietnam. Retrieved from https://www.researchgate.net/publication/228618756_An_empirical_analysis_of_primary_teacher_standards_in_Vietnam

Griffin, P., Rickards, F., & Francis, M. (2018). Competence assessment: A clinical approach. In P. Griffin (Ed.), *Assessment for teaching* (2nd ed., pp. 43-77). Cambridge: Cambridge University Press.

Guerrero, D., & De los Rios, I. (2012). Professional competences: A classification of international models. *Procedia - Social and Behavioral Sciences, 46*, 1290-1296.

Gupta, U. G., & Clarke, R. E. (1996). Theory and applications of Delphi technique: A bibliography (1975 - 1994). *Technological Forecasting & Social Change, 53*, 185-211.

Guskey, T. R. (2000). *Evaluating professional development*. Thousand Oaks, C.A: Corwin Press.

Hager, P. (1995). Competency standards - a help or a hindrance? An Australian perspective. *The Vocational Aspect of Education, 47*(2), 141-151.

Hager, P. (2004). The competence affair, or why vocational education and training urgently needs a new understanding of learning. *Journal of Vocational Education and Training, 45*(3), 409-433.

Hager, P. (2017). The integrated view on competence. In M. Mulder (Ed.), *Competence-based vocational and professional education: Bridging the worlds of work and education* (Vol. 23, pp. 203-228). Switzerland: Springer.

Hager, P., & Beckett, D. (1995). Philosophical underpinnings of the integrated conception of competence. *Educational Philosophy and Theory, 27*(1), 1-24.

Hager, P., & Gillis, S. (1995). *Assessment at higher levels*. Paper presented at the National Vocational Education and Training Research Conference, Adelaide, Australia.

Hager, P., & Gonczi, A. (1993). Attributes and competence. *Australian and New Zealand Journal of Vocational Education Research, 1*(1), 36-45.

Hager, P., & Gonczi, A. (1998). Development of professional competencies: A case study in the complexities of corporatist policy implementation. In A. Yeatman (Ed.), *Activisim and the policy process* (pp. 56-80). Sydney: Allen & Unwin.

Hager, P., Gonczi, A., & Athanasou, J. (1994). General issues about assessment of competence. *Assessment and evaluation in higher education, 19*(1), 3-16.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis: A global perspective* (7th ed.). Harlow, Essex: Pearson Education Limited.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items (3rd ed.)*. Mahwah, NJ: Lawrence Erlbaum.

Haladyna, T. M., & Downing, S. M. (1989). The validity of a taxonomy of multiple-choice item writing rules. *Applied Measurement in Education, 1*, 51-78.

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item. *Educational and Psychological Measurement, 53*(4), 999-1010.

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*, 309-334.

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

Hallam, S. (2010). Transitions and the development of expertise. *Psychology Teaching Review, 16*, 3-32.

Hambleton, R. K., & Jirka, S. (2006). Anchor-based method for judgmentally estimating item statistics. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 399-420). Mahwah, NJ: Lawrence Erlbaum.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. New York: Springer Business Media, LLC.
Hamp-Lyons, L., & Lynch, B. K. (1998). Perspective on validity: A historical analysis of language testing. In A. J. Kunnan (Ed.), *Validation in language assessment: Selected paper from the 17th Language Testing Research Colloquium* (pp. 253-276). Long Beach, Mahwah, NJ: Lawrence Erlbaum.
Hasson, F., Keeney, S., & McKenna, H. P. (2000). Research guidelines for the Delphi survey technique. *Journal of Advanced Nursing, 32*(4), 1008-1016.
Hearn, G., Close, A., Smith, B., & Soutley, G. (1996). Defining generic professional competencies in Australia: Towards a framework for professional development. *Asia Pacific Journal of Human Resources, 34*(1), 44-62.
Hemphill, J. F. (2003). Interpreting the magnitudes of correlation coefficients. *American Psychologist, 58*, 78-79. doi:10.1037/0003-066X.58.1.78
Heywood, L., Gonczi, A., & Hager, P. (1992). *A guide to development of competency standards for professions*. National Office of Overseas Skills Recognition Research Paper No.7, DEET. Canberra: Australian Government Publishing Service.
Hoang, V. V. (2011). The current situation and issues of the teaching of English in Vietnam. Retrieved from https://www.researchgate.net/publication/326718789_The_Curent_Situation_and_the_Teaching_of_English_in_Vietnam
Holden, R. R., Fekken, G. C., & Cotton, D. H. G. (1991). Assessing psychopathology using structured test-item response latencies. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 3*(1), 111-118.
Homer, C., Passant, L., Kildea, S., Pincombe, J., Thorogood, C., Leap, N., & Brodie, P. (2007). The development of national competency standards for the midwife in Australia. *Midwifery, 23*, 350-360.
Huddle, T. S., & Heudebert, G. R. (2007). Taking apart the art: The risk of anatomizing clinical competence. *Academic Medicine, 82*(6), 536-541.
Hytönen, K., Palonen, T., Lehtinen, E., & Hakkarainen, K. (2016). Between two advisors: Innterconnecting academic and workplace settings in an emerging field. *Vocations and Learning, 9*(3), 333-359.
Ingham, K. (2008). *The Cambridge ESOL approach to item writer training: The case of ICFE listening*. Retrieved from http://www.cambridgeenglish.org/images/23151-research-notes-32.pdf
Ingvarson, L. (1996). Professional standards for the teaching of science: What roles for ASTA? *Australian Science Teachers Journal, 42*(2), 7-12.
Ingvarson, L., & Kleinhenz, E. (2003). A review of standards of practice for beginning teaching. Retrieved from http://research.acer.edu.au/teaching_standards/5
Ingvarson, L., Meiers, M., & Beavis, A. (2005). Factors affecting the impact of professional development programs on teachers' knowledge, practice, student outcomes and efficacy. *Education Policy Analysis Archives, 13*(10), 1-29.
Iramaneerat, C. (2012). The impact of item writer training on item statistics of multiple-choice items for medical student examination. *Siriraj Medical Journal, 64*, 178-182. Retrieved from https://www.semanticscholar.org/paper/The-Impact-of-Item-Writer-Training-on-Item-of-Items-Iramaneerat/952b22c39205c3b805d346a5cbe00ca836eb1371
Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house medical school examinations. *Academic Medicine, 77*(2), 156-161.
Kiessling, C., Dieterich, A., Fabry, G., Hölzer, H., Langewitz, W., Mühlinghaus, I., . . . Schubert, S. (2010). Communication and social competencies in medical education in

German-speaking countries: The Basel consensus statement - result of a Delphi survey. *Patien Education and Counseling, 81*, 259-266.
Kim, J., Chi, Y., Huensch, A., Jun, H., Li, H., & Roullion, V. (2010). A case study on an item writing process: Use of test specifications, nature of group dynamics, and individual item writers' characteristics. *Language Assessment Quarterly, 7*(2), 160-174.
King, J. D. (2010). *Criterion-referenced assessment literacy of educators*. Unpublished PhD Thesis. The University of Southern Mississippi.
Kirk, M., Tonkin, E., & Skirton, H. (2014). An iterative consensus-building approach to revising a genetics/genomics competency framework for nurse education in the UK. *Journal of Advanced Nursing, 70*(2), 405-420.
Klemp, G. (1988). The meaning of success. *Liberal Education, 74*(3), 37-41.
Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. Englewood Cliffs, NJ: Prentice-Hall.
Krathwohl, D. R., Bloom, B. S., & Masia, B. B. (1964). *Taxonomy of educational objectives, the classification of educational goals. Hanbook II: Affective Domain*. New York, NY: David Mckay Co., Inc.
Kuiken, F., & Vedder, I. (2014). Rating written performance: What do raters do and why? *Language Testing, 31*(3), 329-348.
Kwakman, K. (2003). Factors affecting teachers' participation in professional learning activities. *Teaching and Teacher Education, 19*, 149-170. doi:0.1016/S0742-051X(02)00101-4
Lai, C. P. (1998). *Statistical modelling of students' performance in multicomponent tasks*. Unpublished PhD Thesis. Department of Learning and Educational Development, The University of Melbourne.
Lajoie, S. P. (2003). Transitions and trajectories for studies of expertise. *Educational researcher, 32*(8), 21-25.
Landeta, J. (2006). Current validity of the Delphi method in social sciences. *Technological Forecasting & Social Change, 73*, 467-482.
Lane, S., Raymond, M. R., & Haladyna, T. M. (2015). *Handbook for test development*. New York: Routledge.
Lane, S., Raymond, M. R., Haladyna, T. M., & Downing, S. M. (2016). Chapter 1: Test development process. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 3-18). New York&London: Routledge.
Le, L. (2012). Item point-biserial discrimmination. Retrieved from Retreived from https://www.acer.org/files/Conquest-Notes-5-ItemPointBiserialDiscrimination.pdf
Le, V. C. (2015). English language education innovation for the Vietnamese secondary school: The Project 2020. In B. Spolsky & K. Sung (Eds.), *Secondary school English education in Asia: From policy to practice* (pp. 182-200). New York, NY: Routledge.
Lester, S. (2014). Professional competence standards and frameworks in the United Kingdom. *Assessment and evaluation in higher education, 39*(1), 38-52.
Lester, S. (2015). Competence standards and frameworks: some lessons from the United Kingdom. *Polish Journal of Continuing Education, 3*(90), 132-141.
Levy-Vered, A., & Alhija, F. N. (2015). Modelling beginning teachers' assessment literacy: The contribution of training, self-efficacy, and conceptions of assessment. *Educational Research and Evaluation, 21*(5-6), 378-406.
Lieberman, A. (1995). Practices that support teacher development. *Phi Delta Kappan, 76*, 591-596.
Lim, G., S. (2009). *Prompt and rater effects in second language writing performance assessment.* (PhD). The University of Michigan, USA,

Linacre, J. M. (1994a). Constructing measurement in a many-facet Rasch model. In M. Wilson (Ed.), *Objective measurement: Theory in practice* (2nd ed., Vol. II). New Mark, NJ: Ablex.
Linacre, J. M. (1994b). Sample size and item calibration stability. *Rasch Measurement Transactions, 4*, 231-242.
Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement, 3*(2), 103-122.
Lindberg, M., Lundström-Landegren, K., Johansson, P., Lidén, S., & Holm, U. (2012). Competencies for practice in renal care: A national Delphi study. *Journal of Renal Care, 21*(7-8), 1145-1154.
Linstone, H. A., & Turroff, M. (2002). The Delphi method: Techniques and applications. Retrieved from http://is.njit.edu/pubs/delphibook/delphibook.pdf
Little, J. W. (1993). Teachers' professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis, 11*(2), 129-151.
Lloyd-Jones, G., Fowell, S., & Bligh, J. G. (1999). The use of the nominal group technique as an evaluation tool in medical undergraduate education. *Medical Education, 33*, 8-13.
Lord, F. M. (1983). Small N justified Rasch model. In D. J. Weiss (Ed.), *New horizons in testing* (pp. 51-62). New York: Academic Press.
Louden, W. (1994). Setting standards for teaching. In L. Ingvarson & R. Chadbourne (Eds.), *Valuing teachers' work: New directions in teacher appraisal*. Camberwell: ACER.
Louden, W., & Wildy, H. (1998). *Circumstance and proper timing: Context and the contruction of a standard framework for school principals' performance*. Paper presented at the International Conference on Professional Standards and the Status of Teaching, Fremantle.
Lum, G. (2004). On the non-discursive nature of competence. *Educational Philosophy and Theory, 36*(5), 485-496.
Lynch, B. K., & Davidson, F. (1994). Criterion-referenced language test development: Linking curricula, teachers, and tests. *TESOL Quarterly, 28*(4), 727-743.
MacGregor, D., Kenyon, D., Christenson, J., & Louguit, M. (2008). *Predicting item difficulty: A rubric-based approach*. Paper presented at the The American Association of Applied Linguistics, Washington, DC.
Magee, B. (1998). *Confession of a Philosopher*. London: Phoenix.
Mahony, P., & Hextall, I. (2000). *Reconstructing teaching: Standards, performance and accountability*. London: Routledge Falmer.
Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment and evaluation in higher education, 37*(8), 919-931.
Marton, F. (1981). Phenomenography: Describing conceptions of the world around us. *Instructional Science, 10*(2), 177-200.
Masters, G. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.
Masters, G. (2003). Standards and assessment for students and teachers: A developmental paradigm. In V. Zbar & T. Mackay (Eds.), *Leading the education debate: Selected papers from a decade of the IARTV seminar series* (pp. 168-180). Jolimont, VIC: IARTV.
Masters, G., Adams, R., & Wilson, M. (1990). Charting of student progress. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies. Supplementary Volume 2* (pp. 628-634). Oxford: Pergamon Press.
Masters, J., Hulsmeyer, B., Pike, M., Leichty, K., Miller, M., & Verst, A. (2001). Assessment of multiple-choice questions in selected test banks accompanying text books used in nursing education. *Journal of Nursing Education, 40*(1), 25-32.

McKenna, H. P. (1994). The Delphi techniques: a worthwhile research approach for nursing? *Journal of Advanced Nursing, 19*, 1221-1225.
McNamara, T. (1996). *Measuring second language performance*. London, New York: Longman.
McNamara, T. (2008). The socio-political and power dimensions of tests. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education, 2nd Edition, Volume 7: Language testing and assessment* (pp. 415-427). New York, London: Springer.
McNamara, T., & Roever, C. (2006). *Language testing: The social dimension*. Oxford, UK: Blackwell Publishing.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd Ed)* (pp. 13-103). New York, NY: American Council on Education.
Messick, S. (1995). Validity in spychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741-749.
Mislevy, R. J. (2003). Substance and structure in assessment arguments. *Law, Probability and Risk, 2*, 237-258.
Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence-centred design.* Retrieved from https://www.ets.org/Media/Research/pdf/RR-03-16.pdf:
Mislevy, R. J., & Haertel, G. (2006). *Implication for evidence-centred design for educational testing (Draft PADI Technical Report No.17)*. Retrieved from http://padi.sri.com/downloads/TR17_EMIP.pdf
Mislevy, R. J., & Yin, C. (2013). Evidence - centred design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing*. London&New York: Routledge: Taylor & Francis.
Moaveni, A., Gallinaro, A., Conn, L., Callanhan, S., Hammond, M., & Oandasan, I. (2011). Delphi approach to developing a core competency framework for family practice registered nurses in Ontario. *Nursing Leadership, 23*(4), 45-60.
Moon, Y. K., Kim, E. J., & You, Y. M. (2013). Study on exertise development process based on arête. *International Journal of Information and Education Technology, 3*(2), 226-230.
Moreno, R., Martínez, R. J., & Muñiz, J. (2015). Guidelines based on validity criteria for the development of multiple choice items. *Psicothema, 27*(4), 388-394.
Morrow, K. (1986). The evaluation of tests of communicative performance. In M. Portal (Ed.), *Innovations in language testing*. London: NFER/Nelson.
Myers, K. K., & Oetzel, J. G. (2003). Exploring the dimensions of organisational assimilation: Creating and validating a measure. *Communication Quarterly, 51*(4), 438-457.
Naeem, N., van der Vleuten, C., & Alfaris, E. A. (2012). Faculty development on item writing substantially improves item quality. *Advances in Health Sciences Education Theory and Practice, 17*, 369-376.
Nash, J. M., & Larkin, K. T. (2012). Geometric models of competency development in specialty areas of professional psychology. *Training and Education in Professional Psychology, 6*(1), 37-46.
Nguyen, C. T. K. (2002). *Family and student influences on withdrawal from lower secondary school in rural Vietnam*. Unpublished PhD Thesis. Assessment Research Centre, The University of Melbourne.
Nguyen, N. (2012). How English has displaced Russian and other foreign languages in Vietnam since "Doi Moi". *International Journal of Humanities and Social Science, 2*(23), 259-266.
Nguyen, N. H. (2013). *National Foreign Language Project 2020*. Paper presented at the Vietnam Engineering Education Conference, Can Tho University, Vietnam.

Nguyen, T. N. Q., & Do, T. M. (2015). *Developing a made-in-Vietnam standardised test of English proficiency for adults: The status-quo and future development*. Paper presented at the The International TESOL Symposium 2015, Danang, Vietnam.

Nicholson, P., Griffin, P., Gillis, S., Wu, M., & Dunning, T. (2013). Measuring nursing competencies in the operating theatre: Instrument development and psychometric analysis using Item Response Theory. *Nurse Education Today, 33*, 1088-1093.

Nokelainen, P., Kaisvuo, H., & Pylvas, L. (2017). Self-regulaltion and competence in formal and informal contexts of vocational and professional education. In M. Mulder (Ed.), *Competence-based vocational and professional education: Bridging the worlds of work and education* (Vol. 23, pp. 756-774). Switzerland: Springer.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.

O'Connell, J., & Gardner, G. (2012). Development of clinical competencies for emergency nurse practitioners: A pilot study. *Australasian Emergency Nursing Journal, 15*(4), 195-201.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance, and other formats*. Boston: Kluwer Academic Publishers.

Palermo, C., Conway, J., Beck, E. J., Dart, J., Capra, S., & Ash, S. (2016). Methodology for developing competency standards for dietitians in Australia. *Nursing and Health Sciences, 18*, 130-137.

Pedaste, M., Leijen, A., & Poom-Valickis, K. (2019). Teacher professional standards to support teacher quality and learning in Estonia. *European Journal of Education: Research, Development, and Policy, 54*(3), 1-11.

Persky, A. M., & Robinson, J. D. (2017). Moving from novice to expertise and its implication for instruction. *American Journal of Pharmaceutical Education, 81*(9), 72-80.

Pham, T. N. (2014). Foreign language policy. In L. Tran, S. Marginson, H. Do, Q. Do, T. Le, N. Nguyen, T. Vu, T. Pham, H. Nguyen, & T. Ho (Eds.), *Higher education in Vietnam: Flexibility, mobility and practicality in the global knowledge economy*. New York, NY: Palgrave Macmillan.

Pintrich, P. R. (2000). The role of goal orientation in self-regulated learning. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 451-502). San Diego: Academic Press.

Potolea, D. (2013). Doctoral studies and research competences. *Procedia - Social and Behavioral Sciences, 76*, 935-946.

Power, M. (1997). *The audit society: Rituals of verification*. Oxford: Oxford University Press.

Putnam, R. T., & Borko, H. (2000). What do new views of knowledge and thinking have to say about research on teacher learning? *Educational Researchers, 29*(1), 4-15.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

Reeves, S., Fox, A., & Hodges, B. D. (2009). The competency movement in the health professions: Ensuring consistent standards or reproducing conventional domains of practice. *Advances in Health Sciences Education, 14*(4), 451-453.

Richter, D., Kunter, M., Klusmann, U., Lüdtke, O., & Baumert, J. (2011). Professional development across the teaching career: Teachers' uptake of formal and informal learning opportunities. *Teaching and Teacher Education, 27*, 116-126.

Riconscente, M. M., Mislevy, R. J., & Corrigan, S. (2016). Evidence-centred design. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 40-63). New York & London: Routledge Taylor and Francis.

Rodolfa, E., Bent, R., Eisman, E., Nelson, P., Rehm, L., & Ritchie, P. (2005). A cube model for competency development: Implications for psychology educators and regulators. *Professional Psychology: Research and Practice, 36*(4), 347-354.

Rodriguez, M. C. (2016). Selected-response item development. In M. R. Raymond, S. Lane, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 259-273). London & New York: Routledge Taylor and Francis Groups.
Rothwell, W. J., & Lindholm, J. E. (1999). Competency identification, modelling and assessment in the USA. *International Journal of Training and Development, 3*(2), 90-105.
Salisbury, K. (2005). *The edge of expertise: Toward an understanding of listening test item writing as professional practice*. King's College Unpublished PhD Thesis. King's College London, UK.
Salisbury, K. (2016). *The edge of expertise: Test writers think aloud*. Paper presented at the The International Conference of Education, Research and Innovation 2016, Seville, Spain.
Schmeiser, C. B., & Welch, C. J. (2006). Test development. In R. L. Brennan (Ed.), *Educational measurement (4th ed)* (pp. 307-353). Washington, DC: American Council on Education.
Schmidt, K. M., & Embretson, S. E. (2013). Item response theory and measuring abilities. In B. I. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology* (2nd ed., Vol. 2, pp. 451-473). Hoboken, New Jersey: Wiley.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350-353.
Schön, D. A. (1987). *Educating the reflective practitioner*. San Francisco, CA: Jossey-Bass.
Shapiro, L. (1995). *English language training in Vietnam in the era of Doi Moi: Ho Chi Minh City - a descriptive case study.* (Master's thesis). The School for International Training Brattleboro, Vermont. Retrieved from http://files.eric.ed.gov/fulltext/ED418605.pdf
Shin, D. (2012). Item writing and writers. In F. Davidson & G. Fulcher (Eds.), *The Routledge Handbook of Language Testing*. Hoboken: Taylor and Francis.
Shohamy, E. (2001). *The power of tests: A critical perspective on the uses of language tests*. London: Pearson.
Snoek, M., Dengerink, J., & de Wit, B. (2019). Reframing the teacher profession as a dynamic multifaceted profession: A wider perspective on teacher quality and teacher competence frameworks. *European Journal of Education: Research, Development, and Policy*, *54*(3), 1-13.
Sonnentag, S. (1998). Expertise in professional software design: A process study. *Journal of Applied Psychology*, *83*(5), 703-715.
Sonnentag, S., & Kleine, B. M. (2000). Deliberate practice at work: A study with insurance agents. *Journal of Occupational and Organisational Psychology*, *73*, 87-102.
Spolsky, B. (1995). *Measure words: The development of objective language testing*. Oxford: Oxford University Press.
Staggers, N., Gassert, C. A., & Curran, C. (2002). A Delphi study to determine informatics competencies for nurses at four levels of practice. *Nursing Research, 51*(6), 383-390.
Sternberg, R. J. (1999). Intelligence as developing expertise. *Contemporary Educational Psychology*, *24*(4), 359-375.
Stewart, J. (2001). Is the Delphi technique a qualitative method? *Medical Education, 35*, 922-923.
Stiggins, R. J., Griswold, M. M., & Wikelund, K. R. (1989). Measuring thinking skills through classroom assessment *Journal of Educational Measurement, 26*, 233-246.
Stone, C., Boud, D., & Hager, P. (2011). Assessment of Osteopaths: Developing a capability-based approach to judging readiness to practice. *International Journal of Osteopathic Medicine, 14*(4), 129-140.
Suhairom, N., Musta'amal, A. H., Mohd Amin, N. F., & Johari, N. K. A. (2014). The development of competency model and instrument for competency measurement: The research methods. *Procedia - Social and Behavioral Sciences, 152*, 1300-1308.

Sutton, F. A., & Arbon, P. A. (1994). Australian nursing moving forward? Competencies and the nursing profession. *Nurse Education Today, 14*(5), 388-393.
Sydorenko, T. (2011). Item writer judgments of item difficulty versus actual item difficulty: A case study. *Language Assessment Quarterly, 8*, 34-52.
Symeonidis, V. (2019). Teacher competence frameworks in Hungary: A case study on the continuum of teacher learning. *European Journal of Education: Research, Development, and Policy, 54*(3), 1-13.
Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Harlow, Essex: Pearson Education Ltd.
Tao, N. (2014). *Development and validation of classroom assessment literacy scales: English as a Foreign Language (EFL) instructors in a Cambodian higher education setting*. PhD Thesis. College of Education, Victoria University, Melbourne, Australia.
Tennant, M. (1991). Expertise as a dimension of adult development: Implications for adult education. *New Education, 13*, 49-55.
Tigelaar, D., Dolmans, D., Wolfhagen, I., & Van der Vleuten, C. (2004). The development and validation of teaching competencies in higher education. *Higher Education, 48*(2), 253-268.
Truong, N., & Wang, C. (2019). Understanding Vietnamese college students' self-efficacy beliefs in learning English as a foreign language. *System, 84*, 123-132.
Tuoi Tre News. (2005). It is time to see English not as a foreign language. Retrieved from https://tuoitre.vn/khong-nen-xem-tieng-anh-la-ngoai-ngu-972822.htm
Turoff, M. (2002). The Policy Delphi. In H. A. Linstone & M. Turoff (Eds.), *The Delphi method: Techniques and applications* (pp. 80-96). Retrieved from http://is.njit.edu/pubs/delphibook/delphibook.pdf
van de Wiel, M. W. J., Szegedi, K. H. P., & Weggeman, M. C. D. P. (2009). Professional learning: Deliberate attempts at developing expertise. In K. A. Ericsson (Ed.), *Development of professional expertise: Toward measurement of expert performance and design of optimal learning environments* (pp. 182-206). New York, USA: Cambridge University Press.
Vietnam Ministry of Education and Training (MOET). (1996). *Guidance No. 6627/TH on "Piloting teaching English at promary schools"*. Ha Noi: MOET
Vietnam Ministry of Education and Training (MOET). (2005). *Decision No. 50 on "Teaching English at primary school"*. Hanoi: MOET
Vietnam Ministry of Education and Training (MOET). (2014). *Circular No. 01/TT-BGDDT on "The six-level English language proficiency standards for Vietnamese"*. Hanoi: MOET
Vollmer, S., Spada, H., Caspar, F., & Burri, S. (2013). Expertise in clinical psychology: The effects of university training and practical experience on expertise in clinical psychology. *Frontiers in Psychology, 4*(141), 1-12.
Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge, Massachusetts, London: Harvard University Press.
Wallin, A., Nokelainen, P., & Mikkonen, S. (2019). How experienced professionals develop their expertise in work-based higher education: A literature review. *High Education, 77*, 359-378.
Warren, C. A. B. (2001). Qualitative interviewing. In J. F. Gubrium & J. A. Holstein (Eds.), *Handbook of Interviewing Research* (pp. 83-103). Thousand Oaks, CA: SAGE.
Watson, R., Stimpson, A., Topping, A., & Porock, D. (2012). Clinical competence assessment in nursing: A systematic review of the literature. *Journal of Advanced Nursing, 39*(5), 421-431.
Weir, C. J. (1990). *Communicative language testing*. London: Prentice Hall.

Welch, C. J. (2006). Item and prompt development in performance testing. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 303-327). Mahwah, New Jersey, London: Lawrence Erlbaum Associates.
White, R. H. (1959). Motivation reconsidered: The concept of competence. *Psychological Review, 66*, 297-333.
Whitehead, N. (2013). *Review of adult vocational qualifications in England*. Wath-upon-Dearne: UK Commission for Employment and Skills.
Wilkinson, S. (1999). Focus groups: A feminist method. *Psychology of Women Quarterly, 23*(2), 221-244.
Wilson, M. (2011). Some notes on the term: "Wright Map". *Rasch Measurement Transactions, 25*(3), 1331.
Wilson, M., Scalise, K., & Gochyyev, P. (2018). Learning in digital networks as a modern approach to ICT literacy. In E. Care, P. Griffin, & M. Wilson (Eds.), *Assessment and teaching of 21st century skills: Research and applications* (pp. 181-210). Switzerland: Springer.
Wise, L. L., & Blake, B. S. (2016). Test design and development following the Standards for Educational and Psychological Testing. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 19-39). New York, NY: Routledge.
Wolf, A. (2001). Competence-based assessment. *Counterpoints, 166*, 453-466.
Wolf, M. K., Everson, P., Lopez, A., Hauck, M., Pooler, E., & Wang, J. (2014). *Building a framework for a next-generation English language proficiency assessment system*. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/ets2.12034/full
Wolfe, E. W., & Smith, E. V. (2007). Instrument development tools and activities for measure validation using Rasch models: Part I-instrument development tools. *Journal of Applied Measurement, 8*(1), 97-123.
Woods, K. (2010). *The design and validation of measures of communication and literacy to support the instruction of students with learning disabilities*. Unpublished PhD Thesis. Melbourne Graduate School of Education, The University of Melbourne.
Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-115.
Wright, B. D. (1995). Diagnose person misfit. *Rasch Measurement Transactions, 9*(2), 430-431. Retrieved from asch.org/rmt/rmt92h.htm
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions, 8*(3), 157-177.
Wright, B. D., Linacre, J. M., Gustafson, J.-E., & Martin-Löf, P. (1994). Reasonable mean square fit values. *Rasch Measurement Transactions, 8*(3), 370. Retrieved from http://www.rasch.org/rmt/rmt83b.htm
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: Mesa Press.
Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
Wu, M. L. (2003). *The application of item response theory to measure problem solving proficiencies*. Unpublished PhD Thesis. Department of Learning and Educational Development, The University of Melbourne.
Wu, M. L., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.
Wu, M. L., Tam, H. P., & Jen, T. (2016). *Educational measurement for applied researchers: Theory into practice*. Singapore: Springer.
Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *LanguageTesting, 31*(4), 501-527.
Yang, R. (2010). *A many-facet Rasch analysis of rater effects on an oral English proficiency test.* (PhD). Purdue University, Indicana, USA,

Yates, G., & Chandler, M. (1991). The cognitive psychology of knowledge: Basic research findings and educational implications. *Australian Journal of Education, 35*, 131-153.
Zhang, Y. (2011). Judgements of oral proficiency by non-native and native English speaking teacher raters: Competing or complementary contructs? *Languag Testing, 28*(1), 31-50.
Zhu, C., & Wang, D. (2014). Key competencies and characteristics for innovative teaching among secondary school teachers: A mixed-methods research. *Asia Pacific Education Review, 15*, 299-311. doi:10.1007/s2564-014-9329-6
Zimmerman, B. J. (1998). Developing self-fulfilling cycles of academic regulation: An analysis of exemplary instructional models. In D. H. Schunk & B. J. Zimmerman (Eds.), *Self-regulated learning: From teaching to self-reflective practice* (pp. 1-19). New York: Guilford.
Zimmerman, B. J. (2000). Attaining self-regulation: A social-cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 13-39). San Diego: Academic Press.
Zimmerman, B. J. (2006). Development and adaptation of expertise: The role of self-regulatory processes and beliefs. In A. S. Erickson, N. Charness, P. J. Feltovich, & R. R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance* (pp. 705-722). Cambridge: Cambridge University Press.

Appendix A. Plain Language Statements

PLAIN LANGUAGE STATEMENT (Focus group)

THE UNIVERSITY OF MELBOURNE

Project title: "The Development and Validation of Competency Standards for English Language Proficiency Test Developers in Vietnamese Higher Education Settings"

Dear Participant,

You are invited to participate in the above research project, which is being conducted by Professor Hamish Coates, Associate Professor Shelley Gillis (supervisors) and Ms. Phuong Viet Ha Ngo (PhD candidate) of the Melbourne Graduate School of Education, at the University of Melbourne. This project is done in fulfilment of Ms. Ngo's PhD dissertation, and has been approved by the Melbourne Graduate School of Education Human Research Ethics Committee.

The aim of the project is to develop and validate a developmental framework that describes the specific knowledge, skills, and transversal competencies required of English Language Proficiency (ELP) test developers to perform major job functions in constructing valid, reliable and fair tests of ELP in Vietnamese higher education settings.

The demands on participants would be a ninety minute focus group discussion on the following aspects of competencies required of ELP test developers:

- Professional expectations, concerns and needs;
- Identifying and confirming the major job functions required of ELP test developers
- Identifying the critical competencies (skills, knowledge, and transversal competencies) required of ELP test developers to perform each of the major job functions;
- Exploring the appropriateness and relevance of these competencies in the context of Vietnam

This study will contribute to supporting the performance and development of teachers' expertise in ELP test development. Your valued input will help support quality development of reliable, valid and fair tests of ELP.

There are no foreseeable risks, other than inconvenience, to participants resulting from participation in this study.

Please be advised that participation in the study is completely voluntary. Should you wish to withdraw at any stage, or to withdraw any unprocessed data you have supplied, you are free to do so without prejudice. All the discussions with participants will be audio-taped. Maintaining confidentiality is important to us. No individuals and/or participating institutions will be identified in any publications from this study. The participants will be referred to by pseudonyms in all publications and interview transcriptions. The confidentiality of your responses will be protected to the fullest extent, within the limits of the law.

HREC: 1648114.1; **Date:23/01/17 Version: 02**.

Melbourne Graduate School of Education
100 Leicester Street | The University of Melbourne Victoria 3010 | Australia
T: +61 3 8344 8285 **F:** +61 3 8344 8213 **W:** www.education.unimelb.edu.au

Data collected during the research process will be stored, according to the University of Melbourne regulations, in private and secure location at the Melbourne Graduate School of Education for a period of five years and will then be destroyed.

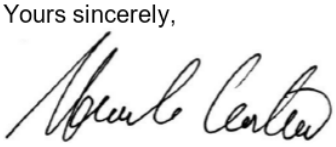
Once the thesis arising from this research has been completed, a brief summary of the findings will be made available to you if your interest is indicated and your contact details are provided in the consent form.

If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Executive Officer, Human Research Ethics, the University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Email: humanethics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team or the name or ethics ID number of the research project.

If you would like to voluntarily participate in this study, could you please respond to the student researcher, Phuong Ngo, by <<insert date>>, either through phone 0405 971 525 or email nvngo@student.unimelb.edu.au. The student researcher will then contact you to arrange a focus group interview with you at a mutually convenient time. You will indicate that you have read and understood this information by signing the accompanying consent form before participating in the data collection procedure.

We trust that you will be able to assist with this project, and we look forward to working with you.

Yours sincerely,



Professor Hamish Coates
Melbourne Centre for the Study of Higher Education
Elisabeth Murdoch Building (Building 134), Spencer Road
The University of Melbourne, Victoria 3010 Australia
T: +61 3 9035 4928 | **M:** +61 423 475 605 | **E:** h.coates@unimelb.edu.au
W: melbourne-cshe.unimelb.edu.au

Associate Professor Shelley Gillis
Centre for Vocational and Educational Policy
Melbourne Graduate School of Education
The University of Melbourne
T: (03) 9035 4925
E: shelley.gillis@unimelb.edu.au

Phuong Ngo (PhD Candidate)
Melbourne Centre for the Study of Higher Education
Elisabeth Murdoch Building, Spencer Road
The University of Melbourne, Victoria 3010 Australia
M: +61 405 971 526
E: nvngo@student.unimelb.edu.au

PLAIN LANGUAGE STATEMENT (Workshops)

THE UNIVERSITY OF MELBOURNE

Project title: "The Development and Validation of Competency Standards for English Language Proficiency Test Developers in Vietnamese Higher Education Settings"

Dear Participant,

You are invited to participate in the above research project, which is being conducted by Professor Hamish Coates, Associate Professor Shelley Gillis (supervisors) and Ms. Phuong Viet Ha Ngo (PhD candidate) of the Melbourne Graduate School of Education, at the University of Melbourne. This project is done in fulfilment of Ms. Ngo's PhD dissertation, and has been approved by the Melbourne Graduate School of Education Human Research Ethics Committee.

The aim of the project is to develop and validate a developmental framework that describes the specific knowledge, skills, and transversal competencies required of English Language Proficiency (ELP) test developers to perform major job functions in constructing valid, reliable and fair tests of ELP in Vietnamese higher education settings.

The demands on participants would be two two-hour workshops on writing the rubrics for the competencies required of ELP test developers, including:

- Identifying the performance indicators for each of the competencies required of ELP test developers;
- Developing the performance levels for each indicator and set them along a progressing continuum of difficulty;

This study will contribute to supporting the performance and development of teachers' expertise in ELP test development. Your valued input will help support quality development of reliable, valid and fair tests of ELP.

There are no foreseeable risks, other than inconvenience, to participants resulting from participation in this study.

Please be advised that participation in the study is completely voluntary. Should you wish to withdraw at any stage, or to withdraw any unprocessed data you have supplied, you are free to do so without prejudice. All the discussions with participants will be audio-taped. Maintaining confidentiality is important to us. No individuals and/or participating institutions will be identified in any publications from this study. The confidentiality of your responses will be protected to the fullest extent, within the limits of the law.

Data collected during the research process will be stored, according to the University of Melbourne regulations, in private and secure location at the Melbourne Graduate School of Education for a period of five years and will then be destroyed.

HREC: 1648114.1; **Date:23/01/17**; **Version: 02**.

Melbourne Graduate School of Education
100 Leicester Street | The University of Melbourne Victoria 3010 | Australia
T: +61 3 8344 8285 **F:** +61 3 8344 8213 **W:** www.education.unimelb.edu.au

Once the thesis arising from this research has been completed, a brief summary of the findings will be made available to you if your interest is indicated and your contact details are provided in the consent form.

If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Executive Officer, Human Research Ethics, the University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Email: HumanEthics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team or the name or ethics ID number of the research project.

If you would like to voluntarily participate in this study, could you please respond to the student researcher, Phuong Ngo, by <<*insert date*>>, either through phone 0405 971 525 or email nvngo@student.unimelb.edu.au. The student researcher will then contact you to arrange the workshops with you at a mutually convenient time. You will indicate that you have read and understood this information by signing the accompanying consent form before participating in the data collection procedure.

We trust that you will be able to assist with this project, and we look forward to working with you.

Yours sincerely,

Professor Hamish Coates
Centre for the Study of Higher Education
Elisabeth Murdoch Building (Building 134), Spencer Road
The University of Melbourne, Victoria 3010 Australia
T: +61 3 9035 4928 | **M:** +61 423 475 605 | **E:** h.coates@unimelb.edu.au
W: melbourne-cshe.unimelb.edu.au

On behalf of:

Associate Professor Shelley Gillis
Centre for Vocational and Educational Policy
Melbourne Graduate School of Education
The University of Melbourne
T: (03) 9035 4925
E: shelley.gillis@unimelb.edu.au

Phuong Ngo
Centre for the Study of Higher Education
Elisabeth Murdoch Building, Spencer Road
The University of Melbourne, Victoria 3010
Australia
M: +61 405 971 526 |
E: nvngo@student.unimelb.edu.au

PLAIN LANGUAGE STATEMENT (Online survey)

THE UNIVERSITY OF MELBOURNE

Project title: "The Development and Validation of Competency Standards for English Language Proficiency Test Developers in Vietnamese Higher Education Settings"

Introduction

This research project is being conducted by Professor Hamish Coates, Associate Professor Shelley Gillis (supervisors) and Ms. Phuong Viet Ha Ngo (PhD candidate) of the Melbourne Graduate School of Education, at the University of Melbourne in fulfilment of Ms. Ngo's PhD dissertation, and has been approved by the Melbourne Graduate School of Education Human Research Ethics Committee.

The aim of the project is to develop and validate a developmental framework that describes the specific knowledge, skills, and transversal competencies required of English Language Proficiency (ELP) test developers to perform major job functions in constructing valid, reliable and fair tests of ELP in Vietnamese higher education settings.

What will I be asked to do?
You will be asked to complete an online questionnaire in two rounds. The questionnaire will take approximately 10 minutes. Your valued input will help to rank sets of competencies for ELP test developers along a developmental learning continuum.

What are the possible risks?
This research study is low risk. If you do experience any discomfort during the online survey you are free to stop the process at any time.

How will my confidentiality be protected?
This survey is completely anonymous and the confidentiality of your responses will be protected to the fullest possible extent, within the limits of the law.
Data collected during the research process will be stored, according to the University of Melbourne regulations, in private and secure location at the Melbourne Graduate School of Education for a period of five years and will then be destroyed.

Will I hear about the results of this project?
Once the thesis arising from this research has been completed, a brief summary of the findings will be made available to you if your interest is indicated and your contact details are provided in the consent form.

Where can I get further information?
If you would like more information about the project, please contact the researchers through the details given below.

Who can I contact if I have any concerns about the project?
If you have any concerns or complaints about the conduct of this research project, which you do not wish to discuss with the research team, you should contact the Manager,

HREC: 1648114.1; Date: 23/01/17 Version: 02.

Melbourne Graduate School of Education
100 Leicester Street | The University of Melbourne Victoria 3010 | Australia
T: +61 3 8344 8285 **F:** +61 3 8344 8213 **W:** www.education.unimelb.edu.au

Human Research Ethics, Office for Research Ethics and Integrity, University of Melbourne, VIC 3010. Tel: +61 3 8344 2073 or Email: humanethics-complaints@unimelb.edu.au. All complaints will be treated confidentially. In any correspondence please provide the name of the research team or the name or ethics ID number of the research project.

How do I agree to participate?
We do hope that you will be able to assist with this project. Please follow the link to start the survey: <<*insert link*>>.

Professor Hamish Coates
Melbourne Centre for the Study of Higher Education
Elisabeth Murdoch Building (Building 134), Spencer Road
The University of Melbourne, Victoria 3010 Australia
T: +61 3 9035 4928 | **M:** +61 423 475 605 | **E:** h.coates@unimelb.edu.au
W: melbourne-cshe.unimelb.edu.au

On behalf of the research team, as follows:

Associate Professor Shelley Gillis
Centre for Vocational and Educational Policy
Melbourne Graduate School of Education
The University of Melbourne
T: (03) 9035 4925
E: shelley.gillis@unimelb.edu.au

Phuong Ngo
Centre for the Study of Higher Education
Elisabeth Murdoch Building, Spencer Road
The University of Melbourne, Victoria 3010
Australia
M: +61 405 971 526 |
E: nvngo@student.unimelb.edu.au

Appendix B. Consent Forms

CONSENT FORM (Focus group)

PROJECT TITLE: "The Development and Validation of Competency Standards for English Language Proficiency Test Developers in Vietnamese Higher Education Settings"

Name of Participant: _____
Names of investigators: Professor Hamish Coates, Associate Professor Shelley Gillis, and Phuong Viet Ha Ngo (PhD Candidate)

1. I understand that this project is for research purposes, and is undertaken by the researchers and a PhD candidate in the Melbourne Graduate School of Education, at the University of Melbourne.

2. I understand the nature of the project and what is expected of me, and I consent to participate on that basis.

3. I acknowledge that:

   a) I have read the written information about the project and have received a copy of that information;

   b) I have received an adequate explanation of all likely risks, effects, discomforts or inconvenience arising from participation in the project.

   c) my participation is voluntary and I have the right to withdraw form participation at any time and that I may withdraw any data I have supplied (up to the point of analysis/publication);

   d) the focus group discussion will be audio-recorded for the purpose of the research only;

   e) the researchers will keep this signed copy of my consent form;

   f) I am satisfied that the confidentiality of the information I have provided will be safeguarded subject of any legal limitations;

   g) I will not be identified in any publication arising from the research. Any information used in this research that might identify me or my workplace will be changed and a pseudonym will be used to protect my identity.

4. I would like to receive a copy of the study results via email or post, I have provided my details below and ask that they be used for this purpose only; *(If NOT, please leave this blank)*

   Name: _____

   Address: _____

   Email Address: _____

   Participant signature: _____ **Date:** _____

HREC: 1648114.1; **Date:23/01/17**; **Version:** 02.

Melbourne Graduate School of Education
100 Leicester Street | The University of Melbourne Victoria 3010 | Australia
T: +61 3 8344 8285 **F:** +61 3 8344 8213 **W:** www.education.unimelb.edu.au

CONSENT FORM (Workshops)

PROJECT TITLE: "The Development and Validation of Competency Standards for English Language Proficiency Test Developers in Vietnamese Higher Education Settings"

Name of Participant: _____

Names of investigators: Professor Hamish Coates, Associate Professor Shelley Gillis, and Phuong Viet Ha Ngo (PhD Candidate)

1. I understand that this project is for research purposes, and is undertaken by the researchers and a PhD candidate in the Melbourne Graduate School of Education, at the University of Melbourne.

2. I understand the nature of the project and what is expected of me, and I consent to participate on that basis.

3. I acknowledge that:

   a) I have read the written information about the project and have received a copy of that information;

   b) I have received an adequate explanation of all likely risks, effects, discomforts or inconvenience arising from participation in the project.

   c) my participation is voluntary and I have the right to withdraw form participation at any time and that I may withdraw any data I have supplied (up to the point of analysis/publication);

   d) my ideas contributed in the workshops will be recorded in written format only;

   e) the researchers will keep this signed copy of my consent form;

   f) I am satisfied that the confidentiality of the information I have provided will be safeguarded subject of any legal limitations;

   g) I will not be identified in any publication arising from the research. Any information used in this research that might identify me or my workplace will be changed and a pseudonym will be used to protect my identity.

4. I would like to receive a copy of the study results via email or post, I have provided my details below and ask that they be used for this purpose only; *(If NOT, please leave this blank)*

   Name: _____

   Address: _____

   Email Address: _____

   Participant signature: _____ **Date:** _____

HREC: 1648114.1; **Date:23/01/17**; **Version:** 02.

Melbourne Graduate School of Education
100 Leicester Street | The University of Melbourne Victoria 3010 | Australia
T: +61 3 8344 8285 **F:** +61 3 8344 8213 **W:** www.education.unimelb.edu.au

CONSENT FORM (Online survey)

PROJECT TITLE: "The Development and Validation of Competency Standards for English Language Proficiency Test Developers in Vietnamese Higher Education Settings"

Name of Participant: _____

Names of investigators: Professor Hamish Coates, Associate Professor Shelley Gillis, and Phuong Viet Ha Ngo (PhD Candidate)

1. I understand that this project is for research purposes, and is undertaken by the researchers and a PhD candidate in the Melbourne Graduate School of Education, at the University of Melbourne.

2. I understand the nature of the project and what is expected of me, and I consent to participate on that basis.

3. I acknowledge that:

   a) I have read the written information about the project and have received a copy of that information;

   b) I have received an adequate explanation of all likely risks, effects, discomforts or inconvenience arising from participation in the project.

   c) my participation is voluntary and I have the right to withdraw form participation at any time and that I may withdraw any data I have supplied (up to the point of analysis/publication);

   d) the researchers will keep this signed copy of my consent form;

   e) I am satisfied that the confidentiality of the information I have provided will be safeguarded subject of any legal limitations;

   f) My identity is completely anonymous and will be protected to the fullest extent, within the limits of the law.

4. I would like to receive a copy of the study results via email or post, I have provided my details below and ask that they be used for this purpose only; *(If NOT, please leave this blank)*

   Name: _____

   Address: _____

   Email Address: _____

   Participant signature: _____ **Date:** _____

HREC: 1648114.1; **Date:**23/01/17; **Version:** 02.

Melbourne Graduate School of Education
100 Leicester Street | The University of Melbourne Victoria 3010 | Australia
T: +61 3 8344 8285 **F:** +61 3 8344 8213 **W:** www.education.unimelb.edu.au

Appendix C. Ethics Approval

15 March 2017

Prof H.B. Coates
Centre for the Study of Higher Education
Melbourne Graduate School of Education
The University of Melbourne

Dear Professor Coates

I am pleased to advise that the Humanities and Applied Sciences Human Ethics Sub-Committee approved the following Project:

Project title:	**The development and validation of competency standards for English Language Proficiency (ELP) test developers in Vietnamese higher education settings**
Researchers:	**Prof H B Coates, P Ngo, A/Prof S A Gillis**
Ethics ID:	**1648114**

The Project has been approved for the period: **13-Mar-2017 to 31-Dec-2017**

It is your responsibility to ensure that all people associated with the Project are made aware of what has actually been approved.

Research projects are normally approved to 31 December of the year of approval. Projects may be renewed yearly for up to a total of five years upon receipt of a satisfactory annual report. If a project is to continue beyond five years a new application will normally need to be submitted.

Please note that the following conditions apply to your approval. Failure to abide by these conditions may result in suspension or discontinuation of approval and/or disciplinary action.

(a) **Limit of Approval:** Approval is limited strictly to the research as submitted in your Project application.

(b) **Variation to Project:** Any subsequent variations or modifications you might wish to make to the Project must be notified formally to the Human Ethics Sub-Committee for further consideration and approval. If the Sub-Committee considers that the proposed changes are significant, you may be required to submit a new application for approval of the revised Project.

(c) **Incidents or adverse effects:** Researchers must report immediately to the Sub-Committee anything which might affect the ethical acceptance of the protocol including adverse effects on participants or unforeseen events that might affect continued ethical acceptability of the Project. Failure to do so may result in suspension or cancellation of approval.

(d) **Monitoring:** All projects are subject to monitoring at any time by the Human Research Ethics Committee.

(e) **Annual Report:** Please be aware that the Human Research Ethics Committee requires that researchers submit an annual report on each of their projects at the end of the year, or at the conclusion of a project if it continues for less than this time. Failure to submit an annual report will mean that ethics approval will lapse.

(f) **Auditing:** All projects may be subject to audit by members of the Sub-Committee.

If you have any queries on these matters, or require additional information, please contact me using the details below.

Please quote the ethics registration number and the title of the Project in any future correspondence.

On behalf of the Sub-Committee I wish you well in your research.

Yours sincerely

Ms Jennifer Hassell - Secretary
Humanities and Applied Sciences HESC
Phone: 90353341, Email: hassell@unimelb.edu.au

RESEARCH, INNOVATION & COMMERCIALISATION
Incorporating UoM Commercial Ltd and the School of Melbourne Custom Programs
The University of Melbourne, Victoria 3010, Australia
T: +61 3 8344 1539 (external) **T:** 40777 (internal) **W:** unimelb.edu.au/research

THE UNIVERSITY OF MELBOURNE

unimelb.edu.au

Appendix D. The Questionnaire

Vietnamese Teacher Competence in Multiple-Choice Test Item Writing

Preamble and instructions for the survey process

The research, ***"Understanding teacher competence in multiple-choice test item writing for English reading and listening skill tests: A case of EFL teachers in Vietnamese higher education settings"****, undertaken by Ngo Viet Ha Phuong, at the University of Melbourne in fulfilment of her PhD dissertation, aims to construct an understanding of Vietnamese teacher competence in multiple-choice test item writing for English reading and listening skill tests and explore teacher-related factors contributing the competence. This research has been approved by the Melbourne Graduate School of Education Human Research Ethics Committee.*

We invite you to complete an anonymous online survey which will take 15 minutes to complete. If you agree to voluntarily participate, you will be asked to complete a self-report assessment against the indicators of the draft Developmental Framework for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests, and to provide information about yourself, and your engagement in and access to professional learning.

Your responses will be anonymous and will not be used individually. You may withdraw from the survey at any time. The survey has also been designed to allow you to start and resume at another time as long as you use the same device. However, as the survey is anonymous, it will not be possible to remove your response from the database set collected should you wish to withdraw them at a later date.

Should you require any further information or have any concerns, please contact Ms. Ngo Viet Ha Phuong at nvngo@student.unimelb.edu.au.
If you agree to participate in the survey, please complete the questions that follow.

PART 1: SELF - REPORT ASSESSMENT

Please respond to the questions below by ticking only one option. You should select the highest level that best describes your typical behaviour.

1. Which of the following options best describes how you tend to **review test specifications**?

□ I check aspects of the test specification (e.g., the skills/subskills to be tested, text length, text type, item formats, quantity, and item difficulty levels).

□ I retrieve from prior experience important aspects of the test to be written.

□ I consider operational implications of the test specification with target examinees in mind.

2. Which of the following options best describes how you tend to **utilise test specifications**?

□ I follow the instructions from the test specifications.

□ I refer to other training materials or resources (e.g., sample tests) to clarify the test specifications.

□ I use experience and available resources to internalise the test specifications.

□ I develop own internal guideline for item writing in accordance with the given test specification and informed by other training materials/resources and prior experiences.

3. Which of the following options best describes how you tend to **search for input materials for tests**?

□ I browse materials of required length on recommended topics

□ I identify texts of specified complexity and difficulty from newspaper, magazines and books written for non-native English speakers.

□ I locate texts from previously used high-yield sources.

□ I source texts from a diverse range of resources with consideration of the target examinees.

4. Which of the following options best describes how you tend to **evaluate the input materials for text-fit-specifications**?

□ I use text analysis tools (e.g., English Profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine text-fit-specifications in terms of difficulty and complexity.

□ I assess texts based on defined characteristics of text contents and difficulty levels.

□ I use multiple sources of evidence regarding text density, lexical difficulty, text potentiality for item development (see items leap from the text) to judge the relevance of input materials.

5. Which of the following options best describes how you tend **to make textual changes to the selected input materials for tests**?

□ I restructure and cut texts to required length.

□ I make lexical and syntactic changes to achieve the target register.

□ I employ an extensive repertoire of strategies to adapt the original text to meet the required textual features in terms of language use, style and their rhetorical functions.

6. Which of the following options best describes **how you tend to adapt texts for texts-fit-items and items-fit-texts?**

□ I design items closely adhering to the generated text/script.

□ I use research- grounded factors/ruses (e.g., use of paraphrase, antonyms and synonyms to avoid dictation effects, add distraction to text/scripts to reduce guessability, word-form shift, secondary text supplement) to adjust texts to fit items.

□ I call upon a range of strategies to transform source texts to fit specified test purposes and test items requirements.

7. Which of the following options best describes **your approach to item writing tasks?**

□ I follow a standardised text/script-first item writing process.

□ I focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items.

□ I multi-task between text/script adaptation and item construction.

8. Which of the following options best describes how you tend to **write item stems**?

□ I replicate items from sample tests provided in the test specifications.

□ I construct items from a pool of item shells/item stems derived from established items in past test materials.

□ I rely on different sources to create items accommodating test specifications and input materials.

9. Which of the following options best describes how you tend to **write item distractors**?

□ I follow the format of the distractors given in the test specifications or sample tests.

□ I write distractors with reference to guidelines on Dos and Don'ts for writing test item options.

□ I utilise a range of paraphrasing strategies to construct distractors.

□ I use flexible repertoire of strategies to write distractors with consideration of typical errors or misconceptions of the subject content by test takers.

10. Which of the following options best describes how you tend to **provide answer keys** for the items you write?

□ I assign an answer key for each item.

□ I provide a reference (e.g., underline chunks of texts from reading passages, listening scripts that helps to justify the answer key) for each answer key.

□ I justify correct and incorrect answer options based on the evidence from text/script.

11. Which of the following options best describes how you tend to **decide on item difficulty**?

□ I align items with the descriptors defined in the item blueprint and/or in the six level English language proficiency framework for Vietnamese learners.

□ I consider the text complexity and cognitive demand/sub-skills required to answer each item.

□ I draw on a range of factors affecting item difficulty from experience, expert colleagues and contemporary research to make judgements of item difficulty.

12. Which of the following options best describes how you tend to **check item content fit**?

□ I examine the conformity of the items to the test specifications regarding item content, wording of items and instructions.

□ I consider abilities required to complete the items to compare what the items are thought to be testing with what is claimed in the test specifications.

□ I attempt items on own to make judgements of item-fits-specifications for content complexity and difficulty levels.

13. Which of the following options best describes how you tend to **verify answer keys and distractors?**

□ I try out items on own to check the original keys.

□ I analyse responses from a self-conducted pilot with a sample of students to identify any miskeys and/or poor performing distractors.

14. Which of the following options best describes how you tend to **respond to feedback on your items?**

□ I incorporate received feedback.

□ I consider the test purposes to make the decision to implement feedback.

□ I use information from experience and research to justify decision to implement or not to implement received feedback.

PART 2: ENGAGEMENT IN PROFESSIONAL LEARNING

1. How often have you engaged in the following professional activities since you were involved in writing items for VSTEPs? *(Please circle the one that best describes you)*

1= never 2= rarely 3= sometimes 4= often

Activities	Frequency Scale			
1.1 I study guidelines provided to item writers.	1	2	3	4
1.2 I read item writing manuals from online resources.	1	2	3	4
1.3 I read research publications on multiple-choice test item writing.	1	2	3	4
1.4 I participate in compulsory training workshops organised by my institutions for teacher item writers.	1	2	3	4
1.5 I attend professional learning opportunities outside my institutions.	1	2	3	4
1.6 I conduct/co-conduct training sessions for item writers.	1	2	3	4
1.7 I tend to write items in a similar way to how I have written then in the past.	1	2	3	4
1.8 I reflect on/recall prior item writing experiences to suggest lessons learnt for future tasks.	1	2	3	4
1.9 I record/document my own best practices in multiple-choice test item writing.	1	2	3	4
1.10 I utilise my previously used strategies for multiple-choice test item writing.	1	2	3	4
1.11 I adjust my test item writing strategies and approaches based on lessons learnt from past experiences.	1	2	3	4
1.12 I experiment new strategies and approaches to multiple-choice test item writing.	1	2	3	4
1.13 I talk about problems with writing multiple-choice test items with colleagues.	1	2	3	4
1.14 I share resources and ideas about multiple-choice test item writing with colleagues.	1	2	3	4
1.15 I support or coach colleagues in multiple-choice test item writing.	1	2	3	4

2. Please indicate if the following induction and training activities are provided to you by your institutions for your development of expertise in multiple-choice test item writing. ***(****Please circle Yes or No)*

2.1 I am provided with a documentation of the task description.	Yes	No
2.2 I am provided with sample tests.	Yes	No
2.3 I am provided with additional guidelines for item writers.	Yes	No

2.4 I am briefed about the test specification.	Yes	No
2.5 I have opportunities to discuss the test specifications with colleagues for further clarification.	Yes	No
2.6 I am provided with a training agenda (i.e., schedule of topics to be presented and times) before the training.	Yes	No
2.7 I am provided with a description of the training with learning objectives prior to training.	Yes	No
2.8 I am provided with readings, activities and/or questions to think about prior to the training.	Yes	No
2.9 I am trained about the principles of multiple-choice test item writing.	Yes	No
2.10 I am trained about techniques for writing different objective test item types	Yes	No
2.11 I have opportunities to engage in discussions with colleagues on item writing techniques.	Yes	No
2.12 I am provided with example test items to analyse.	Yes	No
2.13 I have opportunities to consider the strengths and weaknesses of the items with colleagues.	Yes	No
2.14 I have opportunities to work with colleagues to identify features of sound items.	Yes	No
2.15 I have opportunities to engage in hands-on item writing practice tasks (e.g., select and adapt texts, write item stems, write distractors, estimating cognitive demands for answering each item …) with colleague item writers of varying experience.	Yes	No
2.16 I have opportunities to analyse recent live test papers.	Yes	No
2.17 I have opportunities to engage in feedback on live test item performance with colleagues.	Yes	No
2.18 I have opportunities to learn about basic measurement characteristics of test items (e.g., item difficulty, and item discrimination indices).	Yes	No
2.19 I am trained about using different tools and resources (e.g., CEFR, text analysis tools) to support my item writing.	Yes	No
2.20 I am trained by relevant trainers/have opportunities to work with and learn from experienced/ well qualified trainers.	Yes	No

2.21 I have opportunities to participate in training and development workshops for teacher item writers 1-2 working weeks a year. Yes No

PART 3: BACKGROUND INFORMATION

1. What is your highest academic qualification?

□ Bachelor □ Master □ Doctorate

2. Which discipline best describes your highest degree?

□ TESOL/TEFL

□ Applied Linguistics

□ Education

□ Language Testing and Assessment

□ Others

3. Have you ever taken any coursework in language testing and assessment as part of your university studies?

□Yes □ No

4. If Yes, how well did the coursework prepare you for your multiple-choice test item writing task?

□ Not at all □ Somewhat □ Fairly well □ Very well

5. What is the total number of years you have been working as EFL teachers (either in your current university or (an) other(s)?.....

6. What is the total number of years you have been teaching EFL exam preparation courses (e.g., IELTS, TOEFL iBT, TOEIC, VSTEP, Cambridge Certificate Exams)? .....

7. What is the approximate total number of VSTEP tests you have written so far?

.....

8. What is your English language proficiency level*? (Please circle the test score that applies to you)*

IELTS band score equivalents	5.5	6.0	6.5	7.0	7.5	8.0	8.5	9.0

9. Your current working location: …….

10. Your gender:

□ Male □ Female

Thank you for completing the survey.

Your contribution is greatly appreciated by the research team at the University of Melbourne.

Appendix E. The Initial Draft Framework for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests

PHASES	COMPETENCIES	PERFORMANCE INDICATORS
THE EXPLORATORY PHASE	1.1 Analysing test specifications	1.1.1 Review test specifications (Alderson et al., 1995; Davison & Lynch, 2003, Kim et al., 2010; Lynch & Davison, 1994; Salisbury, 2005)
		1.1.2 Utilise test specifications (Alderson et al., 1995; Davison & Lynch, 2003, Kim et al., 2010; Lynch & Davison, 1994; Salisbury, 2005)
	1.2 Generating materials for tests	1.2.1 Select source texts/materials for the test (Green & Hawkey, 2011; Salisbury, 2005; Sydorenko, 2011)
		1.2.2 Evaluate source texts to make decisions (Green & Hawkey, 2011; Salisbury, 2005; Sydorenko, 2011)
		1.2.3 Pre-edit the generated texts for item writing (Green & Hawkey, 2011; Salisbury, 2005; Sydorenko, 2011)
THE CONCERTED PHASE	2.1 Writing test items	2.1.1 Approach item writing (Salisbury, 2005)
		2.1.2 Write item stems (ALTE, 2005; Beddow, 2012; Haladyna & Downing, 1989; Haladyna et al., 2002; Haladyna & Rodriguez, 2013; Rodriguez, 2016; Moreno et al., 2015)
		2.1.3 Write item distractors (ALTE, 2005; Beddow, 2012; Haladyna & Downing, 1989; Haladyna et al., 2002; Haladyna & Rodriguez, 2013; Moreno et al., 2015; Rodriguez, 2016)
		2.1.4 Provide answer keys (ALTE, 2005; Beddow, 2012; Haladyna & Downing, 1989; Haladyna et al., 2002; Haladyna & Rodriguez, 2013; Moreno et al., 2015; Rodriguez, 2016)
		2.1.5 Estimate item difficulty (Alderson et al., 1995; Sydorenko, 2011)

PHASES	COMPETENCIES	PERFORMANCE INDICATORS
	2.2 Reviewing test items	2.2.1 Review test items (AERA et al., 2014 ; Alderson et al., 1995 ; ALTE, 2005 ; Baranowski, 2006 ; Haladyna & Rodriguez, 2013 ; Lane et al., 2016 ; Livingston, 2006 ; Maau-Aduli & Zimitat, 2012; Osterlind, 1998; Raymond, 2015; Wise & Blake, 2016)
		2.2.2 Verify answer keys and distractors (Alderson et al., 1995 ; AERA et al., 2014 ; ALTE, 2005 ; Baranowski, 2006 ; Haladyna & Rodriguez, 2013 ; Lane et al., 2016 ; Livingston, 2006 ; Maau-Aduli & Zimitat, 2012 ; Osterlind, 1998 ; Raymond, 2015 ; Wise & Blake, 2016)
THE REFINING PHASE	3.1 Revising test items	3.1.1 Respond to item feedback (Alderson et al., 1995; Salisbury, 2005)

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
THE EXPLORATORY PHASE		
1.1 Analysing test specifications	1.1.1 Review test specifications	1.1.1.1Check the test purposes, the test format, the distribution of test items and their targeted levels within the constructs (Salisbury, 2005) 1.1.1.2 Identify targeted test item content, and levels of item difficulty (Salisbury, 2005) 1.1.1.3 Recall important aspects of the test to be written/ Generate ideas about what and how to write the items for the test (Salisbury, 2005) 1.1.1.4 Use mental models of target examinees to trigger operational consideration of the test specifications or Consider/contextualise operational implications of the test specifications (Salisbury, 2005)
	1.1.2 Utilise test specifications	1.1.2.1 Follow the instructions from the test specifications (Davidson & Lynch, 2002; Kim et al., 2010; Salisbury, 2005) 1.1.2.2 Recall my prior (item writing) experiences to interpret the test specifications (Davidson & Lynch, 2002; Kim et al., 2010; Salisbury, 2005) 1.1.2.3 Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications (Davidson & Lynch, 2002; Kim et al., 2010; Salisbury, 2005) 1.1.2.4 Recall previous experience and discuss with colleagues to internalise the content of the test specifications (Davidson & Lynch, 2002; Kim et al., 2010; Salisbury, 2005) 1.1.2.5 Refer to own internal guideline for item writing based on the given test specifications and/or other training materials/resources and prior experiences (Davidson & Lynch, 2002; Kim et al., 2010; Salisbury, 2005)
1.2 Generating materials for tests	1.2.1 Select source text	1.2.1.1 Select text of suitable length (Green & Hawkey, 2011) 1.2.1.2 Select authentic texts from prestigious sources based on an inventory of criteria/characteristics and/or guidelines (Green & Hawkey, 2011) 1.2.1.3 Choose potential texts from published test books (Green & Hawkey, 2011) 1.2.1.4 Search the Internet for materials of own interest 1.2.1.5 Locate texts from previously used high-yield sources (Green & Hawkey, 2011) 1.2.1.5 Source texts from new resources (Green & Hawkey, 2011)

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
	1.2.2 Evaluate source texts to make decisions	1.2.2.1 Seek English native colleague judgements of materials (Kim et al. 2010) 1.2.2.2 Use measures (e.g., English profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine source texts meet requirements for level of lexical difficulty and textual complexity (Green & Hawkey, 2011) 1.2.2.3 Judge the relevance of source texts based on own common sense/intuition/ "see items leap out from the text" (Green & Hawkey, 2011) 1.2.2.4 Assess texts based on defined characteristics of text contents and difficulty levels (Green & Hawkey, 2011)
	1.2.3 Pre-edit the generated texts for item writing	1.2.3.1 Retain the original texts (Green & Hawkey, 2011) 1.2.3.2 Cut texts to required length (Green & Hawkey, 2011) 1.2.3.3 Make lexical and syntactic changes to achieve the desirable register (e.g., to make the text more academic or colloquial) (Green & Hawkey, 2011; Sydorenko, 2011) 1.2.3.4 Amend texts for improved clarity and writing quality (e.g., removing or glossing cultural references or parts requiring technical knowledge; deleting unsuitable offensive contents, deleting irrelevant elements of texts, cutting redundancies and repetitions) (Green & Hawkey, 2011; Sydorenko, 2011)
THE CONCERTED PHASE		
2.1. Developing test items	2.1.1 Approach item writing	2.1.1.1 Follow the standardised text-first item writing process (i.e., to generate texts first and then write items) /use the generated texts/scripts as the basis to write items in accordance with the test requirements (text-first approach) (Salisbury, 2005) 2.1.1.2 Focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items (item-first approach)/ and arrange texts/scripts to fit items/use item-first approach (Salisbury, 2005) 2.1.1.3 Multi-task between text/script adaptation and item construction/ Work with texts and items simultaneously (Salisbury, 2005)
	2.1.2 Write item stems	2.1.2.1 Replicate items from sample tests provided in the test specification (Alderson et al., 1995; Haladyna &Rodriguez, 2013)

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
		2.1.2.2 Construct items from a pool of item shells/item stems derived from established items in past test materials (Alderson et al. 1995; Haladyna & Rodriguez, 2013) 2.1.2.3 Rely on different sources (e.g., test specifications, past exam papers, examples of good items, previously validated items, guidelines) to create items accommodating test specifications and input materials (Alderson et al., 1995; Beddow, 2012; Griffin & Care, 2014; Haladyna &Rodriguez, 2013; Rodriguez, 2016) 2.1.2.4 Build on various resources to innovate items (Beddow, 2012; Elliott et al., 2010; Kettler et al., 2011 cited in Rodriguez, 2016; Rodriguez, 2016)
	2.1.3 Write item distractors	2.1.3.1 Follow the format of the distractors given in the test specifications/ Write test items with adherence to those from past papers (Haladyna &Rodriguez, 2013) 2.1.3.2 Write distractors with reference to guidelines on Dos and Don'ts for test item writing (Haladyna &Rodriguez, 2013) 2.1.3.3 Utilise a range of paraphrasing strategies (e.g., antonyms, synonyms, change sentence structures, change word forms, change word order or use referents) to construct distractors (Haladyna &Rodriguez, 2013) 2.1.3.4 Use typical errors or misconceptions of test takers to write distractors/ Draw on subject matter knowledge and experience with target examinees to write distractors that reflect common errors or misconceptions (Haladyna &Rodriguez, 2013)
	2.1.4 Provide answer keys	2.1.4.1 Assign an answer key for each item (Salisbury, 2005) 2.1.4.2 Provide a reference (e.g., underline chunks of texts from reading passages, listening scripts that help to justify the answer key) for each answer key (Salisbury, 2005)
	2.1.5 Estimate item difficulty	2.1.5.1 Judge item difficulty and discrimination based on experience and common sense as English language teacher (Alderson et al., 1995; Osterlind, 1998) 2.1.5.2 Consider the text complexity and cognitive demand/sub-skills (e.g., reading for factual information, listening for details, and making inferences from text) required to answer each item (Alderson et al., 1995; Osterlind, 1998) 2.1.5.3 Measure item difficulty based on descriptive statistics of piloted data (Mean, Mode, Median, Range, Standard deviation, p-value) (Alderson et al., 1995;

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
		Anderson & Morgan, 2008; Haladyna & Downing, 1993; Haladyna & Rodriguez, 2013; Osterlind, 1998; Raymond, 2015) 2.1.5.4 Use a model –based method (e.g., classical test theory, item response theory) and relevant item analysis programs to compute indexes regarding the distribution of item difficulties, item discrimination and distractor performance (Alderson et al., 1995; Anderson & Morgan, 2008; Haladyna & Downing, 1993; Haladyna & Rodriguez, 2013; Osterlind, 1998; Raymond, 2015)
2.2 Reviewing items	2.2.1 Review test items	2.2.1.1 Edit and proof tests using an editorial style guide (Lane et al., 2016) 2.2.1.2 Check tests for flaws in readability and mechanics (Lane et al., 2016) 2.2.1.3 Check items using a pre-established reviewing checklist (Lane et al., 2016) 2.2.1.4 Conduct collaborative item reviews 1. Have the items reviewed by the target examinees (Baranowski, 2006) 2. Collaborate with colleagues to proofread tests (Malau-Aduli & Zimitat, 2012; Lane et al., 2016) 3. Engage in editorial reviews of items with at least one English native colleague for language naturalness, and authenticity this depends on resources available not person's expertise) (Downing, 2006; Lane et al., 2016) 4. Collaborate with (a) professional editor(s) to edit and proof items using an editorial style guide (Downing, 2006; Lane et al., 2016) 5. Internally review test items for obvious flaws in readability and mechanics (Alderson et al., 1995; Haladyna & Rodriguez, 2013) 6. Review and critique test items collaboratively a SME, an item-writing team leader and an English native colleague for naturalness, understandability and item-fit-specs (Alderson et al., 1995; Haladyna & Rodriguez, 2013) 7. Try out test items as embedded items in operational tests with external SMEs to provide judgements on appropriateness of content, linguistic complexity/difficulty, correctness of answer and verification of reference (Haladyna & Rodriguez, 2013) 8. Co-ordinate rigorous and separate independent reviews with a panel of various expertness to collect constructive and critical feedback and provide justification on test items (Alderson et al., 1995; Haladyna & Rodriguez, 2013)

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
		9 Pre-test drafted items in a test form with targeted representative sample of examinees to collect item responses for analysis and moderation (Alderson et al., 1995; Haladyna & Rodriguez, 2013) 10. Review items based on statistical analyses of item responses (Anderson & Morgan, 2008; Alderson et al., 1995; Raymond, 2015)
	2.2.2 Verify distractors and answer keys	2.2.3.1 Identify distractor discrimination is through statistical analysis of responses to item distractors (Alderson et al., 1995; Haladyna & Downing, 1993; Haladyna & Rodriguez, 2013)
THE REFINING PHASE		
3.1 Revising test items	3.1.2 Make decisions to respond to feedback/Respond to item feedback	3.1.1.1 Incorporate received feedback (Alderson et al., 1995; Salisbury, 2005)

Appendix F. Amendments Made to the Framework following Focus Groups, Workshops with SMEs and Expert Review

Appendix F.1 Amendments to Performance Indicators Following Focus Groups

First draft performance indicators	Concerns	Changes
1.2.1 Select source texts/materials for the test	Suggestion to use the word 'search for' instead of 'select' to reflect the task more comprehensively; suggestion to use 'input materials' as it can be applied to both listening texts/scripts and reading texts	1.2.1 Search for input materials
1.2.2 Evaluate source texts to make decisions	Ambiguity 'make decisions about what?'; Suggestion to use 'input materials' consistently; Suggestion to replace 'make decisions' with 'for text-fit-specifications'	1.2.2 Evaluate the input materials for text-fit-specifications
1.2.3 Pre-edit the generated texts for item writing	'Pre-edit' may be too general; Suggest to replace with a more specific expression 'Make textual changes'	1.2.3 Make textual changes to the selected materials for tests
1.2.4 Adapt source texts	Suggestion to add the purpose for adapting the input materials	1.2.4 Adapt input materials for texts-fit-items and items-fit-texts
2.1.1 Approach item writing	Suggestion to reword for improved clarity	2.1.1 Determine the approach to item writing
2.1.5 Estimate item difficulty	Suggestion to use 'decide' rather than 'estimate' to be more relevant given that test items are qualitatively validated rather than statistically validated	2.1.5 Decide on item difficulty
2.2.1 Review test items	This is a general statement – could be interpreted in many ways; need to specify what kinds of review it is. An editorial review or item content review? or both?	2.2.2 Review item content fit

Appendix F.2 Amendments Made to the Quality Criteria Following Workshops

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
1.2.1 Search for input materials for tests		
1.2.1.1 Select text of suitable length 1.2.1.2 Select authentic texts from prestigious sources based on an inventory of criteria/characteristics and/or guidelines 1.2.1.3 Choose potential texts from published test books 1.2.1.4 Search the Internet for materials of own interest 1.2.1.5 Locate texts from previously used high-yield sources 1.2.1.6 Source texts from new resources	Suggestion to add *'select English texts from domestic sources'* Suggestion to combine *'1.2.1.1 Select text of suitable length'* and *'1.2.1.4 Search the Internet for materials of own interest'*; to change '*own interest'* to *'recommended topics'* as often provided in the test specifications	1.2.1.1 Select English texts that potentially fit the level of difficulty and complexity required from domestic (Vietnamese) newspapers, journals, magazines and books 1.2.1.2 Browse the Internet for materials of required length on recommended topics 1.2.1.3 Select authentic texts from prestigious sources based on an inventory of criteria/characteristics and/or guidelines 1.2.1.4 Choose potential texts from published test books 1.2.1.5 Locate texts from previously used high-yield sources 1.2.1.6 Source texts from new resources
1.2.2 Evaluate the input materials for text-fit-specifications		
1.2.2.1 Seek English native colleague judgement of materials	Suggestions to remove *'1.2.2.3 Seek English native colleague judgement of materials'* on	1.2.2.1 Use measures (e.g., English Profile, Academic Word List, Natural Language

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
1.2.2.2 Use measures (e.g., English Profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine source texts meet requirements for level of lexical difficulty and textual complexity 1.2.2.3 Judge the relevance of source texts based on own common sense/intuition/ "see items leap out from the text" 1.2.2.4 Assess texts based on defined characteristics of text contents and difficulty levels	the grounds that this reflects opportunities and resources available rather than item writers' expertise	Processing, British National Corpus, LexTutor) to determine source texts meet requirements for level of lexical difficulty and textual complexity 1.2.2.2 Judge the relevance of source texts based on own common sense/intuition/ "see items leap out from the text" 1.2.2.3 Assess texts based on defined characteristics of text contents and difficulty levels
1.2.3 Make textual changes to the input materials for tests		
1.2.3.1 Retain the original texts 1.2.3.2 Cut texts to required length 1.2.3.3 Make lexical and syntactic changes to achieve the desirable register (e.g., to make the text more academic or colloquial)	Suggestion to remove *'Retain the original texts'*; item writers hardly ever find an original text that matches the test specifications 100%; making changes to original texts is necessary and also for copyright reasons;	1.2.3.1 Cut texts to required length 1.2.3.2 Make lexical and syntactic changes to achieve the desirable register (e.g., to make the text more academic or colloquial) 1.2.3.3 Amend texts for improved clarity and writing quality (e.g., removing or glossing

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
1.2.3.4 Amend texts for improved clarity and writing quality (e.g., removing or glossing cultural references or parts requiring technical knowledge; deleting unsuitable offensive contents, deleting irrelevant elements of texts, cutting redundancies and repetitions)		cultural references or parts requiring technical knowledge; deleting unsuitable offensive contents, deleting irrelevant elements of texts, cutting redundancies and repetitions)
2.1.4 Provide answer keys		
2.1.4.1 Assign an answer key for each item 2.1.4.2 Provide a reference (e.g., underline chunks of texts from reading passages, listening scripts that help to justify the answer key) for each answer key	Suggestions to add *'2.1.4.3 Justify correct and incorrect answer options based on the evidence from text/script'*	2.1.4.1 Assign an answer key for each item 2.1.4.2 Provide a reference (e.g., underline chunks of texts from reading passages, listening scripts that help to justify the answer key) for each answer key 2.1.4.3 Justify correct and incorrect answer options based on the evidence from text/script
2.1.5 Decide on item difficulty		
2.1.5.1 Consider the text complexity and cognitive demand/sub-skills (e.g., reading for factual information, listening	Suggestions to remove 2.1.5.3 and 2.1.5.4 (the descriptors related to estimating item difficulty using statistical methods such as descriptive statistics and	2.1.5.1 Map each item with descriptors defined in CEFR – VN descriptors indicated in the test specifications

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
for details, and making inferences from text) required to answer each item 2.1.5.2 Judge item difficulty and discrimination based on experience and common sense as English language teacher 2.1.5.3 Measure item difficulty based on descriptive statistics of piloted data (Mean, Mode, Median, Range, Standard deviation, p-value) 2.1.5.4 Use a model –based method (e.g., classical test theory, Item response theory) and relevant item analysis programs to compute indexes regarding the distribution of item difficulties, item discrimination and distractor performance	model-based methods namely classical test theory and item response theory) on the grounds that: item piloting was not conducted due to limited resources and test security reasons; therefore, there was no data on the item responses to inform them of the item difficulty and discrimination. In Vietnam, item analysis was often conducted as a post-test procedure and item writers were not involved in this process, and did not have access to the item response data. Suggestions to add *'Map each item with descriptors defined in CEFR – VN indicated in the test specifications'*	2.1.5.2 Consider the text complexity and cognitive demand/sub-skills (e.g., reading for factual information, listening for details, and making inferences from text) required to answer each item 2.1.5.3 Judge item difficulty and discrimination based on experience and common sense as English language teacher

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
2.2.1 Check item content fit		
2.2.1.1 Edit and proof tests using an editorial style guide 2.2.1.2 Check tests for flaws in readability and mechanics 2.2.1.3. Check items using a pre-established reviewing checklist 2.2.1.4 Conduct collaborative test item reviews	Suggestions to remove 2.2.1.1, 2.2.1.2 and 2.2.1.3; these are integral tasks in the test item content review process; should not be considered a major job function; no need to put in as separate criteria) Suggestions to add *'Align item content to the test specifications'; 'Map item content with the abilities/skills required to complete the items'; 'Try out the test items on own to make judgements of item difficulty'*	2.2.1.1Align item content to the test specifications 2.2.1.2 Map item content with the abilities/skills required to complete the items 2.2.1.3 Try out the test items on own to make judgements of item difficulty
Have the items reviewed by the target examinees Collaborate with colleagues to proofread tests Engage in editorial reviews of items with at least one English native colleague for language naturalness, and authenticity this depends on resources available not person's expertise)	Suggestions to remove all the descriptors related to collaborative test item review on the grounds collaborative item review was an opportunity rather than an expertise. The SME elaborated that it was part of the examination board's strategies and policies for the coordination and monitoring of the item writing process. They argued that whether having their items reviewed collaboratively or not did	

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
Collaborate with a professional editor(s) to edit and proof items using an editorial style guide	not reflect their levels of expertise while the ability to self-review test items demonstrated their levels of expertise that grew with experience, skills and knowledge. The SMEs added that as the majority of VSTEP® item writers in Vietnam worked in isolation or even in lockdown guarded by police officers when writing items for high-stakes tests and were not involved in the post-item writing process, collaborative item reviewing tended to be unrealistic in the context of Vietnam. They also said that the framework was intended to be used for item writers' self-development of expertise, and thus it should focus on their individual expertise, removing cooperative aspects of the item writing process. In the same vein, the SMEs claimed that the same should be applied to those descriptors related to the use of English native consultants for reviewing test items and input test materials and use of statistical analysis of item responses to inform item writers of item difficulty and to verify answer keys and	
Internally review test items for obvious flaws in readability and mechanics Review and critique test items collaboratively a SME, an item-writing leader and an English native colleague for naturalness, understandability and item-fit-specifications Try out test items as embedded items in operational tests with external SMEs to provide judgement on appropriateness of content, linguistic complexity/difficulty, correctness		

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
of answer and verification of reference Co-ordinate rigorous and separate independent reviews with a panel of various expertness to collect constructive and critical feedback and provide justification on test items Pre-test drafted items in a test form with targeted representative sample of examinees to collect item responses for analysis and moderation Review items based on statistical analyses of item responses	distractors. They reiterated that these reflected opportunities rather than expertise.	
2.2.2 Verify answer keys and distractors		
2.2.3.1 Try out items on own to check the original keys 2.2.3.2 Analyse responses from a pilot with a sample of students to	Suggestions to remove *'2.2.3.3 Identify distractor discrimination through statistical analysis of responses to item distractors'* on the grounds that item response data is not	2.2.3.1 Try out items on own to check the original keys

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
identify any miskeys and/or poor performing distractors 2.2.3.3 Identify distractor discrimination through statistical analysis of responses to item distractors	available to item writers; test items are not piloted with the targeted population before being put into operational tests. The SMEs' main argument was that because the items they wrote were not piloted before being put into operational tests for reasons of test security, there were no data on the item responses to inform them of the item difficulty and discrimination. They did not perceive the need for developing their competence in estimating item difficulty based on statistical analysis. This resonated the finding from the literature that the role and the extent to which item writers are involved in the process after they submit their test items also vary (ALTE, 2005; Haladyna & Rodriguez, 2013; Lane et al., 2016); for some item writers, their job finishes when they submit the items and the rest will be taken care of by the editing team and/or the test developers. For some, they may take part in editing meetings and engaging in robust discussions on their items, and revise their	2.2.3.2 Analyse responses from a pilot with a sample of students to identify any miskeys and/or poor performing distractors

Draft performance indicators/ Quality criteria	Recommendations by SMEs	Revised quality criteria
	items based on received feedback. For others, they may even be informed of how their test items perform in the operational or pilot tests (see Section 3.3.4). They recommended removing these descriptors on the grounds that they did not reflect their actual practice and that contextual constraints should be taken into consideration.	

Appendix F.3 Amendments Made to the Wording of the Quality Criteria Following Expert Review

Performance indicators/quality criteria	Revised quality criteria
1.2.3 Make textual changes to the input materials for tests	
1.2.3.1 Cut texts to required length	1.2.3.1 Restructure and cut texts to required length
1.2.3.2 Make lexical and syntactic changes to achieve the desirable register (e.g., to make the text more academic or colloquial)	1.2.3.2 Make lexical and syntactic changes to achieve the target register
1.2.3.3 Amend texts for improved clarity and writing quality (e.g., removing or glossing cultural references or parts requiring technical knowledge; deleting unsuitable offensive contents, deleting irrelevant elements of texts, cutting redundancies and repetitions)	1.2.3.3 Employ an extensive repertoire of strategies to adapt the original text to meet the required textual features in terms of language use, style and their rhetorical functions
2.1.1 Determine the approach to item writing	
2.1.1.1 Follow standardised text-first item writing process (i.e., to generate texts first and then write items) /use the generated texts/scripts as the basis to write items in accordance with the test requirements (Text-first)	2.1.1.1 Follow a standardised text/script-first item writing process
2.1.1.2 Focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items (item first)/ and arrange texts/scripts to fit items/use item-first approach	2.1.1.2 Focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items
2.1.1.3 Multi-task between text/script adaptation and item construction/ Work with texts and items simultaneously	2.1.1.3 Multi-task between text/script adaptation and item construction

Performance indicators/quality criteria	Revised quality criteria
2.1.2 Write item stems	
2.1.2.1 Replicate items from sample tests provided in the test specifications	2.1.2.1 Replicate items from sample tests provided in the test specifications
2.1.2.2 Construct items from a pool of item shells/item stems derived from established items in past test materials	2.1.2.2 Construct items from a pool of item shells/item stems derived from established items in past test materials
2.1.2.3 Rely on different sources (e.g., test specifications, past exam papers, examples of good items, previously validated items, guidelines) to create items accommodating test specifications and input materials	2.1.2.3 Rely on different sources to create items accommodating test specifications and input materials
2.1.2.4 Build on various resources to innovate items	2.1.2.4 Build on various resources to innovate items
2.1.3 Write item distractors	
2.1.3.1 Follow the format of the distractors given in the test specifications/ Write test items with adherence to those from past papers	2.1.3.1 Follow the format of the distractors given in the test specifications or sample tests
2.1.3.2 Write distractors with reference to guidelines on Dos and Don’ts for test item writing	2.1.3.2 Write distractors with reference to guidelines on Dos and Don’ts for writing test item options
2.1.3.3 Utilise a range of paraphrasing strategies (e.g., antonyms, synonyms, change sentence structures, change word forms, change word order or use referents) to construct distractors	2.1.3.3 Utilise a range of paraphrasing strategies to construct distractors
	2.1.3.4 Use a flexible repertoire of strategies to write distractors with consideration of typical errors or misconceptions of the subject content by test takers

Performance indicators/quality criteria	Revised quality criteria
2.1.3.4 Use typical errors or misconceptions of test takers to write distractors/ Draw on subject matter knowledge and experience with target examinees to write distractors that reflect common errors or misconceptions	
2.2.5 Decide on item difficulty	
2.1.5.1 Map each item with descriptors defined in CEFR – VN descriptors indicated in the test specifications 2.1.5.2 Consider the text complexity and cognitive demand/sub-skills (e.g., reading for factual information, listening for details, and making inferences from text) required to answer each item 2.1.5.3 Judge item difficulty and discrimination based on experience and common sense as English language teacher	2.1.5.1 Align items with the descriptors defined in the item blueprint and/or in the six-level English language proficiency framework for Vietnamese learners 2.1.5.2 Consider the text complexity and cognitive demand/sub-skills required to answer each item 2.1.5.3 Draw on a range of factors affecting item difficulty from experience, expert colleagues and contemporary research to make judgements of item difficulty
2.2.1 Check item content fit	
2.2.1.1Align item content to the test specifications 2.2.1.2 Map item content with the abilities/skills required to complete the items 2.2.1.3Try out the test items on own to make judgements of item difficulty	2.2.1.1 Examine the conformity of the items to the test specifications regarding item content, wording of items and instructions 2.2.1.2 Consider abilities required to complete the items to compare what the items are thought to be testing with what is claimed in the test specifications

Performance indicators/quality criteria	Revised quality criteria
	2.2.1.3 Attempt items on own to make judgements of item-fits-specifications for content complexity and difficulty levels
2.2.2 Verify answer keys and distractors	
2.2.3.1 Try out items on own to check the original keys	2.2.3.1 Try out items on own to check the original keys
2.2.3.2 Analyse responses from a pilot with a sample of students to identify any miskeys and/or poor performing distractors	2.2.3.2 Analyse responses from a self-conducted pilot with a sample of students to identify any miskeys and/or poor performing distractors

Appendix F.4 Amendments Made to the Ordering of the Quality Criteria Following Expert Review

Performance indicators/quality criteria	Recommendations by panellists	Revised quality criteria
1.1.1 Review test specifications		
1.1.1.1 Check the test purposes, the test format, the distribution of test items and their targeted levels within the constructs	Suggestions to combine 1.1.1.1 and 1.1.1.2; they show the same level of skills required	1.1 1.1 Check aspects of the test specifications (e.g., the skills/subskills to be tested, text length, text type, item formats, quantity, and item difficulty levels)
1.1.1.2 Identify targeted test item content, and levels of item difficulty	Suggestions to reword for improved clarity	1.1.1.2 Retrieve from prior experience important aspects of the test to be written.
1.1.1.3 Recall important aspects of the test to be written/ Generate ideas about what and how to write the items for the test		1.1.1.3 Consider operational implications of the test specifications with the target examinees in mind
1.1.1.4 Use mental models of target examinees to trigger operational consideration of test specifications or Consider/contextualise operational implications of the test specifications		

Performance indicators/quality criteria	Recommendations by panellists	Revised quality criteria
1.1.2 Utilise test specifications		
1.1.2.1 Follow the instructions from the test specifications	Suggestions to combine 1.1.2.2 and 1.1.2.4 and move 1.1.2.3 to the second position	1.1.2.1 Follow the instructions from the test specifications
1.1.2.2 Recall my prior (item writing) experiences to interpret the test specifications		1.1.2.2. Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications
1.1.2.3 Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications	Suggestions to reword for improved clarity	1.1.2.3 Use experience and available resources to internalise the test specifications
1.1.2.4 Recall previous experience and discuss with colleagues to internalise content of test specifications		1.1.2.4 Develop own internal guideline for item writing in accordance with the given test specifications and informed by other training materials/resources and prior experiences
1.1.2.5 Refer to own internal guideline for item writing based on the given test specification and/or other training materials/resources and prior experiences		

Performance indicators/quality criteria	Recommendations by panellists	Revised quality criteria
1.2.1 Search for input materials for tests		
1.2.1.1 Select English texts that potentially fit the level of difficulty and complexity required from domestic (Vietnamese) newspapers, journals, magazines and books 1.2.1.2 Browse the Internet for materials of required length on recommended topics 1.2.1.3 Select authentic texts from prestigious sources based on an inventory of criteria/characteristics and/or guidelines 1.2.1.4 Choose potential texts from published test books 1.2.1.5 Locate texts from previously used high-yield sources 1.2.1.6 Source texts from new resources	Suggestions to combine 1.2.1.3, 1.2.1.4 and 1.2.1.6; they reflect the same level of skills and different ways to demonstrate the competency	1.2.1.1 Browse materials of required length on recommended topics 1.2.1.2 Identify texts of specified complexity and difficulty from newspaper, magazines and books written for non-native English readers 1.2.1.3 Locate texts from previously used high-yield sources 1.2.1.4 Source texts from a diverse range of resources with consideration of the target examinees

Performance indicators/quality criteria	Recommendations by panellists	Revised quality criteria
1.2.2 Evaluate the materials for text-fit-specifications		
1.2.2.1 Use measures (e.g., English profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine source texts meet requirements for level of lexical difficulty and textual complexity	Suggestions to re-order 1.2.2.2 and 1.2.2.3 to be in line with the Dreyfus and Dreyfus model of skill acquisition which states that experts draw on experience and intuition (Dreyfus & Dreyfus, 1986)	1.2.2.1 Use text analysis tools (e.g., English profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine text-fit-specifications in terms of difficulty and complexity 1.2.2.2 Assess texts based on defined characteristics of text contents and difficulty levels 1.2.2.3 Use multiple sources of evidence regarding text density, lexical difficulty, text potentiality for item development to judge the relevance of input materials
1.2.2.2 Judge the relevance of source texts based on own common sense/intuition/ “see items leap out from the text”	Suggestions to reword for improved clarity	
1.2.2.3 Assess texts based on defined characteristics of text contents and difficulty levels		
1.2.4 Adapt input materials for text-fit-items and items-fit-text		
1.2.4.1 Design items closely adhering to the generated text/script	Suggestions to remove *‘1.2.4.3 Add a second text to supplement text-fit-items’* as it is part of 1.2.4.4	1.2.4.1 Design items closely adhering to the generated text/script 1.2.4.2 Use research- grounded factors/ruses (e.g., use of paraphrase, antonyms and synonyms to avoid
1.2.4.2 Use research- grounded factors/ruses (e.g., use of paraphrase,		

Performance indicators/quality criteria	Recommendations by panellists	Revised quality criteria
antonyms and synonyms to avoid dictation effects, add distraction to text/scripts to reduce guessability, word-form shift, secondary text supplement) to adjust texts to fit items 1.2.4.3 Add a second text to supplement text-fit-items 1.2.4.4 Call upon a range of strategies to transform source texts to fit specified test purposes/requirements and test items/or to make the test workable		dictation effects, add distraction to text/scripts to reduce guessability, word-form shift, secondary text supplement) to adjust texts to fit items 1.2.4.3 Call upon a range of strategies to transform source texts to fit specified test purposes and test item requirements

Appendix G. The Final Draft Framework for Vietnamese Teacher Competence in Multiple-Choice Test Item Writing for English Reading and Listening Skill Tests

PHASES	COMPETENCIES	PERFORMANCE INDICATORS
THE EXPLORATORY PHASE	1.1 Analysing test specifications	1.1.1 Review test specifications
		1.1.2 Utilise test specifications
	1.2 Generating materials for tests	1.2.1 Search for input materials for tests
		1.2.2 Evaluate the input materials for text-fit-specifications
		1.2.3 Make textual changes to the selected input materials for tests
		1.2.4 Adapt input materials for text-fit-items and items-fit-texts
THE CONCERTED PHASE	2.1 Writing test items	2.1.1 Determine the approach to item writing
		2.1.2 Write item stems
		2.1.3 Write item distractors
		2.1.4 Provide answer keys
		2.1.5 Decide on item difficulty
	2.2 Reviewing test items	2.2.1 Check item content fit
		2.2.2 Verify answer keys and distractors
THE REFINING PHASE	3.1 Revising test items	3.1.1 Respond to item feedback

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
THE EXPLORATORY PHASE		
1.2 Analysing test specifications	1.1.1 Review test specifications	1.1 1.1 Check aspects of the test specifications (e.g., the skills/subskills to be tested, text length, text type, item formats, quantity, and item difficulty levels) 1.1.1.2 Retrieve from prior experience important aspects of the test to be written. 1.1.1.3 Consider operational implications of the test specifications with the target examinees in mind
	1.1.2 Utilise test specifications	1.1.2.1 Follow the instructions from the test specifications 1.1.2.2. Refer to other training materials or resources (e.g., sample tests) to clarify the test specifications 1.1.2.3 Use experience and available resources to internalise the test specifications 1.1.2.4 Develop own internal guideline for item writing in accordance with the given test specifications and informed by other training materials/resources and prior experiences
1.3 Generating materials for tests	1.2.1 Search for input materials for tests	1.2.1.1 Browse materials of required length on recommended topics 1.2.1.2 Identify texts of specified complexity and difficulty from newspaper, magazines and books written for non-native English readers 1.2.1.3 Locate texts from previously used high-yield sources 1.2.1.4 Source texts from a diverse range of resources with consideration of the target examinees
	1.2.2 Evaluate the input materials for text-fit-specifications	1.2.2.1 Use text analysis tools (e.g., English Profile, Academic Word List, Natural Language Processing, British National Corpus, LexTutor) to determine text-fit-specifications in terms of difficulty and complexity 1.2.2.2 Assess texts based on defined characteristics of text contents and difficulty levels 1.2.2.3 Use multiple sources of evidence regarding text density, lexical difficulty, text potentiality for item development (see items leap from the text) to judge the relevance of input materials

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
	1.2.3 Make textual changes to the selected input materials for tests	1.2.3.1 Restructure and cut texts to required length 1.2.3.2 Make lexical and syntactic changes to achieve the target register 1.2.3.3 Employ an extensive repertoire of strategies to adapt the original text to meet the required textual features in terms of language use, style and their rhetorical functions
	1.2.4 Adapt input materials for text-fit-items and items-fit-texts	1.2.4.1 Design items closely adhering to the generated text/script 1.2.4.2 Use research- grounded factors/ruses (e.g., use of paraphrase, antonyms and synonyms to avoid dictation effects, add distraction to text/scripts to reduce guessability, word-form shift, secondary text supplement) to adjust texts to fit items 1.2.4.3 Call upon a range of strategies to transform source texts to fit specified test purposes/requirements and test items/or to make the test workable
THE CONCERTED PHASE		
2.1. Developing test items	2.1.1 Determine the approach to item writing	2.1.1.1 Follow a standardised text/script-first item writing process 2.1.1.2 Focus on skills to be tested to write items in accordance with the test requirements and alter texts/scripts to accommodate these items 2.1.1.3 Multi-task between text/script adaptation and item construction
	2.1.2 Write item stems	2.1.2.1 Replicate items from sample tests provided in the test specifications 2.1.2.2 Construct items from a pool of item shells/item stems derived from established items in past test materials 2.1.2.3 Rely on different sources to create items accommodating test specifications and input materials 2.1.2.4 Build on various resources to innovate items
	2.1.3 Write item distractors	2.1.3.1 Follow the format of the distractors given in the test specifications or sample tests 2.1.3.2 Write distractors with reference to guidelines on Dos and Don’ts for writing test item options 2.1.3.3 Utilise a range of paraphrasing strategies to construct distractors 2.1.3.4 Use a flexible repertoire of strategies to write distractors with consideration of typical errors or misconceptions of the subject content by test takers

PHASES/ COMPETENCIES	PERFORMANCE INDICATORS	QUALITY CRITERIA
	2.1.4 Provide answer keys	2.1.4.1 Assign an answer key for each item 2.1.4.2 Provide a reference for each answer key 2.4.1.3 Justify correct and incorrect answer options based on the evidence from text/script
	2.1.5 Decide on item difficulty	2.1.5.1 Align items with the descriptors defined in the item blueprint and/or in the six-level English language proficiency framework for Vietnamese learners 2.1.5.2 Consider the text complexity and cognitive demand/sub-skills required to answer each item 2.1.5.3 Draw on a range of factors affecting item difficulty from experience, expert colleagues and contemporary research to make judgements of item difficulty
2.2 Reviewing test items	2.2.1 Check item content fit	2.2.1.1 Examine the conformity of the items to the test specifications regarding item content, wording of items and instructions 2.2.1.2 Consider abilities required to complete the items to compare what the items are thought to be testing with what is claimed in the test specifications 2.2.1.3 Attempt items on own to make judgements of item-fits-specifications for content complexity and difficulty levels
	2.2.2 Verify answer keys and distractors	2.2.3.1 Try out items on own to check the original keys 2.2.3.2 Analyse responses from a pilot with a sample of students to identify any miskeys and/or poor performing distractors
THE REFINING PHASE		
3.1 Revising test items	3.1.2 Respond to item feedback	3.1.1.1 Incorporate received feedback 3.1.1.2 Consider the test purposes to make the decision to implement feedback 3.1.1.3 Use information from experience and research to justify decision to implement or not to implement received feedback