

**Development and Validation of Classroom Assessment Literacy Scales:
English as a Foreign Language (EFL) Instructors in a
Cambodian Higher Education Setting**

Nary Tao

BEd (TEFL), IFL, Cambodia

MA (TESOL), UTS, Sydney, Australia

Submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy

College of Education

Victoria University

Melbourne, Australia

March 2014

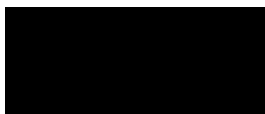
Abstract

This study employed a mixed methods approach aimed at developing and validating a set of scales to measure the classroom assessment literacy development of instructors. Four scales were developed (i.e. Classroom Assessment Knowledge, Innovative Methods, Grading Bias and Quality Procedure). The first scale was a multiple-choice test designed to measure the assessment knowledge base of English as a Foreign Language (EFL) tertiary instructors in Cambodia, whereas the latter three scales were constructed to examine their assessment-related personal beliefs (using a series of rating scale items). One hundred and eight instructors completed the classroom assessment knowledge test and the beliefs survey. Both classical and item response theory analyses indicated that each of these four scales had satisfactory measurement properties. To explore the relationship among the four measures of classroom assessment literacy, a one-factor congeneric model was tested using Confirmatory Factor Analysis (CFA). The results of the CFA indicated that a one-factor congeneric model served well as a measure of the single latent Classroom Assessment Literacy construct. In addition to the survey, in-depth, semi-structured interviews were undertaken with six of the survey participants. The departments' assessment-related policies and their learning goals documents were also analysed. The qualitative phase of the study was used to further explore the assessment related knowledge of the instructors (in terms of knowledge and understanding of the concepts of validity and reliability) as well as their notions of an ideal assessment, their perceived assessment competence, and how this related to classroom assessment literacy. Overall, the results in both phases of the study highlighted that the instructors demonstrated limited classroom assessment literacy, which had a negative impact on their actual assessment implementation. Instructors' background characteristics were found to have an impact on their classroom assessment literacy. The findings had direct implications for assessment policy development in tertiary education settings as well as curriculum development for pre- and in-service teacher education programmes within developing countries.

Declaration

I, Nary Tao, declare that the PhD thesis entitled “Development and Validation of Classroom Assessment Literacy Scales: English as a Foreign Language (EFL) Instructors in a Cambodian Higher Education Setting” is no more than 100,000 words in length, including quotes and exclusive of tables, figures, appendices and references. This thesis contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work.

Signature

A solid black rectangular box used to redact the signature of Nary Tao.

Nary Tao

Date

17 March 2014

Dedication

This study is dedicated to my dad, Sovann Tao, who encouraged me to reach the highest level of education possible throughout my life, and my mum, Chou Pring, who has been very supportive, particularly during this PhD journey, for which I am greatly indebted.

Acknowledgements

My PhD study is a long journey and has presented me with various challenges from the beginning to its completion. I am indebted to a number of people who have provided me with guidance, support and encouragement throughout this journey.

I am especially grateful to my supervisors, Associate Professor Shelley Gillis, Professor Margaret Wu and Dr. Anthony Watt, for their talent and expertise in guiding and keeping me on target, providing me with ongoing constructive feedback needed to improve each draft chapter of my thesis, as well as challenging me to step outside of my comfort zone. Throughout the period of their supervision, I have gained enormously from their knowledge, skills and encouragement, particularly from the freedom of pace and thoughts they permit. Such expert supervision has played a critical role in the completion of this study.

I owe special thanks to Dr. Cuc Nguyen, Mr. Mark Dulhunty, Dr. Say Sok, Ms. Sumana Bounchan, Mr. Chivoin Peou and Mr. Soth Sok for their valuable feedback with regard to the items employed in the Classroom Assessment Knowledge test, questionnaire and semi-structured interviews during the pilot stage.

I thank the Australian government (through AusAID) for providing the generous financial support throughout this doctoral study, as well as for my previous completed master's study at the University of Technology, Sydney (UTS) during the 2005-2006 academic year.

I wish to extend my sincere thanks to the participating instructors from the two recruited English departments within one Cambodian city-based university. Without their voluntary and enthusiastic participation, this study would not have been possible.

I express my deep appreciation to my family for their love, patience, understanding and support, for which I am grateful.

Table of Contents

Abstract	i
Declaration	ii
Dedication	iii
Acknowledgements	iv
Table of Contents	v
List of Figures	x
List of Tables	xii
List of Abbreviations	xiv
Chapter 1: Introduction	1
1.1 Rationale.....	1
1.2 The Demand for English Language in Cambodia: An Overview	11
1.3 English Language Taught in Cambodian Schools and University	12
1.4 Purpose of the Study	13
1.4.1 Research Questions	14
1.5 Significance of the Study	14
1.6 Structure of the Thesis	15
Chapter 2: Classroom Assessment Processes	17
2.1 Classroom Assessment.....	17
2.2 Classroom Assessment Processes	19
2.2.1 Validity	20
2.2.2 Reliability.....	23
2.2.3 Assessment Purposes.....	25
2.2.4 Assessment Methods	31
2.2.5 Interpreting Assessment Outcomes	43
2.2.6 Grading Decision-making.....	48
2.2.7 Assessment Records	50
2.2.8 Assessment Reporting	51
2.2.9 Assessment Quality Management	56

2.3	Summary	59
Chapter 3: Classroom Assessment Literacy		62
3.1	Theoretical Framework	62
3.2	Concepts of Literacy	64
3.2.1	Definitions of Assessment Literacy	65
3.3	Research on Assessment Literacy.....	67
3.3.1	Assessment Knowledge Base	67
3.3.1.1	Self-reported Measures	68
3.3.1.2	Objective Measures.....	74
3.3.2	Assessment Beliefs.....	81
3.3.2.1	Stages of the Assessment Process: Teachers' Beliefs.....	82
3.3.3	Relationship between Assessment Knowledge and Assessment Practice	85
3.3.4	Relationship between Assessment Belief and Assessment Practice	85
3.4	Summary	88
Chapter 4: Background Characteristics Influencing Classroom Assessment Literacy		89
4.1	Background Characteristics Influencing Classroom Assessment Literacy.....	89
4.1.1	Pre-service Assessment Training	89
4.1.2	Teaching Experience	91
4.1.3	Academic Qualification	92
4.1.4	Gender	92
4.1.5	Professional Development	93
4.1.6	Class Size	93
4.1.7	Teaching Hours	94
4.1.8	Assessment Experience as Students	94
4.2	Summary	95
Chapter 5: Methodology		96
5.1	Part One: Mixed Methods Approach	96
5.1.1	Rationale and Key Characteristics of the Mixed Methods Approach	96
5.1.2	Mixed Methods Sequential Explanatory Design.....	98
5.1.3	Advantages and Challenges of the Sequential Explanatory Design	100
5.2	Part Two: Quantitative Phase	100

5.2.1	The Target Sample	100
5.2.1.1	The Sampling Framework	101
5.2.2	Data Collection Procedures	102
5.2.2.1	Response Rate.....	102
5.2.2.2	Test and Questionnaire Administration	103
5.2.3	Test and Questionnaire Development Processes.....	103
5.2.3.1	The Measures.....	105
5.2.4	Quantitative Data Analysis	111
5.2.4.1	Item Response Modelling Procedure	111
5.2.4.2	Structural Equation Modelling Procedure.....	113
5.3	Part Three: Qualitative Phase	121
5.3.1	The Sample	121
5.3.2	Data Collection Procedures	122
5.3.2.1	Departmental Learning Goals and Assessment-related Policies	122
5.3.2.2	Interview Administration	122
5.3.3	Interview Questions Development Processes	123
5.3.3.1	Interview Questions	123
5.3.4	Qualitative Data Analysis	124
Chapter 6: Scale Development Processes		129
6.1	Development of the Scales	129
6.1.1	Development of the Classroom Assessment Knowledge Scale.....	129
6.1.2	Development of the Innovative Methods scale.....	137
6.1.3	Development of the Grading Bias Scale.....	141
6.1.4	Development of the Quality Procedure Scale	144
6.2	Summary Statistics.....	149
Chapter 7: Quantitative Results		151
7.1	Univariate Results.....	151
7.1.1	The Sample	151
7.1.2	Tests of Normality.....	153
7.2	Bivariate Results	155
7.2.1	Interrelationships among the Classroom Assessment Literacy Constructs	155

7.2.2	Classroom Assessment Literacy Variables as a Function of Age	157
7.2.3	Classroom Assessment Literacy Variables as a Function of Teaching Experience	161
7.2.4	Classroom Assessment Literacy Variables as a Function of Teaching Hours	163
7.2.5	Classroom Assessment Literacy Variables as a Function of Class Size	167
7.2.6	Classroom Assessment Literacy Variables as a Function of Gender	170
7.2.7	Classroom Assessment Literacy Variables as a Function of Departmental Status ..	172
7.2.8	Classroom Assessment Literacy Variables as a Function of Academic Qualifications	176
7.2.9	Classroom Assessment Literacy Variables as a Function of Pre-service Assessment Training.....	178
7.3	Multivariate Results	184
7.3.1	Congeneric Measurement Model Development	184
7.3.1.1	One-factor Congeneric Model: Classroom Assessment Literacy.....	184
Chapter 8: Qualitative Results		188
8.1	Learning Goals of University Departments.....	188
8.2	Departmental Assessment-related Policies	189
8.3	Background Characteristics of the Interviewees	193
8.4	Classroom Assessment Literacy	194
8.4.1	Perceived Assessment Competence	195
8.4.2	Notion of the Ideal Assessment	201
8.4.3	Knowledge and Understanding of the Concepts of Validity and Reliability	206
8.5	Summary	219
Chapter 9: Discussion and Conclusion.....		221
9.1	Overview of the Study	221
9.1.1	Review of Rationale of the Study	221
9.1.2	Review of Methodology	224
9.1.2.1	Quantitative Phase	224
9.1.2.2	Qualitative Phase	226
9.2	Discussion	228

9.2.1	Main Research Question: To what extent did assessment related knowledge and beliefs underpin classroom assessment literacy and to what extent could each of these constructs be measured?	228
9.2.2	Subsidiary Research Question 1: To what extent was classroom assessment literacy developmental?	229
9.2.3	Subsidiary Research Question 2: What impact did classroom assessment literacy have on assessment practices?	231
9.2.4	Subsidiary Research Question 3: How did the background characteristics of instructors (i.e., age, gender, academic qualification, teaching experience, teaching hours, class size, assessment training, and departmental status) influence their classroom assessment literacy?	234
9.2.4.1	The Influence of Pre-service Assessment Training	234
9.2.4.2	The Influence of Class Size	235
9.2.4.3	The Influence of Teaching Hours	235
9.2.4.4	The Influence of Departmental Status	236
9.2.4.5	The Influence of Age	237
9.2.4.6	The Influence of Teaching Experience	237
9.2.4.7	The Influence of Gender	238
9.2.4.8	The Influence of Academic Qualification	238
9.2.4.9	The Influence of Professional Development Workshop and Assessment Experience as Students.....	239
9.3	Conclusion.....	239
9.3.1	Implications of the Study Findings	240
9.3.1.1	Implications for Theory.....	240
9.3.1.2	Implications for Policy and Practice	241
9.3.1.3	Implications for the Design of Pre-service Teacher Education Programme	244
9.3.2	Limitations of the Study	248
9.3.3	Future Research Directions.....	249
	References.....	252
	Appendices	292

List of Figures

Figure 2.1 Classroom assessment processes	20
Figure 5.1 Diagram for the mixed methods sequential explanatory design procedures	99
Figure 5.2 Items within the IM scale	109
Figure 5.3 Items within the GB scale.....	110
Figure 5.4 Items within the QP scale	110
Figure 5.5 One-factor congeneric measurement model: Classroom Assessment Literacy	116
Figure 5.6 Interview questions	124
Figure 6.1 Nine standards and associated items within the Classroom Assessment Knowledge scale.....	130
Figure 6.2 Detail of three item analyses.....	131
Figure 6.3 Items 7 & 8	132
Figure 6.4 Variable Map of the CAK scale	135
Figure 6.5 Variable Map of the IM scale	139
Figure 6.6 Variable Map of the GB scale.....	143
Figure 6.7 Variable Map of the QP scale	147
Figure 7.1 Recoded instructor age variable across the band level of the CAK, GB, IM, and QP scales	159
Figure 7.2 Recoded instructor teaching experience variable across the band level of the CAK, GB, IM, and QP scales	162
Figure 7.3 Recoded instructor teaching hour variable across the band level of the CAK, GB, IM, and QP scales.....	165
Figure 7.4 Recoded instructor class size variable across the band level of the CAK, GB, IM, and QP scales	168
Figure 7.5 Recoded instructor gender variable across the band level of the CAK, GB, IM, and QP scales	171
Figure 7.6 Recoded instructor department variable across the band level of the CAK, GB, IM, and QP scales.....	174

Figure 7.7 Recoded instructor academic qualification variable across the band level of the CAK, GB, IM, and QP scales	177
Figure 7.8 Recoded instructor assessment training variable across the band level of the CAK, GB, IM, and QP scales	182
Figure 7.9 One-factor congeneric model: Classroom Assessment Literacy	187
Figure 8.1 The relationship between instructors' classroom assessment literacy, their backgrounds and departmental assessment policies	220

List of Tables

Table 2.1 Main Types of Assessment Purposes	27
Table 2.2 Main Types of Assessment Methods.....	32
Table 3.1 Summary of Studies Examining Teacher Assessment Competence Using Self-reported Measures.....	71
Table 3.2 Summary of Studies that Used Assessment Knowledge Tests to Measure Teacher Assessment Knowledge Base	75
Table 5.1 Instructor Background Information	106
Table 5.2 Nine Standards and Associated Items within the Classroom Assessment Knowledge Scale	108
Table 5.3 Goodness of Fit Criteria and Acceptable Level and Interpretation.....	119
Table 5.4 Table of Specifications for Selecting Six Participants	121
Table 6.1 Calibration Estimates for the Classroom Assessment Knowledge Scale	133
Table 6.2 Interpretation of the Instructor Classroom Assessment Knowledge Levels from Analyses of the CAK Scale	136
Table 6.3 Calibration Estimates for the Innovative Methods Scale.....	138
Table 6.4 Interpretation of the Instructor Innovative Methods Levels from Analyses of the IM Scale	140
Table 6.5 Calibration Estimates for the Grading Bias Scale	141
Table 6.6 Interpretation of the Instructor Grading Bias Levels from Analyses of the GB Scale	144
Table 6.7 Calibration Estimates for the Quality Procedure Scale	145
Table 6.8 Interpretation of the Instructor Quality Procedure Levels from Analyses of the QP Scale	148
Table 6.9 Summary Estimates of the Classical and Rasch Analyses for each Scale	149
Table 7.1 Background Characteristics of the Sample	152
Table 7.2 Mean, Standard Deviation, Skewness and Kurtosis Estimates	154
Table 7.3 Pearson Product-Moment Correlations for the Relationships among the Classroom Assessment Literacy Constructs, Age, Teaching Experience, Teaching Hours, and Class Size	155

Table 7.4 Classroom Assessment Literacy Variables as a Function of Gender	170
Table 7.5 Classroom Assessment Literacy Variables as a Function of Departmental Status	172
Table 7.6 Classroom Assessment Literacy Variables as a Function of Academic Qualifications	176
Table 7.7 Classroom Assessment Literacy Variables as a Function of Pre-service Assessment Training	179
Table 7.8 Classroom Assessment Literacy Variables as a Function of Assessment Training Duration.....	180
Table 7.9 Classroom Assessment Literacy Variables as a Function of the Level of Preparedness of Assessment Training	180
Table 7.10 Maximum-likelihood (ML) Estimates for One-factor Congeneric Model: Classroom Assessment Literacy	185
Table 7.11 Goodness of Fit Measures for One-factor Congeneric Model: Classroom Assessment Literacy	186
Table 8.1 Assessment Policies of the English-major and English Non-major Departments	190
Table 8.2 Background Characteristics of the Interviewees	194
Table 8.3 Self-reported Measure of Instructor Classroom Assessment Competence.....	195
Table 8.4 Validation of the Self-reported Measure of Instructor Classroom Assessment Competence	200

List of Abbreviations

AFT= American Federation of Teachers
ASEAN= Association of Southeast Asian Nations
CAMSET= Cambodian Secondary English Language Teaching
CFA= Confirmatory Factor Analysis
EFL= English as a Foreign Language
ELT= English Language Teaching
ESL= English as a Second Language
ML= Maximum Likelihood
MoEYS= Ministry of Education, Youth and Sport
NCME= National Council on Measurement in Education
NEA= National Education Association
PCM= Partial Credit Model
QSA= Quaker Service Australia
SEM= Structural Equation Modelling
TEFL= Teaching English as a Foreign Language
TESOL= Teaching English to Speakers of Other Languages
UNTAC= United Nations Transitional Authority in Cambodia

Chapter 1: Introduction

1.1 Rationale

In educational settings around the world, school and tertiary teachers are typically required to design and/or select assessment methods, administer assessment tasks, provide feedback, determine grades, record assessment information and report students' achievements to the key assessment stakeholders including students, parents, administrators, potential employers and/or teachers themselves (Taylor & Nolen, 2008; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; McMillan, 2014; Popham, 2014). Research has shown that teachers typically spend a minimum of one-third of their instructional time on assessment-related activities (Stiggins, 1991b; Quitter, 1999; Mertler, 2003; Bachman, 2014). As such, the quality of instruction and student learning appears to be directly linked to the quality of assessments used in classrooms (Earl, 2013; Heritage, 2013b; Green, 2014). Teachers therefore are expected to be able to integrate their assessments with their instruction and students' learning (Shepard, 2008; Griffin, Care, & McGaw, 2012; Earl, 2013; Heritage, 2013b; Popham, 2014) in order to meet the needs of the twenty-first century goals such as preparing students for lifelong learning skills (Binkley, Erstad, Herman, Raizen, Ripley, Miller-Ricci, & Rumble, 2012). That is, they are expected to be able to assess students' learning in a way that is consistent with twenty-first century skills comprising creativity, critical thinking, problem-solving, decision-making, flexibility, initiative, appreciation for diversity, communication, collaboration and responsibility (Binkley et al., 2012; Pellegrino & Hilton, 2012). They are also expected to design assessment tasks to assess students' broader knowledge and life skills (Masters, 2013a) by means of shifting from a testing culture to an assessment culture. A testing culture is associated with employing tests/exams merely to determine achievements/grades whereas an assessment culture is related to using assessments to enhance instruction and promote student learning (Wolf, Bixby, Glenn, & Gardner, 1991; Inbar-Lourie, 2008b; Shepard, 2013).

In other words, there has been an international educational shift in the field of measurement and assessment where teachers need to view assessments as intertwined relationships with their instruction and students' learning. That is, they have to be able to use assessment data to improve instruction and promote students' learning (Shepard, 2008; Mathew & Poehner, 2014; Popham, 2014) in terms of establishing where the students are in learning at the time of assessment (Griffin, 2009; Forster & Masters, 2010; Heritage, 2013b; Masters, 2013a).

To meet the goals of educational reform and the twenty-first century skills in relation to developing students' broader knowledge and skills, a number of assessment specialists have argued that teachers need to be able to employ a variety of assessment methods in assessing students' learning, irrespective of whether the assessment has been conducted for formative (i.e., enhancing instruction and learning) and/or summative purposes (i.e., summing up achievement) (Scriven, 1967; Bloom, Hastings, & Madaus, 1971; Wiliam, 1998a, 1998b; Shute 2008; Griffin et al., 2012; Heritage, 2013b; Masters, 2013a). These methods include performance-based tasks, portfolios and self- and peer assessments rather than exclusively using traditional assessment (e.g., tests/exams). Such assessment methods have been argued to have the potential to promote students' lifelong learning through the assessment of higher-order thinking skills (Leighton, 2011; Moss & Brookhart, 2012; Darling-Hammond & Adamson, 2013), motivate students to learn, engage them in the assessment processes, help them to become autonomous learners and foster their feelings of ownership for learning (Boud, 1990; Falchikov & Boud, 2008; Lamprianou & Athanasou, 2009; Heritage, 2013b; Nicol, 2013; Taras, 2013; Molloy & Boud, 2014).

Despite such perceived benefits, there has been continual reportings of teachers conducting assessments for summative purposes using poorly constructed, objective paper and pencil tests (e.g., multiple-choice tests) that simply measure students' low-level knowledge and skills (Oescher & Kirby, 1990; Marso & Pigge, 1993; Bol & Strage, 1996; Greenstein, 2004). It has been well documented that such poorly designed tests can lead to surface learning, and therefore produce a mismatch between classroom assessment practices and teaching/learning goals (Rea-Dickins, 2007; Binkley et al., 2012; Griffin et al., 2012; Heritage, 2013b).

There have been increasing concerns amongst educational researchers and assessment specialists regarding the impact of teachers' classroom assessment methods on students' motivation and approaches to learning. According to Crooks (1988), Harlen and Crick (2003) and Brookhart (2013a), classroom assessment can have an impact on students in various ways such as guiding their judgement of what is vital to learn, affecting motivation and self-perceptions of competence, structuring their approach to and timing of personal study, consolidating learning, and affecting the development of enduring learning strategies and skills.

Numerous researchers reported that the assessment methods used included objective exams (i.e., the testing questions that are associated with only right or wrong answers like true/false items), subjective exams (i.e., the testing questions that require students to generate written responses like essays) and assignments influencing students' approaches to learning, namely surface versus deep approaches (Entwistle & Entwistle, 1992; Tang, 1994; Marton & Säljö, 1997; Dahlgren, Fejes, Abrandt-Dahlgren, & Trowald, 2009). A surface learning approach refers to students' focusing on facts and details of their course materials when preparing for assessments, whilst a deep learning approach tends to describe preparation activities in which students develop a deeper understanding of the subject matter by integrating and relating the learning materials critically (Entwistle & Entwistle, 1992; Marton & Säljö, 1997; Biggs, 2012; Entwistle, 2012). Additional support can be drawn from Thomas and Bain (1984) and Nolen and Haladyna (1990) who reported that students employed a surface learning approach when anticipating objective tests/exams (e.g., true/false and multiple-choice tests), whereas they used a deep learning approach when expecting subjective tests/exams (e.g., paragraphs/essays) or assignments.

There has also been an anecdotal commentary amongst Western educators that Asian students are merely rote-learners (Biggs, 1998; Leung, 2002; Saravanamuthu, 2008; Tran, 2013a). In other words, Asian students tend to employ surface learning approaches in undertaking the assessment tasks. Such perceptions may be due to many Asian countries' cultures, which share the deeply rooted Confucian heritage, putting greater values on objective paper and pencil tests/exams in assessing students' factual knowledge within teaching, learning and assessment contexts.

Such perceptions have raised further concerns about the assessment of students' learning within developing countries, as these countries tend to have a strong preference for objective paper and pencil tests and norm-referenced testing (i.e., comparison of a student's performance to that of other students within or across classes), despite a worldwide shift to the use of innovative assessment and criterion-referenced framework (Heyneman, 1987; Heyneman & Ransom, 1990; Greaney & Kellaghan, 1995; Tao, 2012; Tran, 2013b). Innovative assessments tend to include performance-based assessments (i.e., which require the students to construct their own response to the assessment task/item) as well as self- and peer assessments which tend to operate within a criterion-referenced framework (i.e., a student's performance that demonstrates his/her specific knowledge and skills against the course learning goals).

Such a shift has also pushed developing countries, including Southeast Asian countries such as Cambodia, Lao and Vietnam, to reform their educational systems in relation to teaching, learning and assessment (particularly within higher education sectors) to meet the needs of the workforce regarding twenty-first century skills (Chapman, 2009; Hirosato & Kitamura, 2009). It has also been argued that higher education institutions have a critical role in providing students with the necessary knowledge and skills needed in the twenty-first century to enable them to meet global world challenges (Chapman, 2009; Hirosato & Kitamura, 2009).

Unfortunately, recent research undertaken in these developing countries, particularly within Cambodian and Vietnamese higher education settings, have shown that graduates are not prepared for developing high independent learning skills, knowledge and attributes needed in the workforce in the twenty-first century, as the assessments employed in their higher education institutions tend to strongly emphasise tests/exams to recall factual information (Rath, 2010; Tran, 2013b). For example, Rath (2010) reported that the students' learning assessments in one Cambodian city-based university were strongly focused on facts and details (i.e., rote-learning) thought to be associated with the limited critical thinking capacities of its student cohort. Similarly, Tran (2013b) found that students who were in their final year of study within the Vietnamese higher education setting reported that their universities failed to equip them with skills needed for the workplace. Students attributed such a lack of skills to their

universities' exam-oriented context, in which the exams were designed for recalling factual information. This led them to memorise factual knowledge for the sake of passing their exams.

A worldwide shift towards the use of innovative assessment, such as performance-based and criterion-referenced assessments, has also presented some challenges for teachers. Although teachers are expected to be consistent when judging students' work (in terms of reliability), it has been widely acknowledged that teachers' assessment of students' work tend to be influenced by other factors that do not necessarily reflect students' learning achievements, even though they have employed explicit marking criteria and standards (Bloxham & Boyd, 2007; Orrell, 2008; Price, Carroll, O'Donovan, & Rust, 2011; Bloxham, 2013; Popham, 2014). These extraneous factors tend to be associated with teachers' tacit knowledge (i.e., their values and beliefs) (Sadler, 2005; Orrell, 2008; Price et al., 2011; Bloxham, 2013). While teachers are expected to positively endorse innovative assessment methods in their assessment practices and judgment, research has shown that teachers demonstrate a strong preference toward the use of traditional assessment methods (i.e., objective tests/exams) rather than innovative assessment methods (Tsgari, 2008; Xu & Lix, 2009), given the latter tends to be plagued with reliability issues (Pond, Ul-Haq, & Wade, 1995; Falchikov, 2004) and heavy workload associated with marking students' work (Sadler & Good, 2006).

In addition to the worldwide shift towards embracing innovative assessments in the classroom, teachers are also expected to have positive endorsements toward employing quality assessment procedures (i.e., quality assurance and/or moderation meetings) in their assessment practices in order to guard against any extraneous factors that can have a potential impact on the accuracy and consistency of assessment results (Maxwell, 2006; Daugherty, 2010). Research, however, has highlighted a tendency for teachers to ignore quality assurance in their assessment practices, particularly those associated with the use of traditional assessment, resulting in poorly developed tests/exams (Oescher & Kirby, 1990; Mertler, 2000). Research has also demonstrated that internal moderation practices (i.e., the process undertaken by the teachers regarding their judgements of students' work to ensure valid, reliable, fair and transparent assessment outcomes) of the teachers tend to be ineffective (Klenowski & Adie, 2009; Bloxham &

Boyd, 2012). As such, it is necessary for teachers to explicitly examine their espoused personal beliefs about assessment.

Fundamentally, teachers need to be classroom assessment-literate in order to implement high quality assessments in assessing students' broader knowledge and skills needed in the twenty-first century workforce. To become classroom assessment-literate, teachers need to possess a sound knowledge base of the assessment process (Price, Rust, O'Donovan, Handley, & Bryant, 2012). For example, they have to be able to identify assessment purposes, select/design assessment methods, interpret the assessment data, make grading decision, and record and report the outcomes of assessment. Furthermore, teachers need to better understand what factors can have a potential impact on the accuracy and consistency of assessment results, as well as demonstrate capabilities to ensure the quality of assessments (Stiggins, 2010; Popham, 2014). Such knowledge and understanding will lead teachers to form holistic viewpoints regarding the interconnectedness of all stages within the entire classroom assessment process. Acquiring greater knowledge and understanding of such a process will also enable teachers to better design a variety of assessment methods to enhance instruction and promote students' learning (i.e., formative purposes) and summarise students' learning achievements (i.e., summative purposes). Becoming assessment-literate requires teachers to not only possess a sound knowledge base of the assessment process, but also to be able to explicitly examine the tensions around their implicit personal beliefs about assessment.

Research, unfortunately, has consistently shown that teachers have a limited assessment knowledge base that can impact their assessment implementation (Mayo, 1967; Plake, 1993; Davidheiser, 2013; Gotch & French, 2013). Equally, a collection of studies have repeatedly highlighted that teachers' implicit personal beliefs about assessment play a critical role in influencing the ways in which they implement their assessments (Rogers, Cheng, & Hu, 2007; Xu & Liu, 2009; Brown, Lake, & Matters, 2011). It could therefore be argued that their assessment beliefs are equally paramount to their assessment knowledge base in implementing high quality assessments; and as such the two are interwoven (Fives & Buehl, 2012) and form the underpinnings of classroom assessment literacy.

Given that there is an internationally increasing recognition of the crucial role of assessment literacy, educational researchers and assessment specialists alike have continuously called for teachers to be assessment-literate (Masters, 2013a; Popham, 2014). A solid understanding of the nature of teachers' classroom assessment literacy is important as they are the key agents in implementing the assessment process (Klenowski, 2013a). As such, their classroom assessment literacy is directly related to the quality of the assessments employed in assessing students' learning (Berger, 2012; Campbell, 2013; Popham, 2014).

In line with trends in international classroom assessment literacy research, recent concerns have been raised in relation to the quality of classroom assessment employed in EFL programmes within Cambodia's higher education sector (Bounchan, 2012; Haing, 2012; Tao, 2012; Heng, 2013b) and the classroom assessment literacy of EFL university teachers, given students' learning are mainly assessed by their teachers' developed assessment tasks (Tao, 2012). These concerns are in line with the top priority goals of the Royal Government of Cambodia regarding: the quality of higher education expected to be integrated into the ASEAN community by 2015 (ASEAN Secretariat, 2009); the goals of the Cambodian Ministry of Education with respect to the quality of teaching and learning stated in its Educational Strategic Plan 2009-2013 (MoEYS, 2010); and the vision for Cambodian higher education 2030 (MoEYS, 2012). Linked with both the 2030 Cambodian higher education vision goals and ASEAN's strategic objectives in advancing and prioritising education for its regional community in 2015, is the need to prepare students for lifelong learning or higher-order thinking skills in order to meet global world challenges. To achieve this crucial goal, teacher preparation programmes have been considered as a national priority by the Royal Government of Cambodia and are significantly supported by funding from international organisations (Duggan, 1996, 1997; MoEYS, 2010) due to the premise that high quality teacher preparation programmes will lead to high quality teaching and learning (Darling-Hammond, 2006; Darling-Hammond & Lieberman, 2012).

Despite the persistent efforts by the Royal Government of Cambodia and international organisations in improving the quality of teacher training, recent studies undertaken within a Cambodian EFL higher education context (Bounchan, 2012; Haing,

2012; Tao, 2012; Heng, 2013b) have shown that students' learning is mainly assessed on low-level thinking skills, such as facts and details, rather than on higher-order thinking skills. Such studies have also demonstrated that students tend to be assessed predominantly through employing final examinations solely for summative purposes. For example, Bounchan (2012) reported that there was no relationship between Cambodian EFL first-year students' metacognitive beliefs (i.e., the students' abilities to reflect on their own learning and make adjustments accordingly) and their grade point average (GPA). The researcher concluded that this result was not surprising, given that student learning was mainly assessed on facts and details (i.e., rote-learning or memorisation). Heng (2013a) found that Cambodian EFL first-year students' time spent on out-of-class course related tasks (e.g., reading course-related materials at home), homework/tasks and active participation in classroom settings significantly contributed to their academic learning achievements. In contrast, the time students spent on out-of-class peer learning (e.g., discussing ideas from readings with other classmates) and extensive reading (e.g., reading books, articles, magazines and/or newspapers in English) was found to have no impact on their academic learning achievements. These results were consistent with Heng's (2013b) subsequent study conducted with Cambodian EFL second-year students. The researcher therefore concluded that such findings were not uncommon, given the predominantly exam-oriented emphasis in Cambodian higher education institutions. Haing (2012) further found that Cambodian EFL tertiary teachers' predominant use of final examinations and a lack of assessment tasks throughout the course period contributed to the low quality of students' learning. Similar to Haing (2012), Tao (2012) reported that Cambodian EFL tertiary teachers in one city-based university mainly employed tests and exams to assess students' learning, as well as incorporated students' attendance and class participation into their course grades. Furthermore, the teachers' self-reported that their assessment purposes had predominantly formative functions, yet Tao (2012) argued that the assessments employed served largely summative functions. The grades obtained from such assessments were primarily used to pass or fail students in their courses. The researcher concluded that such assessment practices could be interpreted as limited classroom assessment literacy on the part of teachers. That is, because of their limited classroom assessment literacy, these teachers were unable to

distinguish the differences between formative and summative purposes for their assessments. Furthermore, they strongly relied on using tests and exams in assessing students' learning and incorporated students' non-academic achievement factors (e.g., attendance) into their course grades. Such poor assessment implementation can inflate students' actual academic achievements. The researcher then called for studies on classroom assessment literacy to be conducted within EFL programmes in a Cambodian higher education setting in order to shed light on the nature of teachers' classroom assessment literacy.

There have been increasing calls amongst educational researchers worldwide for EFL/ESL teachers to become classroom assessment-literate within the language education field (Davies, 2008; Inbar-Lourie, 2008a; Fulcher, 2012; Malone, 2013; Scarino, 2013; Green, 2014; Leung, 2014). Yet, while it is apparent that a large number of studies undertaken to measure either teachers' classroom assessment knowledge base or their personal beliefs about assessment within the general education field, there is a paucity of this kind of research conducted within the EFL/ESL context, particularly at the tertiary level. Thus, there is a need for further research focusing on classroom assessment literacy of EFL/ESL tertiary teachers in terms of their assessment knowledge base and personal beliefs about assessment. This type of study should provide a better understanding of the nature of the classroom assessment literacy construct.

Because there is an increasing recognition of the critical role for EFL programmes in both Cambodian schools and higher education sectors, the introduction of the Cambodian annual conference on English Language Teaching (ELT) titled "CamTESOL" was initiated in 2005 by IDP Education, Cambodia. This conference, held in late February, aims to: (1) provide a forum for the exchange of ideas and dissemination of information on good practice; (2) strengthen and broaden the network of teachers and all those involved in the ELT sector in Cambodia; (3) increase the links between the ELT community in Cambodia and the international ELT community; and (4) showcase research in the field of ELT (Tao, 2007, p. iii). Despite this initiative, there is still little research conducted within both Cambodian EFL schools and higher education settings. Of the limited research conducted, it has predominantly focused on issues surrounding the development of English language teaching policies and/or status (Neau, 2003;

Clayton, 2006; Clayton, 2008; Moore & Bounchan, 2010), learning and/or teaching strategies (Bounchan, 2013; Heng, 2013a) and classroom assessment practices (Tao, 2012). There is an apparent lack of research examining the classroom assessment literacy of Cambodian EFL tertiary teachers. The lack of research in this area is a concern, given that other aligned studies provide sufficient evidence of the direct relationships between the quality of classroom assessments used and the quality of instruction and student learning (Black & Wiliam, 1998a; Shute, 2008; Stiggins, 2008; Wiliam, 2011).

There are numerous reasons given as to why it is important to examine the classroom assessment literacy development of university teachers within EFL programmes in a higher education setting, as these programmes play a critical role in the Cambodian tertiary educational system. Students' enrolment in such programmes is expected to significantly increase (The Department of Cambodian Higher Education, 2009). Bounchan (2013) has recently asserted that it is not uncommon to find Cambodian undergraduate students who have enrolled for two university degrees simultaneously: typically a Bachelor of Education in Teaching English as a Foreign Language (TEFL) degree or a Bachelor of Arts in English for Work Skills (EWS) degree. It is further anticipated that EFL programmes in Cambodian higher education institutions are continuously growing, given that Cambodia is expected to be integrated into the ASEAN community by 2015 (ASEAN Secretariat, 2009). As such, the use of English language has been suggested to have a direct relationship with students' long-term academic and occupational needs: locally, regionally and internationally (Ahrens & McNamara, 2013; Bounchan, 2013). Ahrens and McNamara (2013), who have been the advocates of Cambodian higher education reforms for over a decade, have convincingly argued that "English [language] must be taught and taught extensively and well if Cambodia does not want its students to fall behind those of those of [sic] the Association of South-East Asian Nations (ASEAN) regional partners" (p. 56). These advocates have also recommended employing English language as the medium of instruction, particularly in years three and four within undergraduate programmes in all Cambodian higher education institutions and they argue that such instruction will enhance students' learning (i.e., through having access to a variety of academic materials) as well as improve the opportunities for students' future employment when they graduate. Thus, English language is seen as the

most important medium of communication in Cambodian society. Many teachers and students perceive that English could be considered as a second language in Cambodia (Moore & Bounchan, 2010). Due to its vital role, English language is therefore taught in all Cambodian schools, as well as in most higher education institutions. To give a sense of how the use of English language continues to grow in Cambodia, the following sections (see 1.2 and 1.3) will provide an overview of the demands of English language in Cambodian society, and a snapshot of English language taught in schools and university settings.

1.2 The Demand for English Language in Cambodia: An Overview

The introduction of the English language in Cambodia's workforce can be traced back to three major developments. The first development was the arrival of a range of international agencies in Cambodia. In the late 1980s when the Cambodian government moved towards democracy and opened its doors to the free market, numerous international agencies arrived in Cambodia to provide aid to assist Cambodia to integrate its economic and political transition. As the majority of these international agencies employed English as a main medium for communication, there was the need for Cambodian people to possess sufficient levels of English language proficiency to actively and fully engage with the donors' aid related activities (Clayton, 2006). The second phase was the establishment of the United Nations Transitional Authority in Cambodia (UNTAC). When the Cambodian government signed the Paris Peace Accord in 1991, UNTAC was formed to ensure future stability in facilitating Cambodia for its upcoming 1993 election. The UNTAC comprised 20, 000 personnel spread across Cambodia when they arrived in 1992. As most of the UNTAC personnel used English as their main medium for communication, there was an increased demand for Cambodian people to acquire an adequate level of English language proficiency (Neau, 2003; Clayton, 2008; Howes & Ford, 2011). The last phase was the integration of Cambodia into the Association of Southeast Asian Nations (ASEAN). Becoming a member of ASEAN, the need for being proficient in English in Cambodian society became more demanding due to the fact that the use of English had been mandated in article 34 as the only working language of communication by all ASEAN members (Clayton, 2006; Association of

Southeast Asian Nations, 2007). In addition, the use of English had been promoted as “an internal business language at the work place” which was one of ASEAN’s plans for integrating its regional community in 2015 (ASEAN Secretariat, 2009, p. 3). Thus, in order to fully cooperate and actively engage with the ASEAN community, there was a societal need for Cambodians to be proficient in English language communication.

1.3 English Language Taught in Cambodian Schools and University

Given the increased need for being proficient in the English language in Cambodian society, it was officially permitted to be taught in Cambodian secondary schools for five hours per week from grade 7 onwards in 1989. The newly established English subject, however, faced some challenges due to the lack of teaching and learning resources as well as the shortage of teachers of English language to teach this new language in all Cambodian secondary schools (Neau, 2003; Clayton, 2006).

To facilitate the implementation of this new language policy, an Australian non-government organisation, namely Quaker Service Australia (QSA) funded by the Australian government, set up the Cambodian English Language Training Programme to provide training to both Cambodian government staff and English language teachers in secondary schools. The QSA project was undertaken within three distinct phases: 1985-1988, 1988-1991 and 1991-1993. Owing to the demands for English Language Training in Cambodia, the Bachelor of Education in Teaching English as a Foreign Language (TEFL) programme was established at the University of Phnom Penh in 1985 (Suos, 1996). In line with the Australian government’s Cambodian secondary school English language teaching project, the British government sponsored the Cambodian-British Centre for Teacher Education and the Cambodian Secondary English Language Teaching project (CAMSET) from 1992 to 1997 to kick-start their English language programmes with the aim of training Cambodian teachers of English as a foreign language (EFL) for secondary schools (Kao & Som, 1996). Eventually, these trained EFL teachers were also provided with opportunities to teach at university level, given the lack of English language teachers within the tertiary setting (Suos, 1996).

As a result of these initiatives implemented by both the Australian and British governments, since the early 1990s all Cambodian secondary school students as well as

most tertiary students were provided with opportunities to study English as a Foreign Language (EFL). Recently, the Cambodian Ministry of Education announced that English language was permitted to be taught in primary schools, starting from grade 4 onwards, and this new language programme began in late 2013 (Kuch, 2013). Given its popularity, some public and private universities have set up a bachelor's degree in teaching English as a Foreign Language (TEFL) and a master's degree in Teaching English to Speakers of other Languages (TESOL) to continually train more teachers for both school and tertiary settings. Furthermore, given the majority of teaching and learning resources across all discipline areas are written in English, most Cambodian universities that offer bachelor, master and doctorate degrees in fields other than TEFL/TESOL also require their students to take English language courses in addition to their major courses. Thus, acquiring an adequate level of English language proficiency has been seen as critical for Cambodian people in order to enable them to fully participate and actively engage within both everyday activities in their society, higher education studies, employment and also within the ASEAN community. The following section provides the purpose for the current study and the research questions employed.

1.4 Purpose of the Study

The primary purpose of the current study was to develop and validate a set of scales to examine the classroom assessment literacy development of instructors within EFL programmes in a Cambodian higher education setting. The study examined the interrelationships amongst the four constructs (i.e., Classroom Assessment Knowledge, Innovative Methods, Grading Bias and Quality Procedure) that were thought to underpin classroom assessment literacy of the instructors. It also sought to examine the level of instructors' classroom assessment literacy and its associated impact on their actual assessment implementation. It further investigated the influence of instructors' background characteristics on their classroom assessment literacy. To gain further insights regarding the nature of instructors' classroom assessment literacy development, the study employed a mixed methods approach.

1.4.1 Research Questions

Given that previous exploratory research has highlighted that classroom assessment knowledge base and personal beliefs about assessment are the underpinnings of classroom assessment literacy of instructors, there is a need to examine the interrelationships amongst these variables that form the instructors' classroom assessment literacy construct. The present study tested a hypothesised conceptual measurement model to confirm whether or not the model could represent relations amongst the four constructs of classroom assessment literacy. It also sought to investigate the instructors' classroom assessment literacy level and its associated impact on assessment practices. It further examined the influence of the instructors' background characteristics on their classroom assessment literacy.

The main research question explored in this study was: *“To what extent did assessment related knowledge and beliefs underpin classroom assessment literacy and to what extent could each of these constructs be measured”*? Subsidiary research questions comprised:

1. To what extent was classroom assessment literacy developmental?
2. What impact did classroom assessment literacy have on assessment practices?
3. How did the background characteristics of instructors (i.e., age, gender, academic qualification, teaching experience, teaching hours, class size, assessment training, and departmental status) influence their classroom assessment literacy?

1.5 Significance of the Study

This research is the first empirical study of EFL classroom assessment literacy within a tertiary education setting in Cambodia. It is one of the few studies that have employed a mixed methods approach in measuring EFL instructors' classroom assessment literacy development within a classroom-based context. Despite the fact that this study has been undertaken in a specific language educational setting, the findings will contribute to the general understanding of classroom assessment literacy in tertiary education. It could also make a contribution to the development of the classroom

assessment literacy scales in the field. Given the desire of achieving high quality classroom assessments, educational researchers and assessment specialists alike are looking to the factors that underpin classroom assessment literacy of the instructors. High quality classroom assessments have the potential to enable students to acquire lifelong learning skills and/or higher-order thinking to fulfil the goals of educational reform and equip them with skills needed for the twenty-first century. It is therefore essential to better comprehend the nature of instructors' classroom assessment literacy development, so that appropriate remedies can be used to address the issues in a timely manner, given instructors are the key agents in the assessment process. Thus, the development and validation of a set of scales to measure the instructors' classroom assessment literacy progression, undertaken within the current study, could address these needs. The findings from the present study further provide important implications for theory, policy and practice, and the design of pre-service teacher education programmes. The study also provides a valuable framework for future classroom assessment literacy research.

1.6 Structure of the Thesis

The thesis is organised into nine chapters. Chapter one provides the rationale for the study, an overview of the demands of English language in Cambodian society, a snapshot of English language taught in Cambodian schools and university settings, and outlines the purpose of the study, the proposed research questions and significance of the study. Chapter two explores the key stages of classroom assessment processes, together with the body of studies on classroom assessment practices as they relate to the assessment process. Chapter three proposes the theoretical framework that underpins the design of the study. This chapter also explores the concept of literacy in general and various definitions of classroom assessment literacy, and further documents the key factors that underpin classroom assessment literacy. Chapter four explores a range of background characteristics of instructors thought to impact on their classroom assessment literacy development. Chapter five presents the methodology employed in the study, in terms of a mixed methods approach including quantitative and qualitative methods. Chapter six documents the development and validation of a set of scales underpinning the study. Chapter seven presents the univariate, bivariate and multivariate results from the

quantitative phase of the study. Chapter eight presents the results from the qualitative phase of the study. Chapter nine integrates the results from both quantitative and qualitative phases of the study and discusses the implications of the findings according to theory, policy and practice, and the design of pre-service teacher education programmes. This chapter also discusses the study's limitations and future directions for research in the area of classroom assessment literacy.

Chapter 2: Classroom Assessment Processes

This chapter explores current research and development activities in the field of educational assessment, and in particular classroom-based assessment, which can be applied within a range of contexts including language and general courses within higher education programmes. Where possible, lessons learnt from other educational settings, such as the school sector, have been explored. The chapter has been structured in terms of the key stages within the assessment process, namely: assessment purposes, methods for gathering evidence of student performance, interpretation frameworks, grading decision making, recording and reporting formats. Within each of these key stages, factors that impact on the validity and reliability of the assessment have been explored. Finally, this chapter explores a range of theoretical frameworks for quality management of the assessment process.

2.1 Classroom Assessment

There are disparate views on the notion of classroom assessment and this is largely due to the fact that the terms “assessment,” “measurement,” “testing” and “evaluation” are used interchangeably in both the literature and in practice, despite the fact that each has a specific meaning (Griffin & Nix, 1991; Miller, Linn, & Gronlund, 2013; Mathew & Poehner, 2014).

The interchangeable use of these terms may be due to the fact that they are all involved in a single process (Griffin & Nix, 1991; Miller et al., 2013). An “assessment” is typically associated with the procedures used to describe the characteristics of an individual or something. In contrast, a “measurement” is relevant to the comparison of an observation such as assigning numbers/marks for particular questions in the test. Unlike its counterparts, “testing” refers to an attempt used to determine the worth of an individual’s effort and it typically contains a set of questions to be administered during a specific period of time (Griffin & Nix, 1991; Miller et al., 2013). An “evaluation” however tends to be associated with making judgments of worth of an individual or something (Griffin & Nix, 1991). Thus, an analysis of each of the meanings of these

terms indicates that assessment is much broader, and can include testing, measurement and evaluation in the processes employed to collect information about any individual's characteristics (Griffin & Nix, 1991; Miller et al., 2013).

The term “classroom assessment” is used to emphasise a classroom-based context and to avoid connotations of the term “testing” with standardised paper and pencil tests and/or large-scale tests, since the term “assessment” tends to be used synonymously with the term “testing” in the literature (Rea-Dickins, 2007). For example, Huerta-Macias (2002) and Brookhart (2004) distinguish classroom assessment from large-scale tests and/or standardised paper and pencil tests in that it can be embedded within instruction. Rea-Dickins (2007) and Mathew and Poehner (2014) also refer to classroom assessment as the procedures by which students' performance are interpreted in terms of learning goals and instruction processes, as opposed to a finished product measured by large-scale tests. Cumming (2010), Stobart and Gipps (2010) and Hill and McNamara (2012) further identify classroom assessment as the assessments employed to enhance instruction, promote learning and report achievement. Thus, the meaning of classroom assessment is relevant to assessments conducted to enhance instruction and learning as well as to report achievement, and it is typically undertaken by the teachers during their teaching time rather than being administered separately during a fixed period of time as per large-scale tests. As such, the term “classroom assessment” is used for school and higher education settings, while the term “assessment” can be applied to a broader range of contexts, other than school and higher education institutions.

The term “assessment” has its roots in the Latin word *assessare* which means “to impose a tax or to set a rate” (Athanasou & Lamprianou, 2002, p. 2). According to the *Cambridge Advanced Learner's Dictionary* (2008), the term “assessment” has been defined as the determination or evaluation judged by any individuals on the nature and degree of an object and/or thing surrounding them. Given its important role, the term “assessment” quickly spread to education (Athanasou & Lamprianou, 2002). Within school and higher education contexts, this term typically refers to the process of gathering and organising evidence of student learning for making inferences about teaching and learning activities (Lamprianou & Athanasou, 2009; Chappuis, Stiggins, Chappuis, & Arter, 2012; Russell & Airasian, 2012; McMillan, 2014; Popham, 2014). As such,

assessments can be conducted in a variety of settings, including language and general education within the higher education and/or school sector, vocational education, as well as external environments or within the workplace. Given the current study is situated within EFL higher education programmes, the discussion of each key stage within the assessment process is therefore specific to a classroom-based assessment context. Assessments conducted within language and general education follow similar procedures, despite the fact that language education emphasises students' language achievements rather than focuses on achievements more broadly, as in general education (Rea-Dickins, 2007). The next section explores the theoretical underpinnings of classroom assessment processes, which can be applied within a range of educational settings. In the following discussions, the term "teacher" is used throughout this chapter to refer to school teacher and/or tertiary instructor.

2.2 Classroom Assessment Processes

An assessment process within an educational setting typically encompasses the following key components: defining the purposes of the assessment, constructing or selecting assessment methods to collect evidence of learning, interpreting assessment outcomes collected, grading decision making, recording assessment information, and reporting assessment results to relevant stakeholders comprising students, parents, administrators, potential employers and/or teachers themselves (Gillis & Griffin, 2008; Lamprianou & Athanasou, 2009; Chappuis et al., 2012; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014). Moreover, there is consensus that the assessment process must cover validity and reliability characteristics, given they play a crucial role in providing accuracy, fairness, and appropriateness of the interpretations and uses of assessment results (Cizek, 2009; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014). Furthermore, it has been argued that quality management should be integrated into the assessment process to achieve accuracy, appropriateness, fairness and transparency of assessment outcomes in order to ensure comparability of standards between classes and within schools/universities (Dunbar, Koretz, & Hoover, 1991; Gipps, 1994b; Harlen, 1994, 2007; Gillis, Bateman, & Clayton, 2009). Hence, validity, reliability and quality

management must be taken into consideration within each stage of the whole assessment process (see Figure 2.1).

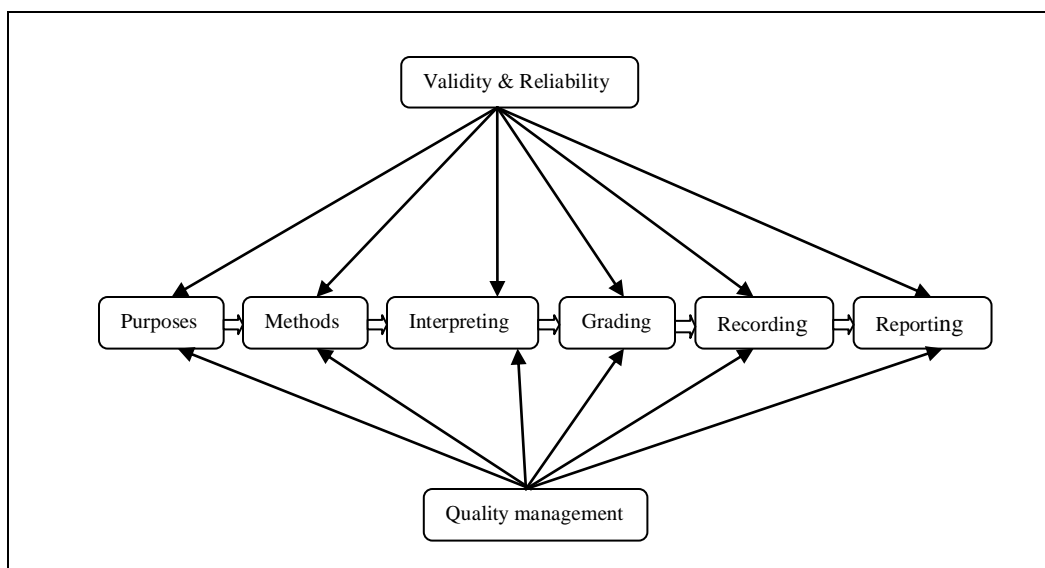


Figure 2.1 Classroom assessment processes

As illustrated in Figure 2.1, the concepts of validity and reliability will be explored, followed by each stage within the classroom assessment process; section 2.2.9 then explores a range of theoretical frameworks for quality management of the assessment process.

2.2.1 Validity

The concept of “validity” has been referred to as “the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores” (Messick, 1989, p. 13). Despite the fact that this definition is dated (nearly twenty-five years’ old) and there is widespread acceptance in the literature that assessment is more than just test scores, interpretations are still meaningful and crucial to modern day educational assessments. Various types of validity have been proposed including content, construct, consequential, face and criterion (Messick, 1989; Bachman, 1990; Bachman & Palmer, 1996; Kane, 2006; Gillis

& Griffin, 2008; Lamprianou & Athanasou, 2009; Miller et al., 2013) and these will be discussed next.

Content validity has been defined as the extent to which the assessment tasks provide a relevant and representative sample of the learning domains to be measured (Messick, 1989; Kane, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013; Popham, 2014). To enhance content validity for the purpose of achieving accurate measures of students' learning achievements, it has been argued that teachers and/or assessment developers should take into consideration four key steps in developing their assessment tasks (Lamprianou & Athanasou, 2009; Chappuis et al., 2012; Miller et al., 2013; Popham, 2014). Firstly, they should identify the intended domain of the learning outcomes (i.e., assessment purposes). Secondly, they should prioritise the learning goals and objectives to be measured through creating a table of specifications for learning aspects to fulfil the identified purpose(s). Thirdly, they should construct or select the assessment items/tasks based on the table of specifications. Finally, they should assign weightings for each assessment item/task based on its importance in achieving the learning goals and curriculum objectives. Hence, content validity is crucially important as it reflects the course learning objectives/goals. As such, when assessment is conducted for summative purposes (i.e., awarding certificates/degrees), a high level of content validity is required.

Construct validity has been referred to as the extent to which an assessment task can be interpreted as a meaningful measure of some characteristics or qualities of the student (Messick, 1989; Kane, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). That is, construct validity is concerned with the degree to which the assessment task adequately represents the intended construct, as well as the degree to which the students' performance has been influenced by other factors that are irrelevant to the intended construct (Messick, 1989; Kane, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). When the assessment task does not adequately measure the intended knowledge and/or skills of students (e.g., short test), this issue is known as construct underrepresentation. When students' performance has been influenced by other factors (e.g., personal interest) that are irrelevant to the intent of the assessment tasks, this issue has been known as construct-irrelevant variance.

Consequential validity is concerned with the extent to which the assessment results can achieve intended assessment purposes and avoid unintended or negative impacts on teaching and learning (Messick, 1989; Kane, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). Consequential validity comprises intended consequences (i.e., using assessment results to enhance instruction and improve learning) and unintended consequences (i.e., teaching to the assessment tasks that may result in reducing learning and narrowing the curriculum).

Face validity has been associated with the appearance of the assessment (Messick, 1989; Kane, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). Face validity is concerned with the degree to which the assessment tasks are likely to be a reasonable measure of the learning domain and tends to be based on the superficial examination of the tasks. As such, face validity appears to be less important than content, construct and consequential validities. In particular, in a higher education classroom-based assessment of language skills, face validity is not as important as other measures of validity, but this is not the case for all educational sectors. For instance, in applied courses or vocational education, face validity is extremely important- otherwise stakeholders will not accept the results. That is, the assessments of practical skills need to simulate the real world, profession and/or workplace.

Finally, criterion validity has been referred to as the extent to which the assessment task predicts students' future performance and/or estimates students' performance on some measures other than the assessment task itself. Criterion validity has been divided into two types: predictive and concurrent (Messick, 1989; Kane, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). Predictive validity is associated with predicting the relationship between two measures over an extended period of time, whereas concurrent validity refers to the relationship between two measures obtained concurrently. In contrast to other types of validity, criterion validity has been argued to be irrelevant to classroom teachers due to its impractical nature (Lamprianou & Athanasou, 2009; Miller et al., 2013). In other words, within a classroom-based assessment context, criterion validity is not as important as other measures of validity, because teachers rarely use their assessment results to relate to other measures and/or to predict future performance of the student. The criterion validity, however, is important for EFL

programmes that use externally developed standardised tests. Thus, criterion validity is important in a standardised test, as its results are typically employed to predict the likely performance of the student in other settings (Miller et al., 2013). Nevertheless, it should be noted that the current study is limited to examining classroom assessment where the locus of control for assessment task development is at the teacher level.

As it is unlikely for classroom assessment to satisfy all five types of validity (i.e., due to practicalities) discussed above, the importance for classroom assessments to have demonstrated content, construct and consequential validities has been well documented (Lamprianou & Athanasou, 2009; Miller et al., 2013). It is thought that such validity types help to provide for sufficiency, fairness, appropriateness of the interpretations and uses of assessment results to key stakeholders. Within the EFL higher education programmes, the key stakeholders of the assessment results typically are teachers, students, parents, administrators and/or relevant employers.

2.2.2 Reliability

In addition to determining the extent to which the interpretation and use of classroom assessments are valid, the reliability aspect needs to be equally addressed. It is nonetheless noted that while reliability has been considered as necessary, it does not provide sufficient conditions for the validity of the assessment results (Lamprianou & Athanasou, 2009; Miller et al., 2013). The concept of “reliability” has been defined as the accuracy or precision of the measurement (Cronbach, 1951, 1990). That is, reliability relates to the results of assessment rather than the assessment instrument itself. Reliability is typically determined using statistical indices. There are six types of reliability:

1. test-retest;
2. equivalent forms (also referred to as parallel or alternative forms);
3. split-half;
4. Kuder-Richardson or coefficient alpha (Cronbach, 1951, 1990);
5. intra-rater; and
6. inter-rater.

(Haertel, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013).

The test-retest method is associated with administering the same assessment tasks to the same group of students twice, with a sufficient interval time between these two periods of administration. The assessment results are then correlated and the correlation coefficient obtained is used to provide evidence on how stable the assessment results are over these periods of time. Similar to the test-retest method, the equivalent forms method is conducted through administering two equivalent forms of assessment, having similar content and levels of difficulty, to the same group of students with two different periods of time. Then the assessment results obtained from these two equivalent forms of assessment are correlated. The correlation coefficient obtained suggests the extent to which these two assessment tasks are assessing the same aspects of behaviour.

In contrast to test-retest and equivalent forms methods, the split-half method is undertaken by administering assessment tasks at a single point in time to a group of students. Subsequently, the assessment tasks are divided into two equivalent parts during the marking period, and typically the odd- and even-numbered assessment tasks are marked separately. Through this procedure, each student receives two different scores and the correlation coefficient of the two scores provides evidence of internal consistency. The Kuder-Richardson or coefficient alpha method is similar to the split-half method, where assessment tasks are administered once to a group of students. The coefficient alpha obtained from the assessment tasks provides evidence of internal consistency (Cronbach, 1951, 1990; Haertel, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013).

Intra-rater and inter-rater reliability indices are relevant to assessments that involve subjective judgment by teachers in marking students' work (e.g., essays, assignments and performance) (Haertel, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). Intra-rater reliability refers to consistency in marking the students' work/performance by the same teacher at different times. In contrast, inter-rater reliability is relevant to the extent to which the consistency of marking students' responses by two or more teachers can be achieved. In examining inter-rater consistency, the scores given by one teacher are usually correlated with those given by another teacher. To achieve an acceptable level of inter-rater or intra-rater consistency, it requires

the teachers to fully understand the marking criteria and standards before assessing the students' work/performance and consensus amongst teachers needs to be reached to avoid unfair treatment (Lamprianou & Athanasou, 2009; Miller et al., 2013).

Of the six reliability methods discussed, the first four (test-retest, equivalent forms, split-half, Kuder-Richardson or coefficient alpha) are relevant to paper and pencil testing, while the latter two (intra-rater and inter-rater) relate to performance-based assessment, in subjective judgement is exercised by teachers. With regard to classroom assessment, test-retest and equivalent forms methods of reliability are less relevant to classroom teachers, given it is unusual to administer assessment tasks to a group of students twice (Lamprianou & Athanasou, 2009; Miller et al., 2013).

2.2.3 Assessment Purposes

In implementing classroom assessment, teachers firstly need to take into consideration assessment purposes. There is general agreement on a variety of common functions in conducting classroom assessment including:

- instructional purposes (i.e., to adjust instruction to student level) (Chappuis et al., 2012; Russell & Airasian, 2012; McMillan, 2014; Popham, 2014);
- placement purposes (i.e., to put students in different levels) (Hughes, 1989; Bachman & Palmer, 1996; Shute & Kim, 2014);
- evaluation purposes (i.e., to determine progress in learning) (Chappuis et al., 2012; Russell & Airasian, 2012; McMillan, 2014; Popham, 2014); and
- accountability purposes (i.e., to provide information to administrators) (Chappuis et al., 2012; Russell & Airasian, 2012; Popham, 2014).

Other assessment specialists classify classroom assessment purposes into two broad types: formative and summative (Bloom, Hastings, & Madaus, 1971; Harlen & James, 1997; Harlen, 2005a; Wiliam, 2010; Brookhart, 2011b; Chappuis et al., 2012; McMillan, 2014). Assessment used for a formative purpose is typically associated with enhancing instruction and improving learning, whereas a summative purpose is relevant to summing up learning achievements to be communicated to administrators and/or other

relevant stakeholders. The terms “formative” and “summative” assessments were first used by Michael Scriven (1967) in connection with the improvement of curriculum. Subsequently, Bloom et al. (1971) have extended the definition of “formative assessment” to one that assists curriculum developers, teachers and students to improve teaching and learning. In contrast, the term “summative assessment” has been referred to as the type of assessment employed at the end of units, mid-term or the end of a semester/course for grading, certification, evaluation of progress, or even for researching the effectiveness of a curriculum (Bloom et al., 1971). Wiliam and Black (1996) also define formative assessment as providing evidence for improving students’ performance in specific activities, whereas summative assessment sums up the evidence of learning achievements.

More recently, classroom assessment purposes have been classified into four types labelled as: “assessment is for teaching” (Care & Griffin, 2009), “assessment as learning” (Earl, 2013), “assessment for learning” and “assessment of learning” (Lamprianou & Athanasou, 2009; Chappuis et al., 2012; Popham, 2014). For instance, Care and Griffin (2009) have argued that “assessment is for teaching” because assessments can be employed to identify the students’ zone of proximal development (Vygotsky, 1978). That is, assessments are designed to identify the point at which the students are most ready to learn, so that teaching interventions will have the greatest impact on students’ learning. In contrast, Earl (2013) has argued for “assessment as learning”, because assessments can be used to provide students with opportunities to actively involve themselves in the assessment process, in order to develop and support their metacognition. Metacognition is associated with the students’ ability to reflect on their own thinking with respect to their learning and make adjustments accordingly. That is, students act as agents in the assessment process, linking assessment and learning. Others (Lamprianou & Athanasou, 2009; Chappuis et al., 2012; Popham, 2014) have argued for “assessment for learning” and “assessment of learning”, given that assessment can be employed to enhance teaching and learning as well as to determine students’ achievements. Assessment for learning is typically conducted to plan future instruction, diagnose students’ needs, and offer feedback in improving their work quality and engaging them in the assessment process. In contrast, assessment of learning is

undertaken to gather evidence to determine students' achievements/grades at a single point in time or to make decision about programmes. Thus, "assessment is for teaching," "assessment as learning" and "assessment for learning" are terms for formative assessment while "assessment of learning" is a term for summative assessment. Table 2.1 below summarises the main functions for conducting classroom-based assessment.

Table 2.1 Main Types of Assessment Purposes

Formative Purposes	Summative Purposes
instruction	assessment of learning
placement	evaluation
assessment is for teaching	accountability
assessment as learning	
assessment for learning	

The main purposes for conducting classroom assessment can be classified as either formative or summative, depending on the intended use of results (see Table 2.1). Classroom assessment could be used for summative purposes, but at the same time it could also serve as feedback for teaching and learning improvement or vice-versa (Harlen & James, 1997; Broadfoot, 2005; Harlen, 2005a; Taras, 2005; Brookhart, 2010; Black, 2013; Earl, 2013; Sambell, McDowell, & Montgomery, 2013; Leung, 2014).

Research has revealed that teachers employ their classroom assessments to serve a variety of purposes. For instance, within language education, Cheng and colleagues conducted longitudinal studies on the assessment practices in three ESL/EFL tertiary settings (Cheng, Rogers, & Hu, 2004; Cheng & Wang, 2007; Rogers, Cheng, & Hu, 2007; Cheng, Rogers, & Wang, 2008). Employing questionnaires in phase one and interviews in phase two of their studies, it was found that there were significantly different assessment purposes employed by the teachers at these tertiary institutions. Canadian teachers reported that their assessment purposes were to gain information on students' progress, and offer feedback to students and identify their strengths and weaknesses, whereas Chinese teachers stated that their assessment purposes were to prepare students for standardised tests. In contrast to their counterparts, Hong Kong

teachers reported that their assessment purposes were to give information to the administrators and to determine final grades for students' achievements. Similarly, Xu and Lix (2009), employing a narrative inquiry with one college EFL teacher in China, found that this teacher's assessment purpose had a summative function.

These different purposes for classroom assessment can have a close relationship with validity and reliability characteristics. That is, low validity and reliability tend to be acceptable in relation to assessment conducted for formative purposes (i.e., enhancing instruction and improving learning), whereas it has been strongly argued that assessment conducted for summative purposes (i.e., giving certificates/degrees) must possess both high validity and reliability, given the decisions associated with assessment results have significant consequences on the lives of students and/or teachers (Lamprianou & Athanasou, 2009; Douglas, 2010; Miller et al., 2013).

It has been argued that assessment with summative purposes must possess high content validity. Assessment results must reflect the learning outcomes specified in the curriculum to accurately indicate students' actual learning achievements throughout the courses (Lamprianou & Athanasou, 2009; Chappuis et al., 2012; Miller et al., 2013; Popham, 2014). It has also been asserted that assessment purposes and how the evidence is interpreted and used by assessment stakeholders are inextricably linked to the consequential validity of the assessment (Messick, 1989). Generally, different people have various interests in the outcomes of assessment. This can lead to different stakes of an assessment: low versus high (Gillis, 2003; Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Miller et al., 2013).

It has been further pointed out that the nature of the stakes of an assessment can be determined by the consequences of the intended use of assessment results. Low stakes assessment tends to be associated with conducting assessment for motivation and diagnostic purposes (i.e., formative functions), whereas high stakes assessment tends to be employed for placement, selection or evaluative purposes (i.e., summative functions) (Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Miller et al., 2013). Research has provided sufficient evidence concerning the relationship between high stakes assessment and the negative and unintended consequences of assessment on teaching and learning (Alderson & Hamp-Lyons, 1996; Bailey, 1996; Shohamy, Donitsa-

Schmidt, & Ferman, 1996; Cheng, 2005; Tsagari, 2009). These negative and unintended consequences have been referred to as undesirable washback/backwash effects from assessment (Alderson & Wall, 1993; Wall & Alderson, 1993; Cheng, 2008; Cheng & Curtis, 2012; Wall, 2012; Green, 2014). Studies on high stakes assessment have shown a tendency for teachers to teach to the assessment tasks, resulting in reducing learning and narrowing the curriculum (Popham, 1991; Nolen, Haladyna, & Haas, 1992; Bailey, 1996; Shohamy et al., 1996; Cheng, 2005; Luxia, 2007; Amengual-Pizarro, 2009; Tsagari, 2009). In such situations, textbooks used predominantly influence the content of teaching and classroom tests (referred to as content washback) (Cheng, 2005; Tsagari, 2009) and students typically employ a surface learning approach (i.e., memorising facts and details of the learning materials to handle the assessment tasks) rather than using a deep learning approach (i.e., understanding, integrating and relating the learning materials critically in order to successfully complete assessment tasks) (Marton & Säljö, 1976a, 1976b; Thomas & Bain, 1984; Nolen & Haladyna, 1990; Entwistle & Entwistle, 1992; Scouller, 1998; Dahlgren et al., 2009). Thus, high stakes assessment can induce both teachers and students to focus exclusively on what is thought to be covered in the assessment tasks, and ignore all the important learning goals and objectives, resulting in reducing student learning and narrowing the school curriculum (Binkley et al., 2012; Miller et al., 2013).

To minimise the undesirable consequences of assessments on teaching and learning (i.e., unintended consequential validity), research has revealed a number of ways to foster intended consequences (Hughes, 1989; Heyneman & Ransom, 1990; Kellaghan & Greaney, 1992). Hughes (1989) suggests that assessment tasks should reflect the course learning objectives/goals (i.e., content validity). Heyneman and Ransom (1990) argue that assessment tasks should be designed to include more open-ended tasks rather than selected-response methods, as the former is better suited to measuring higher-order thinking skills (i.e., construct validity). Kellaghan and Greaney (1992) go further and argue that the assessment tasks should sample across course learning objectives and comprise a variety of written, oral, aural and practical skills. Messick (1996) adds that the assessments should consist of tasks that are criterion-referenced. Thus, ensuring content and construct validities can assist teachers overcome the unintended consequences of assessments.

Research has further revealed that high stakes assessment tends to cause teachers to display unethical behaviours such as disclosing the content of a test (Nolen, Haladyna, & Haas, 1992; Lai & Waltman, 2008) and helping students to cheat (Gay, 1990; Herold, 2011). Students also engaged in cheating during test administration time (Lin & Wen, 2007; Zimny, Robertson, Bartoszek, 2008; Eastman, Iyer, & Reisenwitz, 2011). As such, these unethical behaviours (i.e., construct-irrelevant factors) can impact on the validity of assessment results. This raises an issue about the degree to which assessment results can accurately represent students' actual learning achievements. Hence, there is widespread agreement within the educational assessment literature that decisions with regard to student achievement (e.g., pass/fail and/or to award certificates/degrees) should not be merely based on a single assessment task/examination result (Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Miller et al., 2013). Instead, various sources of student learning achievements collected during the course should be considered, prior to making any decisions, particularly for high stakes assessments (Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Miller et al., 2013).

The various purposes of classroom assessment are also relevant to the ethicalness or fairness of assessment results used for the students. The main role of ethics in classroom assessment is to take into consideration the fairness of assessments as well as their appropriate and inappropriate overall use (Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Douglas, 2010; Brown, 2012; Masters, 2013a; Miller et al., 2013; Tierney, 2013; Kunnan, 1999, 2014). Davies (1997a) has argued that the key role of ethics needs to concern the balance between social and individual justice (i.e., the fairness of the use of the assessment results for individual students and the social consequences associated with their assessment results). Davies then raises an issue as to whether teachers or assessment developers should take any responsibility beyond the construction of assessment tasks. In responding to Davies' concern, Shohamy (2001) has identified five key responsibilities that teachers or assessment developers must be aware of: (1) responsibility for making others aware; (2) responsibility for all assessment consequences; (3) responsibility for imposing sanctions; (4) shared responsibility; and (5) ethical responsibility. Nevertheless, it has been argued that it is impossible for teachers or assessment developers to undertake all responsibilities concerning the consequences of

the use of the assessment results for unintended purposes (Davies, 1997b). It has been further argued, however, that teachers or assessment developers primarily have an ethical responsibility to involve students in the assessment process (Taras, 2013), to ensure that assessment supports student learning (Tierney, 2013) and to establish where students are in their learning at the time of assessment (Silis & Izard, 2002; Griffin, 2009; Forster & Masters, 2010; Heritage, 2013b; Masters, 2013a). Irrespective of the results of assessment for summative purposes, assessment can still be employed to feed into learning, and therefore can also serve a formative function although the primary function is summative.

Douglas (2010) has also suggested that decisions made on the basis of assessment results must be a true reflection of student learning achievements. Spolsky (2014) has added that decision making based on assessment results should be dependent on a variety of information regarding student learning. Miller et al. (2013) have supported Douglas's (2010) and Spolsky's (2014) perspectives by pointing out that the negative and unintended consequences of assessment on individual students, caused by the misinterpretations and misuse of assessment results, must be avoided, particularly in summative purposes within a high stakes context. Guskey (2013) has further asserted that all types of assessments comprise some degree of error, both random and systematic and he therefore recommends that decisions regarding high stakes for students based on assessment results must always be made with caution and care. Hence, assessment ethicalness appears to be relevant to both validity (i.e., content, construct and consequential) and reliability (i.e., coefficient alpha and inter-rater or intra-rater). As such, it is necessary for teachers to ensure their assessments have sufficient levels of validity and reliability characteristics in order to obtain ethical assessment or fairness for their students prior to making high stakes decisions (i.e., summative purposes) based on assessment results.

2.2.4 Assessment Methods

Having decided on the purposes of classroom assessment, the second step is to select or develop the types of assessment methods that match the identified assessment

purpose(s). A number of assessment specialists have categorised classroom assessment into two broad types: selected-response assessments (e.g., true-false, matching and multiple-choice questions) and constructed-response assessments (e.g., gap-filling and short answer questions) (Brown & Hudson, 1998; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014). The main types of assessment methods have been displayed in Table 2.2 below.

Table 2.2 Main Types of Assessment Methods

Traditional/Objective assessment	Innovative/Alternative assessment
<p>Selected-response assessments include:</p> <ul style="list-style-type: none"> • True-false items • Matching items • Multiple-choice items <p>Constructed-response assessments include:</p> <ul style="list-style-type: none"> • Gap-filling items • Short answer items 	<p>Authentic assessments include:</p> <ul style="list-style-type: none"> • Performance-based assessments • Self- and peer assessments • Portfolio assessments

As can be seen from Table 2.2, true-false, matching, multiple-choice, gap-filling and short answer methods are regarded as traditional/objective assessment, whereas performance-based assessment, self- and peer assessment and portfolio assessment methods are regarded as innovative/alternative assessment (Brown & Hudson, 1998; Fox, 2008; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014). Typically, traditional or objective assessment tends to be associated with measuring lower-order thinking skills. In contrast, innovative or alternative assessment has been associated with measuring higher-order thinking skills (Brown & Hudson, 1998; Stobart & Gipps, 2010; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014). Innovative or alternative assessment has also been referred to as an authentic assessment. Authentic assessment has been defined as one which corresponds with the features of the tasks involved in students' everyday lives (Bachman & Palmer, 1996; Lamprianou & Athanasou, 2009; Miller et al., 2013). Authentic assessment tasks have been further argued to have higher

face validity than traditional assessment (Bachman & Palmer, 1996; Lamprianou & Athanasou, 2009; Miller et al., 2013).

The concept of “authenticity” was formally used in the learning and assessment context by Archbald and Newmann (1988) in connection with “authentic achievement”. Authentic achievement refers to student learning outcomes that have been assessed by authentic assessment tasks. Subsequently, Newmann and Archbald (1992) have expanded their notions of “authenticity” by stating that the quality and use of assessment rely on the degree to which outcomes can represent worthwhile, appropriate and meaningful accomplishment. Further, authentic assessment is also associated with construct validity, as it can have the potential to influence students’ actual performance (Spolsky, 1985; Bachman & Palmer, 1996). If students perceive an assessment task as a means to an end, which is not meaningful and does not involve real-life knowledge and skills, they are likely to put less effort in their performance due to their lack of interest. This construct-irrelevant factor can impact the validity of assessment results. It raises an issue about the generalisability of the interpretation of such results, as well as fairness issues in relation to how assessment results are used. For the purpose of clarity and consistency, in the following discussions, the terms “traditional assessment” will be used to refer to “objective assessment”, and “innovative assessment” will refer to “alternative assessment” respectively.

There are various types of traditional assessment methods commonly employed to assess student learning (Brown & Hudson, 1998; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014). True-false items/tasks are one format of a traditional assessment. This assessment requires students to choose one of two choices: true or false. The main aim here is to offer simple and direct implications as to whether or not a specific point has been understood. Similarly, in matching items, students are provided with two lists of words/phrases, and they are required to choose the words/phrases in one list that matches those in the other list. This type of item format is mainly used to measure students’ abilities to associate one set of facts with another. In alignment with true or false and matching items, multiple-choice items tend to present students with a number of options and they are required to choose the answer from a listing of plausible alternatives (with typically one correct

response and a number of distracters). This item format can be employed to assess a range of precise learning points. Similarly, in gap-filling items, students are given a text with part of the context removed, and replaced with a blank and they are required to fill in the blanks. This tends to be employed to assess students' abilities to produce a brief written response. In line with gap-filling items, short answer items present students with a question/statement and they are required to respond with phrases or sentences. This item format is used to assess a few phrases or sentences of the students' responses.

Employing traditional assessment has a number of benefits, including standardised administration, and it can be fast and easy to score the students' work (Brown & Hudson, 1998; Gibbs, 2006; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; Guskey, 2013; Miller et al., 2013; McMillan, 2014; Popham, 2014). Critics, however, have argued that using traditional assessment can only assess lower level thinking skills, surface or memorising learning, permits guessing and does not reflect meaningful and real-life knowledge and skills students are likely to encounter in their everyday lives (hence low face validity) (Boud, 1990; Russell & Airasian, 2012; Miller et al., 2013). Critics have further claimed that the use of traditional assessment tends to define the curriculum, as teachers and students emphasise the narrow topics/aspects in assessment rather than mastering the course objectives/goals, resulting in a narrowing of the school curriculum as well (Boud, 1990; Russell & Airasian, 2012; Miller et al., 2013).

However, unlike other methods of traditional assessment, the multiple-choice method has been argued for as being superior on the grounds that if it is well-constructed, it can measure different types of knowledge and complex learning outcomes, such as critical thinking and higher-order thinking skills, and can provide diagnostic feedback on learning areas needing improvement (Miller et al., 2013; McMillan, 2014; Popham, 2014). The plausible alternative options within each multiple-choice item can be used to identify weaknesses and/or difficult learning areas (Kehoe, 2002; Frary, 2002). Constructing high quality multiple-choice items is time-consuming and demands expertise on the part of the teachers and/or assessment developers (Miller et al., 2013). Unfortunately, it is unlikely that teachers have such skills, as pre-service teacher training programmes do not typically focus on educational measurement and testing (Stiggins,

1991b, 1999; Griffin et al., 2012; Leung, 2014). It has also been reported that teachers received little training on constructing multiple-choice items, resulting in a lack of confidence in using such methods (Kehoe, 2002).

In summary, traditional assessment methods developed by classroom teachers have the potential to negatively impact on student learning, but their attractiveness lies in the ease of enhancing reliability. Hence, validity can be compromised in the pursuit of reliability. As such, decisions based exclusively on the results of traditional assessment methods within a classroom-based context need to be cautioned, given the potential for these results to inaccurately reflect the learning objectives/goals specified in the curriculum.

With respect to innovative assessment, self- and peer assessments are one of the commonly employed methods, particularly within a higher education classroom-based context (Boud, 1995; Boud & Molloy, 2013; Raes, Vanderhoven, & Schellens, 2013). Self-assessment requires students to assess their own work while peer assessment requires students to assess the work of their classmates. Unlike traditional assessment, self- and peer assessments have been argued to: develop students' higher order thinking skills; motivate them to learn; involve them in the assessment process; assist them to become autonomous learners; and foster their feeling of ownership for their learning (Boud, 1990; Falchikov & Boud, 2008; Lamprianou & Athanasou, 2009; Nicol, 2013; Taras, 2013). These assessment methods tend to be associated with conducting assessments for formative purposes. Research has shown that the use of self- and/or peer assessment can motivate students to learn (Schunk, 1996; Munns & Woodward, 2006), improve their learning achievements and develop self-regulation skills (e.g., self-monitoring) that are important to their lifelong learning skills (Schunk, 1996; Munns & Woodward, 2006; Andrade, Du, & Wang, 2008; Ramdass & Zimmerman, 2008; Andrade, Du, & Mycek, 2010; Brown & Harris, 2013; Topping, 2013). This method can also reduce the teachers' own assessment workload (Sadler & Good, 2006).

However, there has been a concern associated with the social effects (e.g., friendship) on the markers' judgments in peer assessment (Pond, Ul-Haq, & Wade, 1995; Falchikov, 2004) and the issues of accuracy in relation to the use of self- and/or peer assessments for summative purposes (Pond, Ul-Haq, & Wade, 1995; Sluijsmans, Dochy,

& Moerkerke, 1999). To address such social effects, researchers have proposed a variety of strategies:

- use blind marking by substituting numbers for the actual student names (Sadler & Good, 2006) in peer assessment in order to eliminate the effects of friendship;
- train students in the necessary skills prior to undertaking self- and/or peer assessments in order to increase accuracy in student marking (Schunk, 1996; Sluijsmans et al., 1999); and
- involve students in constructing the scoring criteria used, and train them to mark their work and/or their peer's work using the scoring criteria they generate (Sadler & Good, 2006; Andrade et al., 2008; Andrade et al., 2010).

Hence, using self- and peer assessments tends to have a positive impact on students' learning, as it can empower them to be the owners of their learning and foster them with lifelong learning abilities (i.e., high validity), despite reliability issues raised in relation to their use for summative assessment purposes (Boud, 1990; Topping, 1998; Falchikov & Boud, 2008; Lamprianou & Athanasou, 2009; Brown & Harris, 2013; Nicol, 2013; Taras, 2013). As such, it has been widely recommended within the literature that teachers should carefully monitor and provide training and/or guidance to students prior to implementing such forms of assessment, as well as involve the students in generating the scoring criteria to foster their understanding of what is being expected from their work. This process can enhance the reliability of the assessments, particularly when they are employed for summative purposes (Boud, 1995; Brown & Harris, 2013; Topping, 2013). Teachers should bear in mind that low reliability of these assessments are still acceptable when they are used for formative functions, whereas high reliability is an important requirement when they are employed for summative functions (Topping, 2013).

Portfolio assessment is classified as another type of innovative assessment (Fox, 2008; Lamprianou & Athanasou, 2009; Klenowski, 2010; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014). This method involves a collection of samples of students' work that display their skills, efforts, abilities and achievements

during the course. The use of portfolio assessment has been thought to provide opportunities for both teachers and students to be involved in the assessment process, as well as to work together and reflect on evaluating learning growth (Klenowski, 2010; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014). Furthermore, it requires teachers to act as coaches/mentors to offer students insights in their learning, practice and revision processes, irrespective of the purposes of assessment (i.e., formative or summative functions). Thus, this method tends to provide high validity, as assessment results are based on a variety of learning aspects during the course. The main concerns regarding portfolio assessment, however, are related to the determination of the most appropriate way to show students' work, ability and improvement as well as marking consistency with the same teacher (i.e., intra-rater reliability) and/or across teachers (i.e., inter-rater reliability) (Brown & Hudson, 1998; Fox, 2008; Lamprianou & Athanasou, 2009; Klenowski, 2010; Russell & Airasian, 2012; Miller et al., 2013).

Performance-based assessment is also associated with innovative assessment (Brown & Hudson, 1998; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; McMillan, 2014; Popham, 2014). This method requires students to use a combination of knowledge and skills, such as speaking, listening, reading and writing, to accomplish given tasks. Tasks or activities include assignments, interviews, problem-solving tasks, communicative pair-work, role playing, observations, journal writing and group discussions. These types of method are employed to assess students' knowledge and skills through demonstrating their actual performance. The perceived strengths of performance-based assessment are associated with authentic activities (i.e., real-life tasks) and relevance to the school curriculum (i.e., students' knowledge and skills are measured against the learning goals and objectives specified in the curriculum). Performance-based assessment has the potential to achieve high validity of the assessment results in relation to student learning (Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; McMillan, 2014; Popham, 2014). However, there are a number of perceived weaknesses of performance-based assessments including marking consistency (within and across teachers) as well as difficulties in selecting sample tasks that represent all learning goals and objectives (i.e., challenging content validity) given its time-consuming nature (Lamprianou & Athanasou, 2009; Russell & Airasian, 2012;

McMillan, 2014; Popham, 2014). Another perceived weakness of performance-based assessment has been related to construct validity. Construct validity typically impinges on the way teachers define the construct in question, namely knowledge or skills. Generally, knowledge or skills are determined by the construct definition, that is, the terms in which they are described and the nature of the tasks and criteria employed to assess the students' performance. As such, various construct definitions may generate various assessment criteria (Brindley, 2000). Gipps (1994a) contends that there has been inadequate attention to construct validity and the definition of the domain of assessment within performance-based assessment. Gipps further contends that if the assessment criteria do not sample parts of the performance domain they are supposed to represent, or if the performance criteria rely on knowledge or skills from outside that domain, their construct validity is threatened.

It can therefore be seen that the main difference between traditional assessment and innovative assessment is in terms of the method of scoring. While innovative assessment (e.g., performance-based assessments) involves judgement on the part of teachers to score students' responses using marking criteria and standards, traditional formats of assessment, such as true-false, matching and multiple-choice items, do not (Hughes, 1989; Brown & Hudson, 1998; Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014). It should, however, be acknowledged that open-ended essays are also a traditional form of assessment and judgment is required in scoring such essays.

The role of judgment in classroom-based assessment has attracted a great deal of attention in the literature. For example, when using innovative assessments to judge students' work/performance, it has been argued that teachers are often influenced by various factors that do not necessarily reflect students' learning achievements (Price, 2005; Sadler, 2005; Bloxham & Boyd, 2007; Orr, 2008; Orrell, 2008; Popham, 2014). Such extraneous factors can include:

- student's gender (Spear, 1984; Bennett, Gottesman, Rock, & Cerullo, 1993; Harlen, 2005b; Read, Francis, & Robson, 2005; Malouff, 2008);
- student's behaviour (Bennett et al., 1993; Harlen, 2005b);
- student's general ability (Hoge & Butcher, 1984);

- teacher's tacit knowledge (i.e., values and beliefs) (Sadler, 2005; Orrell, 2008; O'Connor, 2009; Price et al., 2011; Bloxham, 2013);
- teacher's overall impression of students' characteristics (known as the halo effect) (Rudner, 1992; Dennis, 2007; Malouff, 2008; Popham, 2014);
- teacher's tendency to rate students' work higher than warranted (known as generosity error) and to underrate the quality of student's work (referred to as the severity error) (Popham, 2014); and
- teacher's tendency to give a similar assessment result for the current work due to the influence of students' past assessment outcomes (referred to as the spill-over effect) (Brown, 1998; Malouff, 2008; Orr, 2008).

To overcome the influence of such extraneous factors on assessment grading, it has been suggested that it is important for teachers to:

- develop explicit rubrics prior to marking students' work (Malouff, 2008; Lamprianou & Athanasou, 2009; Brookhart, 2013b; Selke, 2013); and
- mark students' work anonymously (Dennis, 2007; Malouff, 2008).

Rubrics have been referred to as the coherent sets of criteria used to mark students' work and the descriptions of levels of performance on these criteria are commonly known as standards (O'Connor, 2009; Brookhart, 2013b; Selke, 2013). Sadler (2007) also cautions teachers not to confuse the term "criteria" with the term "standard", given they tend to be used interchangeably in the literature, despite each having a specific meaning. He defines "criteria" as the properties or characteristics employed for judging the suitability of student work, whereas he associates "standards" with the levels in which students' work qualifies for a particular designation such as meeting the criteria for a specific grade. The attainment of an in-depth understanding of the rubrics can contribute to improving consistency in teachers' judgments of students' work. Grading students' work anonymously can further avoid and/or at least minimise teachers' influence of students' characteristics on their judgments of students' work. Given the rubrics can contribute to the improvement of reliability and validity of an assessment employed, it is important for teachers to construct explicit rubrics that reflect learning objectives/goals

(Miller et al., 2013). According to McNamara (2000), whether or not a student is assessed as meeting particular criteria relies on which teacher assesses his/her performance. He then suggests that the teachers or “raters may not be even self-consistent from one assessed performance to the next, or from one rating occasion to another” (p.38). Brindley (2000) further adds that teachers or “raters appear to differ markedly in severity”. He suggests that in order to ensure marking consistency amongst teachers, it is crucial to “constantly monitor the consistency with which ratings are administered” (p.32). It has been further argued that an explicit rubric is not in and of itself adequate, teachers’ tacit knowledge needs to be explicitly shared or discussed in order to make teachers aware of the potential bias in their judgments or at least to understand their own views (Price, 2005; Sadler, 2005; Bloxham, 2013). There is also a need to reflect on teachers’ practice as assessors because such reflections have the potential to ensure consistency and reliability in marking students’ work (Bloxham & Boyd, 2007). Sadler (2005) defines tacit knowledge as the implicit knowledge that expert teachers carry with them (mostly in their heads) when marking students’ work. Bloxham (2013) suggests that the more experience the teachers have in marking, the more their judgments become increasingly intuitive, and they tend to be unable to articulate their own tacit knowledge for marking students’ work.

However, the decisions for the selection of assessment tasks and aspects of the curriculum to be assessed are indeed subjective. It has been argued that all types of assessment have some element of subjectivity, even traditional assessment (Broadfoot, 2005; Harlen, 2005a). Despite the fact that innovative assessment tends to be plagued with more concerns associated with reliability issues than traditional assessment, assessment experts often recommend the use of innovative assessment more frequently than traditional assessment. This is largely due to the perception that innovative assessment has the potential to:

- assess students’ high-order thinking skills (Lamprianou & Athanasou, 2009; Russell & Airasian, 2012; Miller et al., 2013; McMillan, 2014; Popham, 2014);
- assess students’ broader range of knowledge, skills and attributes (O’Connor, 2009; Earl, 2013; Heritage, 2013b) essential for their lives in the twenty-first

century such as creativity, critical thinking, problem-solving, decision-making, flexibility, initiative, appreciation for diversity, communication, collaboration and responsibility (Binkley et al., 2012; Pellegrino & Hilton, 2012).

- provide students with opportunities to become the owners of their learning (Boud, 1990, Falchikov & Boud, 2008; Lamprianou & Athanasou, 2009; Nicol, 2013; Taras, 2013); and
- involve students in the assessment process aligned with constructive learning perspectives (Boud, 1995; Falchikov & Boud, 2008; Lamprianou & Athanasou, 2009; Nicol, 2013; Taras, 2013).

It has also been reported that there is some misunderstanding with respect to the use of traditional assessment to serve solely summative purposes, while the use of innovative assessment has been thought to exclusively serve formative purposes (Davison & Leung, 2009; Heritage, 2010). Such a misunderstanding may have arisen as traditional assessment has been predominantly related to the use of test/exam formats, whereas innovative assessment has been largely relevant to the use of performance-based assessment, portfolio assessment and self- and peer assessment methods. Because tests/exams usually provide limited information with respect to students' learning, they are often associated with summative functions. In contrast, given that performance-based assessments, portfolio assessments and self- and peer assessments have the potential to provide a variety of information regarding students' learning, these methods are generally associated with formative functions. Assessment experts, however, have pointed out that both traditional and innovative assessments can be employed to serve either formative or summative purposes, but the process in which these assessments have been implemented will determine the purpose (Harlen, 2005a; Davison & Leung, 2009; Heritage, 2010, 2013a).

Research has shown that school teachers are more likely to employ traditional assessment methods than innovative assessment formats when assessing students' learning within a general education setting (Gullickson & Ellwein, 1985; Oescher & Kirby, 1990; Entwistle & Entwistle, 1992; Bol & Strage, 1996; Greenstein, 2004; Allal,

2013). For example, Fleming and Chamber (1983) and Marso and Pigge (1993) analysed teacher developed tests in various subject areas. They reported that most of the items in the test assessed students' learning achievements at recall or knowledge level through using short answer, multiple-choice and matching items. Entwistle and Entwistle (1992) further reported that examinations set by teachers appeared to measure narrow forms of understanding and could force students into superficial learning activities a few weeks prior to the exam day. In line with Fleming and Chamber (1983), Marso and Pigge (1993), and Entwistle and Entwistle (1992), Bol and Strage (1996) employed interview and document analysis to explore the assessment processes of high school biology teachers and their instructional goals. The findings showed that teachers did not realise the difference between their assessment practices and their teaching goals. They thought that they had measured their students' higher thinking skills (i.e., integration and application of content); however, their test questions appeared to assess recognition of details and facts. Gullickson and Ellwein (1985), Marso and Pigge (1987), and Oescher and Kirby (1990) further found that teachers did not consistently employ the table of specifications in constructing their test items. This earlier research represents an important stage in the overview of traditional assessment.

Despite the strong preference for traditional methods, there is evidence of increasing use of innovative assessments among classroom teachers. For example, Greenstein (2004) found that teachers employed traditional assessments in assessing students' learning comprising multiple-choice, gap-filling and matching items, as well as innovative assessments including performance-based assessments, projects, journals, self-assessment and portfolios. The findings further revealed there were some variations associated with the use of innovative assessments, with performance-based assessment the most commonly used, whereas the least commonly employed were portfolios and self-assessment. A conclusion from the study was that the majority of teachers used traditional assessment to determine their course grades.

A wealth of studies undertaken within a language education context in both school and higher education settings have also provided evidence that teachers prefer to use objective tests/exams rather than innovative assessment methods in assessing students' learning (Cheng et al., 2004; Rogers et al., 2007; Tsagari, 2008; Xu & Lix,

2009). For instance, Cheng and her colleagues (Cheng et al., 2004; Rogers et al., 2007) found that EFL/ESL university teachers reported using matching, true/false and multiple-choice items as the commonly used assessment formats. In general, these research findings highlighted that most Chinese teachers utilised tests and examinations as the major results (80%) of their classroom assessment practice. Consistent with findings by Cheng and her colleagues, Tsagari (2008) surveyed EFL primary and secondary school teachers in Greece and reported that teachers used little innovative assessment (e.g., self- and peer assessments) and their assessment methods strongly focused on testing vocabulary and grammar. These teachers preferred using objective tests or exams and relied on available print sources.

Hence, the findings documented by the previous body of studies conducted within both language and general education settings have highlighted that teachers predominantly use traditional assessment, such as objective tests and exams, rather than innovative assessment (e.g., performance-based assessments) in assessing students' learning. Teachers' greater preference for objective tests/exams may be due to the perception that this method tends to achieve high reliability of marking and can be fast and easy to score the students' work. However, they show less preference towards innovative assessment such as performance-based, portfolios and self- and peer assessments. And this may be due to the fact that with these methods, there are concerns regarding consistencies or reliability of marking. These methods are also more time-consuming and therefore increase teachers' workload. As such, teachers tend to show a strong preference towards using more traditional assessment than innovative assessment, irrespective of the educational setting.

2.2.5 Interpreting Assessment Outcomes

The third step of the assessment process is to consider how the assessment outcomes of the students' work or performance can be interpreted. Interpreting assessment outcomes refers to the way in which teachers interpret the assessment information collected, using either a norm-referenced or criterion-referenced framework

(Popham & Husek, 1969; Hambleton & Novick, 1973; Haertel, 1985; Reynolds, Livingston, & Willson, 2009; Sadler, 2009a; Miller et al., 2013; Popham, 2014).

Norm-referenced interpretation is associated with comparison of a student's performance to that of others within or across classes and/or schools/universities (Lamprianou & Athanasou, 2009; Reynolds et al., 2009; Sadler, 2009a; Miller et al., 2013; Popham, 2014). That is, the norm-referenced interpretation shows the student's ranking relative to others in a norm group.

In contrast, the criterion-referenced interpretation, first introduced by Glaser (1963), is relevant to the description of student performance that demonstrates his/her specific knowledge and skills against the specified course learning outcomes. That is, with the use of objective tests, criterion-referenced interpretation emphasises the percentage of items/tasks obtained correctly, known as a percentage-correct score (Lamprianou & Athanasou, 2009; Miller et al., 2013).

When interpreting test scores, Griffin, Care, Robertson, Crigan, Awwal, and Pavlovic (2013) caution that interpretation of students' performance should not merely examine the raw score (the score observed by aggregating all the items answered correctly). Raw test scores are linked with the test's content. Students' high score on any test indicates that they have mastered the content of that particular test, but only provides limited information about the students' skills. Such interpretation of the assessment data is meaningless for teaching intervention and for informing students' learning. Research, however, indicates that teachers tend to focus mainly on the raw score when interpreting their assessment data. For example, Hoover and Abrams (2013) administered a web-based survey to 656 elementary, middle and high school teachers in a large, suburban school district in central Virginia to explore the extent to which teachers used summative assessment data in formative ways to enhance instruction. The study indicated that teachers were unable to make use of summative data to inform their instruction, as their interpretation of the assessment data simply emphasised the average score. Such findings highlight Griffin et al.'s (2013) caution.

Given such concerns, Griffin (2007), building on the work of Spearritt (1982), combined the work of Glaser (1963) with that of Rasch (1960) and Vygotsky (1978) to propose a probabilistic interpretation of competence. Under this model, each student's

performance is interpreted in terms of competence levels. The score is simply a code for his/her level of development and helps to indicate the student's zone of proximal development (Vygotsky, 1978). When the zone of proximal development (i.e., the point at which students are most ready to learn) has been identified, there is more likelihood that teaching intervention will have a positive impact on students' learning. Such interpretation can serve formative purposes (i.e., assessment is for teaching) (Care & Griffin, 2009). Criterion-referenced interpretation therefore is seen as beneficial for planning instruction, whereas norm-referenced interpretation is useful for selecting or grouping students based on their relative learning achievements (Lamprianou & Athanasou, 2009; Miller et al., 2013). Norm-referenced interpretation is limited to assessments conducted for summative purposes whereas criterion-referenced interpretation can be applied to summative and formative functions (see Gillis & Griffin, 2008).

A criterion-referenced interpretation can be applied to innovative assessment, such as performance-based assessments, portfolios and self- and peer assessments, largely through the use of analytic and holistic scoring to mark students' work (Lamprianou & Athanasou, 2009; Sadler, 2009a; Miller et al., 2013; Popham, 2014). Analytic scoring methods require teachers to emphasise each different aspect of students' work separately such as content, vocabulary, grammar and organisation within an essay. In contrast, holistic scoring methods are associated with judgment based on the overall quality of students' work (Lamprianou & Athanasou, 2009; Sadler, 2009c; Miller et al., 2013; Popham, 2014). Similarly, both analytic and holistic scoring methods require teachers to use the rubrics when marking students' work. While the analytic scoring method has been purported to provide specific feedback about students' learning (i.e., for diagnostic purposes) more than the holistic scoring method, its limitation has been associated with its time-consuming nature (Lamprianou & Athanasou, 2009; Miller et al., 2013). Furthermore, it restricts itself to preset criteria, without taking into account the overall quality of students' work (Sadler, 2009a). However, the holistic scoring method emphasises the overall quality of students' work, whereas its limitation has been relevant to reliability issues (Lamprianou & Athanasou, 2009; Miller et al., 2013). It has been further suggested that if the holistic scoring method is conducted properly, it can serve as

both an analytic and holistic judgment, as teachers have to pay attention to particular aspects of students' work prior to making judgments on the quality of the work as a whole (Sadler, 2009a). It has also been proposed that teachers tend to make holistic judgments although they employ analytical scoring methods in marking students' work (Bloxham, 2013).

In relation to the scoring method employed, studies undertaken within language and general higher education settings have shown some variations in teachers' assessment practices. For instance, Cheng and her colleagues (Cheng et al., 2004; Cheng & Wang, 2007; Rogers et al., 2007; Cheng et al., 2008) found that Canadian and Hong Kong EFL/ESL university teachers indicated they used analytical scoring, whereas Chinese EFL/ESL university teachers tended to employ holistic scoring methods. Similarly, Bloxham, Boyd, and Orr (2011) employed think-aloud and semi-structured interviews to examine the way in which 12 teachers from two universities in the UK marked their students' assignments within general education. The results showed that teachers used holistic rather than analytical judgments. The majority were influenced by students' efforts and did not make use of rubrics in their grading decisions, but instead used the rubrics to check or justify their holistic judgments after the marks had been awarded.

The use of norm- and criterion-referenced interpretations has faced some criticism in the literature. Norm-referenced interpretation has been criticised as unethical or unfair on students (i.e., consequential validity) because they have no control over other students in a norm group, yet their grades are based on relative position in a norm group (Price, 2005; Orrell, 2008; Sadler, 2009a; Bloxham, 2013). Criterion-referenced interpretation, however, has been largely criticised on the grounds of inconsistency in grading innovative assessment, such as performance-based assessments and portfolio assessments, as teachers tend to utilise their different, unarticulated tacit knowledge (i.e., values and beliefs) in conjunction with the rubrics when judging students' work (Price, 2005; Orrell, 2008; Sadler, 2009a; Bloxham, 2013).

It has been further argued that teachers' use of a criterion-referenced framework cannot be separated from the use of a norm-referenced framework, as teachers tend to be influenced by their tacit knowledge (Orr, 2008; Wiliam, 2008). Teachers tend to compare

students' work with their previous attainments, despite employing explicit rubrics in marking their work or making comparisons between students (Orr, 2008). Orrell (2008) supports Orr's (2008) and Wiliam's (2008) perspectives by pointing out that teachers' interpretations of students' work tend to be largely influenced by their own values and beliefs (i.e., tacit knowledge) rather than focusing solely on the desirable attributes stated in the rubrics. Such arguments reinforce Angoff's (1974) assertion that "if you scratch a criterion-referenced interpretation, you will very likely find a norm-referenced set of assumptions underneath" (p.4).

Research has shown variations in the ways teachers interpret assessment data using criterion- and/or norm-referenced frameworks, as well as the influence of their tacit knowledge in judging students' work within general and language education. For instance, Greenstein (2004) surveyed 115 high school teachers in Connecticut about their assessment practices within general education, following up with semi-structured interviews, and reported that the majority of teachers (64%) created their rubrics for criterion-referenced interpretation. They always employed them when marking students' work, whereas 46% of the teachers sometimes used them. Davison (2004) employed verbal protocols, individual and group interviews and questionnaires to explore how 24 Hong Kong and Australian ESL teachers marked students' work. The findings revealed differences in marking methods. Teachers not only took into consideration the rubrics, but also their tacit knowledge of the students (e.g., efforts). Consistent with Davison's (2004) study, Cooksey, Freebody, and Wyatt-Smith (2007), using a think-aloud protocol to examine 20 Queensland primary school teachers making judgments on 50 different pieces of student writing using a rating scale (i.e., 1= Poor level of achievement to 5= Excellent level of achievement). They found that there was variation regarding teachers' judgements of the same piece of work. Schneider and Gowan (2013) assigned 23 elementary mathematics teachers to one of three conditions: analysing items and student responses without rubrics, analysing items and student responses with rubrics, or analysing items and student responses with rubrics after watching a professional development programme on providing feedback to students. The authors reported that teachers in all three contexts had difficulty interpreting their assessment data.

In summary, criterion-referenced interpretation can be applied to both traditional and innovative assessments and it can serve both summative and formative purposes. Criterion-referencing has greater potential to enhance validity (i.e., content validity) than norm-referenced interpretation, as it can closely align with the course learning objectives/goals. However, it has been criticised on the grounds of consistency (i.e., intra-reliability and/or inter-reliability), given that teachers are more likely to be influenced by their tacit knowledge when judging students' work. Such issues (i.e., consistency) need to be appropriately addressed, in particular when assessments are undertaken primarily for summative purposes.

2.2.6 Grading Decision-making

The next step of the assessment process to be considered is relevant to grading decision-making. Grading decision-making is associated with the way in which teachers summarise a set of individual marks to arrive at a grade regarding overall achievement (O'Connor, 2009; Sadler, 2010; Guskey & Jung, 2013; Quinn, 2013; Waugh & Gronlund, 2013). Sadler (2005) has identified four different grading decision models that have been used worldwide in the higher education sector. The first grading model is "achievement of course objectives". This model requires grades to represent how well the students have progressed towards learning outcomes. The second grading model is "overall achievement as measured by score totals," commonly known as percentage grading. Within this grading model, grades for different assessment tasks are typically tallied to give an overall grade for the course. The third grading model is "grades reflecting patterns of achievement". Although this model is somewhat similar to the previous model, it allows teachers to combine grades in varying ways in order to recognise students' overall learning achievements. The fourth grading model is "specified qualitative criteria or attributes". This model involves specifying qualitative criteria or attributes with respect to grading. That is, the typical university grade descriptors that state the requirements for obtaining a particular grade. In relation to these grading decisions, it has been widely stated within the literature that employing broad categories in grading has the likelihood to increase the relative amount of error present in the measures

on which the grades are based (Ebel, 1969; Cresswell, 1986; Rom, 2011; Bradshaw & Wheeler, 2013).

Out of these models, the second grading model “overall achievement as measured by score totals” remains widely used, as it gives the impression of definiteness and precision and is easy to operationalise (Sadler, 2005, Guskey & Jung, 2013; Quinn, 2013; Waugh & Gronlund, 2013). However, the grade cut-off scores in model two are not generally related to the mastery of specific skills or learning outcomes and it is typically left to the teachers’ decisions. As such, the main concern associated with this model is how the marks are generated in the first place in terms of validity, sampling adequacy, assessment task quality, marking standards and marking reliability (O’Connor, 2009; Sadler, 2010; Guskey & Jung, 2013; Quinn, 2013; Waugh & Gronlund, 2013).

In relation to an “overall achievement as measured by score totals” for grading decision-making, numerous studies have shown that teachers tend to combine the students’ academic learning achievement factors with their non-academic achievement factors when determining their overall course grades within both general and language education (Brookhart, 1993; Cross & Frary, 1999; McMillan & Nash, 2000; McMillan, 2001; Greenstein, 2004; Sun & Cheng, 2013). For instance, Brookhart (1993) surveyed two groups of school teachers about their grading practices. One group of teachers had received assessment training, while the second group had no assessment training. Results indicated that both groups of teachers gave below-average students a passing grade, as they believed these students put in adequate efforts in undertaking the assessment tasks, whereas they did not give extra grades to average or above average students. The study concluded that the teachers did not use students’ academic achievements exclusively in their course grade determination, since they considered students’ motivation, self-esteem and the consequences of giving these grades. In line with Brookhart’s (1993) study, Greenstein (2004) reported that teachers incorporated students’ class participation, effort and attitude into their course grades. Consistent with these findings, Sun and Cheng (2013) administered a questionnaire to 350 junior and senior EFL Chinese teachers to explore their grading practices, and found that they included both academic and non-academic achievement factors in determining their course grades. The researchers further

revealed that teachers awarded more weight to students' non-academic achievement factors comprising effort, study habits and homework.

Hence, such decision-making (i.e., combining both academic and non-academic achievement factors to determine an overall course grade) can raise considerable concerns related to the validity and reliability of the students' actual academic learning achievements, particularly for summative purposes, as they have been distorted by non-academic achievement factors. As such, it appears that assessment records have a vital role to play in dealing with such grading decision-making issues.

2.2.7 Assessment Records

Recording assessment information is the next step of the assessment process after grading decision making. Assessment specialists classify assessment records into three types: anecdotal data (e.g., recording of critical incidents, learning behaviours and reflections of teachers), folio information (e.g., samples of student work), and statistical data (e.g., grades) (Griffin & Nix, 1991; Airasian, 2000; Russell & Airasian, 2012; Witte, 2012; Miller et al., 2013). Assessment records provide important information to the teachers/assessment developers and other relevant stakeholders as evidence for making judgments of students' learning achievements (Griffin & Nix, 1991; O'Connor, 2009; Bachman & Palmer, 2010; Chappuis et al., 2012; Miller et al., 2013). Thus, it is essential for the assessment records to be valid and reliable across various assessment tasks, and across various groups of students and assessment outcomes. Specifically, anecdotal assessment records, such as the teachers' notes on their observation sheets and checklists concerning students' learning behaviours, must be accurate across different groups of students, across different classes and across different times. Inaccurate assessment records have been proposed to provide limited information of students' actual learning achievements and could have led to unfairness to students within a high-stake assessment context (Bachman & Palmer, 2010; Miller et al., 2013).

To obtain accurate anecdotal assessment records, various steps have been recommended (Miller et al., 2013). First, each observed incident should be written down as soon as possible. Second, the descriptions of the observed incident should be recorded

on separate cards or separate pages in a paper notebook for each student. Third, the description of an incident should be separated from any interpretation of the actual behaviour observed. Fourth, the observation should be restricted to specific behaviours that cannot be measured by other assessment methods (e.g., tests/exams). Fifth, limiting the observation of all students at any given time to just a few types of behaviours needed.

Furthermore, assessment records of students' learning achievements should be separated into two distinct information aspects- formative and summative for appropriate purposes (Chappuis et al., 2012). Assessment records of students' learning achievements must also be separated from other non-academic learning factors (e.g., effort and behaviour), given the latter is more difficult to measure and can alter the meaning of the actual grade awarded for students' work (Wormeli, 2006; O'Connor, 2009; Brookhart, 2011b; Chappuis et al., 2012; Miller et al., 2013; Popham, 2014; Schimmer, 2014). Recording of students' learning achievements must accurately reflect their learning outcomes (O'Connor, 2009; Brookhart, 2011b; Chappuis et al., 2012; Sadler, 2009b, 2013a; Miller et al., 2013; Popham, 2014; Schimmer, 2014). Research, however, has shown that teachers tend to record students' academic learning achievements together with their non-academic learning factors (Frary, Cross, & Weber, 1993; Brookhart, 1994; Cizek, Rachor, & Fitzgerald, 1995; Greenstein, 2004).

2.2.8 Assessment Reporting

The last step of the assessment process is reporting. Reporting refers to the communication of the outcomes of student learning to the students themselves and/or other relevant stakeholders. There are three main ways to report assessment results including letter grades, numerical scoring and descriptive feedback (Brookhart, 1999; CANEP, 2002; Linn & Miller, 2005; Lamprianou & Athanasou, 2009; Chappuis et al., 2012; Miller et al., 2013; Popham, 2014; Schimmer, 2014). Izard (2006) and Masters (2013b) recommends reporting students' achievements in terms of their learning progression, such as what they currently know, understand and can do as well as how much progress they have made over a semester and/or school year.

However, research has indicated that teachers tend to report predominantly in terms of grades. For instance, Ruiz-Primo and Li (2013), analysing 26 elementary and middle school teachers' written feedback practices in students' science notebooks, found that 61% of the feedback pieces were grades and only 33% were descriptive comments. The researchers concluded that, of the descriptive comments provided, only 14% of feedback had the potential to help students improve their work.

Studies have also shown that the types of grades to be reported and the feedback provided to students can impact their motivation to learn (Crooks, 1988; Black & Wiliam, 1998a; Harlen & Crick, 2003; Nicol & Macfarlane-Dick, 2006; Timperley, 2013). While reports comprising numerical scores have been thought to be concise and easily computed (Linn & Miller, 2005), their communicative power is limited as:

- it does not provide students with meaningful information on their learning progress (Black & William, 1998a; Brookhart, 1999); and
- it can induce students to focus on performance goals in passing the assessment tasks rather than on mastering learning goals (Dweck, 1999).

Harlen and Crick (2003) have suggested that such reports can have a “negative impact on [learners'] motivation for learning that militates against preparation for lifelong learning” (p.169). Lifelong learning is associated with students' capabilities to apply content knowledge to critical thinking, problem solving and analytical tasks throughout their whole education (Binkley et al., 2012).

In contrast, there is widespread agreement in the literature regarding the benefits of using descriptive reports to communicate students' strengths and weaknesses (Black, 1993; Brookhart, 2008; Chappuis et al., 2012; Earl, 2013; Timperley, 2013; Popham, 2014). This use of descriptive reporting is relevant to the assessment undertaken for formative purposes. Providing the student with descriptive reporting has been shown to have positive impacts (i.e., intended consequential validity) on students' learning as it can foster their lifelong learning capacities (Black & Wiliam, 1998a; Wiliam, 2013). However, for such descriptive reports to be effective in improving students' learning, Sadler (1989) has argued three conditions must be met. Firstly, the student must know what high quality work is. Secondly, s/he must possess the necessary skills to compare

the quality of his/her work with the high standard. Finally, the student must know how to close the gap between his/her current and good performance. Nicol and Macfarlane-Dick (2006, p. 205) have supported Sadler's notion by stating that good feedback practice:

- helps clarify what good performance is;
- facilitates the development of self-assessment in learning;
- delivers high quality information to students about their learning;
- encourages teacher and peer dialogue around learning;
- encourages positive, motivational beliefs and self-esteem;
- provides opportunities to close the gap between current and desired performance; and
- provides information to teachers that can be used to help shape teaching.

Izard and Jeffery (2004) have further asserted that the good feedback practice in the formative assessment process should provide information regarding what each student already knows. In other words, it helps in identifying what has to be taught. Such assessment information has the potential to inform the teacher and the student of what assessment tasks can be attempted successfully, what knowledge and skills are being established currently, and what knowledge and skills are not yet within reach. Having such information can assist teachers emphasise what students need to learn and students demonstrate evidence of learning. As Black and Wiliam (1998a, 1998b) have claimed, employing good feedback practice in formative assessment could both enhance teaching effectiveness and produce student academic achievement gains ranging from .4 to .7 magnitude of effect size.

It has been recommended that feedback on students' work should be devoid of judgmental language as it leaves no room for students' responses to improving their work (Boud & Molloy, 2013). Research has indicated that teachers' reporting feedback on their students' work were more frequently words of encouragement rather than specific comments on students' strengths and weaknesses and on the attainment of learning goals (Greenstein, 2004). It is argued that such reports are meaningless as they do not offer students any descriptive information to help them improve their performance (Izard & Jeffery, 2004; Nicol & Macfarlane-Dick, 2006; Molloy & Boud, 2014). Molloy and Boud

(2013) have also argued for renewing the conception of feedback to acknowledge the agency of students in the feedback process. That is, students serve their dual roles as “feedback generators” and “feedback seekers”, rather than considering them as merely “passive recipients” of comments/feedback from their teachers (Molloy & Boud, 2013, p. 25). Nicol (2013) has further argued for students to be proactive, rather than reactive, in the feedback process. When teachers offer feedback, students should be prompted to not only evaluate the feedback, but also to make a structured response. Students should also be given opportunities to engage in both self- and peer reviewing activities. Carless (2013a), Jolly and Boud (2013), McArthur and Huxham (2013) and Sadler (2013b) have gone even further to argue that feedback should be more dialogic or be a two-way interchange, rather than one-way communication. If students receive feedback for an assessment task that will not be repeated, such feedback is meaningless for both teaching and learning. Such feedback has generally been referred to as one-way communication (Carless, 2013b, Jolly & Boud, 2013; McArthur & Huxham, 2013; Sadler, 2013b) and it tends to be relevant to assessments conducted for summative purposes (Brookhart, 2008). Brookhart (2008), however, has argued that students can benefit from feedback on summative assessments if they are provided with the opportunity to incorporate it in other subsequent, similar assessment tasks. Carless (2013a) has further argued for building trust in dialogic feedback, in which students and teachers value the perspectives of others and respond empathetically, and students can feel free to take risks. To effectively involve students in the feedback process, it has been suggested that students need to be trained in their self-evaluative abilities (Boud & Molloy, 2013; Carless, 2013b) in terms of how to judge quality and modify their work during the production stage through using macro appraisal (i.e., assessing how their work is coming together as a whole) and micro appraisal (i.e., judging work aspects that require further improvement) (Sadler, 2013b).

Reports can be either written or oral. In the school sector, the most common way of reporting is through sending report letters/cards to parents of the students (Chappuis et al., 2012; Miller et al., 2013; Popham, 2014). It is also common within the school sector for face-to-face student-teacher conferences and/or parent-teacher conferences to be held. This type of face-to-face reporting has been thought to provide an important supplement to the written report letters/cards. Under this type of reporting, teachers, students and/or

parents have opportunities to share information about learning progress, discuss their concerns and clarify any misunderstandings in relation to the written report letters/cards (Chappuis et al., 2012; Miller et al., 2013; Popham, 2014).

The communication of students' learning achievements varies according to assessment purposes and interests of the stakeholders (Griffin & Nix, 1991). Within language and general education- both for schools and higher education institutions- key assessment stakeholders are teachers, students, parents, administrators and potential employers (Griffin & Nix, 1991; CANEP, 2002; O'Connor, 2009; Chappuis et al., 2012). Each has their own reporting requirements. The use of assessment data for teachers is to collect information to enhance instruction and to summarise students' achievements in a reportable format for relevant stakeholders. The reporting need for students is to obtain feedback to improve their learning outcomes in the course. Similarly, the reporting need for parents is to receive information regarding students' learning achievements (Griffin & Nix, 1991; Anderson, 2003; Chappuis et al., 2012). In contrast, the reporting need for administrators is for decision-making such as passing/failing students and/or awarding certificates/degrees (Linn & Gronlund, 2000; Rea-Dickins, 2000; Chappuis et al., 2012; Miller et al., 2013). Unlike other stakeholders, the reporting need for the potential employers is to obtain the necessary information in relation to knowledge and skills the graduate possesses to fulfil the position offered (Falchikov & Boud, 2008; O'Connor, 2009).

It has been further proposed that assessment reports need to remain confidential in order to protect the rights of the students, given the requirement of fundamental fairness (Bachman & Palmer, 2010; Miller et al., 2013). That is, releasing students' academic achievements publicly has been regarded as unethical or unfair (Bachman & Palmer, 2010).

The timing of the reporting is also important if assessments are undertaken for formative purposes (Gibbs & Simpson, 2004; Molloy & Boud, 2014). According to Gibbs and Simpson (2004), when students are not provided with immediate feedback regarding their work, particularly for formative purposes, they will move on to new content and feedback received subsequently is therefore irrelevant to their ongoing learning. Molloy and Boud (2014) also add that feedback cannot serve formative

functions if it is provided to students too late (i.e., at the end of semester). Research, however, has indicated that teachers often do not provide students with timely feedback. For instance, Chen, Sok, and Sok, (2007) surveyed 200 Cambodian university teachers from the top six higher education institutions within general education. They found that the majority of teachers (90%) returned the results to students long after the assessments were taken. Such late feedback is unlikely to be useful, as it may not be relevant to students' current learning activities.

In summary, assessment reporting plays an important role in providing information regarding students' learning outcomes to all relevant assessment stakeholders for their decision making. As such, it appears that quality management has a role to ensure high quality assessments prior to reporting assessment outcomes, particularly for summative purposes.

2.2.9 Assessment Quality Management

Assessment quality management needs to be taken into consideration at each stage of the assessment process in order to ensure quality results. Assessment quality management has been defined as the process by which to ensure the reliability, validity, ethicalness or fairness and transparency of the assessment outcomes in order to ensure comparability of standards in conducting assessments across classes and within schools/universities (Dunbar et al., 1991; Gipps, 1994b; Maxwell, 2006; Harlen, 1994, 2007; Gillis et al., 2009). In Australia, a professional code of practice for validation and moderation within the vocational education and training setting states: to achieve national comparability of standards in undertaking assessment, there is a need to consider three main aspects to quality management: quality assurance, quality control and quality review (Gillis et al., 2009). Quality assurance is concerned with the quality of assessment by emphasising the input into the assessment process (e.g., policies and assessment standards). Quality control deals with monitoring and, where necessary, making adjustment to assessor judgements before assessment results are reported (i.e., moderation). In contrast, quality review focuses on the review of the assessment results and processes in order to make recommendations for future improvements (Harlen, 2007;

Gillis et al., 2009; Maxwell, 2010). Thus, assessment quality management occurs at three distinct times including the period before (i.e., quality assurance), during (i.e., quality control) and after the assessment takes place (i.e., quality review). That is, quality assurance is an input approach, quality control is an active approach and quality review is a retrospective approach to managing the quality of the assessment process. As such, assessment quality management is the key factor in achieving high quality assessment processes through the enhancement of accuracy and consistency.

Within school and higher education sectors, implementing procedures and processes to ensure high quality assessments is particularly important prior to reporting assessment outcomes to all relevant stakeholders (via quality assurance and quality control mechanisms), particularly for high-stakes assessment (Maxwell, 2006; Daugherty, 2010). In relation to enhancing the quality of assessments from an input approach (i.e., quality assurance), teachers have the responsibilities to ensure that their assessments reflect the learning goals/outcomes specified in the curriculum (Izard, 1998; Chappuis et al., 2012; Miller et al., 2013; Popham, 2014). With respect to traditional assessment, it requires teachers to establish adequate levels of internal consistency (i.e., coefficient alpha) for tests/exams in order to improve quality (i.e., quality assurance) prior to using them to assess students' learning, particularly for summative purposes (Haertel, 2006; Lamprianou & Athanasou, 2009; Miller et al., 2013). Research, however, has shown that teachers ignore quality assurance in their assessment practices regarding the use of traditional assessment (Gullickson, 1984; Gullickson & Ellwein, 1985; Oescher & Kirby, 1990; Mertler, 2000). With regard to the use of innovative assessment (e.g., performance-based assessments), this requires teachers to develop rubrics in order to enhance consistency when they mark students' work (Brookhart, 2013b; Selke, 2013).

Quality control by means of moderation has been the most common practice to enhance the reliability, validity, fairness and transparency of assessment outcomes within the classroom-based assessment context. Moderation refers to the process for monitoring the quality of assessment and to ensure appropriate assessment procedures have been adhered to, as well as to check on the interpretations of assessment outcomes, particularly how the scoring criteria have been applied consistently in marking students' work (Maxwell, 2007; Lawson & Yorke, 2009; Klenowski & Wyatt-Smith, 2010; Sadler,

2013a; Smaill, 2013). Two types of moderation have been commonly implemented: internal and external (Pennycuick, 1991; Bloxham, 2009; Smith, 2012). Internal moderation is relevant to the process undertaken by teachers during the delivery of their courses to ensure valid, reliable, fair and transparent assessment outcomes are obtained. Internal moderation typically comprises a double marking process, which involves two or more teachers assigning marks to the same piece of student writing using developed marking criteria and standards. Thus, internal moderation is relevant to the use of innovative assessments (e.g., performance-based and portfolio assessments). In contrast, external moderation has been referred to the process conducted by external examiners for the purpose of ensuring that standards used are comparable with similar awards, as well as ensuring regulations and assessment procedures employed by schools and universities are effective and fairly applied (Pennycuick, 1991; Bloxham, 2009; Smith, 2012). As such, external moderation has been argued to be irrelevant to teachers, conducted by expert assessors outside schools and higher education institutions (Bloxham, 2009; Smith, 2012).

Studies, however, have shown that internal moderation practices conducted by school and university teachers within general education tend to be ineffective (Klenowski & Adie, 2009; Wyatt-Smith, Klenowski, & Gunn, 2010; Connolly, Klenowski, & Wyatt-Smith, 2012). For example, Bloxham and Boyd (2012) employed think-aloud and semi-structured interviews to explore 12 teachers' grading practices in two UK universities. The findings highlighted that although moderation had some power to ensure standards within the groups of teachers that adhered to the guidelines for marking criteria and standards, there were still some variations in their judgements. Such variations were due to different individual criteria including strong introductions and conclusions of the written paragraphs/essays, particular source material, the evaluative quality of the work, and sources, content and expression. These findings reveal poor internal moderation practice undertaken by teachers and they raise considerable concerns about accuracy, consistency and fairness associated with the assessment outcomes obtained.

Such ineffective internal moderation practices have been attributed to the focus on by-product (i.e., a corrective activity that compensates for inadequacies or mistakes of the teachers in making their judgments of students' work) and the lack of opportunities for

teachers to challenge and clarify their interpretation and application of standards (Bloxham & Boyd, 2007; Maxwell, 2010; Smaill, 2013). It has been suggested however that internal moderation can be effectively undertaken if moderation focuses on teachers' professional learning processes (Bloxham & Boyd, 2007; Maxwell, 2010; Sadler, 2013a; Smaill, 2013). This process values consensus in terms of facilitating shared understanding of standards and debate about assessment practices.

Quality control also plays a vital role in terms of checking teachers' overall course grade decisions and their assessment record-keeping prior to reporting to relevant stakeholders, particularly for summative purposes. Quality control checks can be undertaken by teachers themselves through having discussions within their community of practice (Lave & Wenger, 1991; Wenger, 1998). Within this community of assessment practice, teachers have opportunities to openly justify their course grade decisions (e.g., including students' non-academic learning factors into their final course grades) and to discuss the way in which assessment results are recorded. Such practice can therefore help teachers to be fully aware of the influence of extraneous factors on their grading decision-making and it can minimise the variation of their grading practices.

2.3 Summary

In summary, the classroom assessment process comprises assessment purposes (i.e., formative versus summative), assessment methods (i.e., traditional versus innovative), interpreting assessment outcomes (i.e., norm-referenced versus criterion-referenced), grading decision-making, assessment records (i.e., anecdotal data and statistical data) and assessment reporting. Within each stage of the assessment process, validity (e.g., content, construct and consequential), reliability (e.g., coefficient alpha and intra-rater and inter-rater) characteristics and assessment quality management (e.g., quality assurance and quality control) have to be addressed, particularly for summative purposes in which the stakes are high for the students. Hence, all stages within the assessment process are interrelated rather than independent, yet they are sequential. That is, the assessment purposes are inextricably linked to the selection of assessment methods used. The choices of methods selected are then connected with the way in which the assessment outcomes are interpreted, grading decisions are made, results recorded, and

outcomes reported. The extent to which the reliability and validity characteristics and the assessment quality management are taken into consideration within each stage of the assessment process also depends on the purpose(s) of assessments. When classroom assessment is employed for formative purposes, the reliability and validity characteristics and the assessment quality management within each stage of the assessment process tend to be under less scrutiny. In contrast, when the assessment is undertaken for summative purposes, enhancing the validity and reliability characteristics and ensuring the quality within each stage in the assessment process becomes important as such assessments can have significant consequences on the academic lives of students.

As an outcome of the literature review, it appears that there is a gap between the actual classroom assessment practices implemented by teachers and the ideal theoretical underpinnings of the assessment process. Overall, there are considerable concerns with teachers' assessment practices in relation to each key stage within the assessment process, within both language and general education. The main concerns in relation to assessment practices amongst teachers across school and university settings are the tendency for teachers to:

- ignore the validity and reliability implications of their assessments;
- design tasks that assess students' learning at lower level thinking skills;
- employ traditional assessment methods like tests/exams and little innovative assessment, such as performance-based assessments, portfolio and self- and peer assessments in assessing students' learning;
- be inconsistent in their judgments of students' work (i.e., creating intra-rater reliability and inter-rater reliability concerns);
- be influenced by students' various non-academic learning factors (e.g., effort and attitude) when marking their work;
- combine students' academic learning achievement factors with their non-academic learning factors (e.g., effort), to determine an overall course grade based on an overall achievement, as measured by score totals commonly known as percentage grading;
- not maintain accurate records of the assessment process and outcomes;

- report academic learning achievements to students mainly in terms of numerical grades rather than descriptive feedback, as well as provide students with late feedback; and
- ignore quality assurance and use ineffective quality control.

These assessment practices can be interpreted as evidence of limited classroom assessment literacy of the teachers. Hence, teachers' classroom assessment literacy can be argued to have significant impacts and/or consequences on the quality of teaching and learning. Research focusing on classroom assessment literacy of teachers is warranted in order to better comprehend the nature of their assessment literacy, so that appropriate remedies can be used to address these issues in time.

The next chapter explores the factors that underpin classroom assessment literacy and how such determinants impact on teachers' actual assessment practices.

Chapter 3: Classroom Assessment Literacy

This chapter firstly presents the theoretical framework underpinning the study. Secondly, it explores the key aspects associated with concepts of literacy, followed by definitions of assessment literacy previously discussed in the literature. Thirdly, it examines the underpinnings of classroom assessment literacy (i.e., assessment knowledge base and assessment beliefs) and how such determinants impact on assessment practices within classroom-based settings. In the following sections, the term “language education” has been specifically referred to language learning achievements (English) while the term “general education” is associated with a variety of domains such as biology, chemistry, history and so forth within classroom-based settings. The term “teacher” is also used throughout this chapter to refer to both school teacher and/or tertiary instructor.

3.1 Theoretical Framework

The present study positions itself within agentic theory (Davidson, 1963, 2001; Mayr, 2011; Kögler, 2012), reasoned action theory (Fishbein & Ajzen, 1975, 2010), planned behaviours theory (Ajzen, 1991, 2005) and social cognitive theory (Bandura, 1986, 1997) in examining the underpinnings of teachers’ classroom assessment literacy.

Within agentic theory (Davidson, 1963, 2001; Mayr, 2011; Kögler, 2012), individual teachers are considered to be vital and powerful agents of their own behaviours/performances. In other words, teachers are the key actors and contributors to the success and effectiveness of conducting classroom assessment. As such, teachers play a crucial role in implementing high quality assessments.

Within reasoned action (Fishbein & Ajzen, 1975, 2010) and planned behaviours theories (Ajzen, 1991, 2005), personal beliefs or attitudes of the individual teachers have been argued as a good predictor of their behaviours/ performances. That is, personal beliefs or attitudes of teachers can powerfully influence the ways in which they implement their classroom assessment. Fishbein and Ajzen (1975, 2010) defined personal beliefs/attitudes as the representation of information that individuals hold about any object, thing and other people surrounding them. Green (1971), Rokeach (1972) and

Lamont (2013) defined personal beliefs broadly as propositional attitudes, which refer to the attitudes of individuals toward a proposition about any object, thing and other people. Dasgupta (2013) also asserted that personal beliefs or attitudes could powerfully influence the individuals' judgments, decisions and actions without intention. Thus, teachers' personal beliefs about assessment can have a powerful impact on their assessment implementation.

Within social cognitive theory (Bandura, 1986, 1997), self-efficacy has been argued to influence every phase of the individuals's personal evolution by regulating their behaviours/performances through their cognitive, motivational, affective and decisional processes (Zimmerman, 1995; Bandura, 1997; Schunk & Pajares, 2004; Zimmerman & Cleary, 2006). Self-efficacy determines whether individual teachers think of altering their assessment implementation, and whether they have the motivation and perseverance needed to succeed with their assessment implementation. Bandura (1997) defined self-efficacy as the individual's perceived ability for performing tasks. Bandura (1997) identified four sources of self-efficacy including mastery experience, vicarious experience, social persuasions, and physiological and emotional states. Mastery experience is associated with the notion that success in performing tasks can raise self-efficacy of individuals, whereas failure can lower it. Vicarious experience is concerned with observing the success or failure of peers or models in performing the tasks and these observations can have a powerful influence on the individual's subsequent task performance. Social persuasions are related to verbal messages that individuals receive from others. And these messages can help them to exert the extra effort and persistence required to succeed with tasks. Physiological and emotional states are associated with individuals judging their capabilities through interpreting their stress reactions and tension as a contribution to task performance. Of the four sources of self-efficacy, mastery experience and vicarious experience have been indicated as the most powerful in instilling a sense of efficacy. Given people tend to either overestimate or underestimate their actual abilities (Zwozdiak-Myers, 2012), it is important for them to be aware of their own self-efficacy. It has been suggested that such self-awareness can influence their task performances, the effort they exert in those performances, and the extent to which they use their knowledge and/or skills in performing such tasks (Bandura, 1997). Individual

teacher's self-efficacy therefore plays an important role in his/her assessment implementation.

Theories of social cognitive, reasoned action and planned behaviours, however, have acknowledged that lack of requisite knowledge and/or skills of any individuals can prevent them from carrying out their intended tasks (Bandura, 1989, 1997; Ajzen, 1991, 2005; Fishbein & Ajzen, 1975, 2010). When individual teachers lack knowledge and/or skills to implement the assessment, it is more likely they have little or no intentions of carrying out the intended assessment, even though they have favourable personal beliefs/attitudes. Both the individual teachers' knowledge/skills and personal beliefs/attitudes are therefore considered as the underpinnings of classroom assessment literacy that can enable them to implement assessments effectively and successfully.

Hence, the current study hypothesised that classroom assessment knowledge base and personal beliefs about assessment are important facets to reflect teachers' classroom assessment literacy, given the study positioned itself within agentic, social cognitive, reasoned action and planned behaviours theories. The concepts of literacy in general and definitions of assessment literacy will be considered next.

3.2 Concepts of Literacy

Given its prominent role, the concept of literacy has been continually evolving over decades. As such, there does not appear to be any explicit single definition with regard to the concept of literacy (Collins, 1995; Brockmeier & Olson, 2009; Olson, 2009; Street, 2009; Wagner, 2009). Consequently, a few definitions associated with the concept of literacy have been proposed by a number of researchers. For example, Olson (2009) has offered three distinct definitions of literacy. First, literacy has been commonly defined as an individual's ability to read and write text, which is known as literacy as a basic personal competence. Second, literacy has been referred to as an individual's capability to handle and/or engage with a variety of texts in order to unpack meanings, authority and identity embedded in the texts, namely linguistic or academic literacy. Third, literacy has been associated with an individual's capacity to carry out various activities based on explicit rules, norms and formal procedures of the society that have been written down in documents and/or manuals, known as societal literacy. Wagner

(2009) goes further to expand the notion of literacy by calling for information and communication technology (ICT) literacy. ICT literacy has been referred to as individuals' abilities in using technology for communicating and accomplishing tasks/activities in their everyday lives and/or workplaces. To continually broaden the concept of literacy, Street (2009) has proposed another type of literacy, namely cultural literacy. Cultural literacy has been associated with individuals' capabilities to apprehend the task/activity practiced in their social and cultural settings. That is, the culture that is associated with the individuals' beliefs, ways of thinking, behaving and remembering shared by members of their community (Nostrand, 1989; Kramsch, 1995; Tudge, Doucet, Odero, Sperb, Piccinni, & Lopes, 2006).

Hence, there appear to be various literacy competencies in the literature (e.g., linguistic, technological and cultural literacies). As Brockmeier and Olson (2009) have urged:

...it is necessary to abandon the notion of a single competence that we may think of as literacy and to embrace the broad range of particular competencies and practices that, in turn, may be analysed in linguistic, cognitive, semiotic, technological, and cultural terms (p. 5).

3.2.1 Definitions of Assessment Literacy

The concept of literacy has quickly spread to other fields. With regard to educational measurement and assessment, the notion of assessment literacy has been introduced firstly into general education and eventually being actioned as a component of language education. As with the concept of literacy discussed above, the notion of assessment literacy has also evolved over time, owing to an increasing interest amongst researchers to continue to explore this concept in the field. In line with the concept of literacy in general, there is still no unique definition in relation to the notion of assessment literacy (Taylor, 2009; Walters, 2010; Fulcher, 2012). As such, numerous definitions have been offered by various researchers. For instance, within the general education literature, Richard Stiggins (1991a, 1995) was the first educator who coined the term "assessment literacy" of classroom teachers to describe whether they know the

difference between sound and unsound assessment. Stiggins has added that assessment-literate teachers know:

- what they are measuring and why they are measuring it (i.e., the construct and the purpose of assessment);
- how to measure the knowledge/skill of interest by adequately designing and administering tasks to sample students' performance; and
- what may go wrong with the assessment and how to prevent that from occurring.

Stiggins has further argued that assessment-literate teachers are also aware of the potential harmful consequences of inaccurate assessment data. Similarly, Popham (2006, 2009) defined assessment literacy as the teachers' understanding of fundamental assessment-related principles and procedures that can impact their educational decision making. That is, assessment-literate teachers know the way to develop and/or select more suitable assessment tasks, employ a variety of assessment methods and interpret accurate assessment data, as well as know how to deal with any bias that may creep into their self-made assessment tasks.

Within the language education literature, however, the concept of assessment literacy had just begun to emerge (Inbar-Lourie, 2008a; Malone, 2008) and therefore was still in its infancy (Fulcher, 2012). Recently, there was a call for the inclusion of "assessment beliefs" into the notion of "assessment literacy" (Scarino, 2013). Such a call highlights an acknowledgement of the important role of personal beliefs about assessment in addition to the "assessment knowledge base", previously defined within the general education literature. For example, Davies (2008) and Fulcher (2012) defined assessment literacy as EFL/ESL teachers' acquaintance with theoretical knowledge, practical skills, and understanding of assessment related principles and procedures. Malone (2013) specifically defined assessment literacy as teachers' familiarity with testing definitions and the application of this knowledge to their classroom practices. Inbar-Lourie (2008a) offered further definitions of assessment literacy as teachers' capabilities to apprehend the social role of assessment and the nature of language knowledge in relation to assessment practices. Scarino (2013) went as far as to call for

the inclusion of teachers' personal beliefs about assessment into the notion of assessment literacy. Scarino's call echoes Wolf et al.'s (1991), Inbar-Lourie's (2008b) and Shepard's (2000, 2013) notion of assessment culture, in which personal, espoused beliefs of the teachers are recognised as crucial reflections of their actual assessment practices, given they are key agents in the assessment process (Rea-Dickins, 2004; Scott, 2007; Klenowski, 2013a). As Wyatt-Smith and Klenowski (2013) pointed out, social and cultural practice was inherent in teachers' assessment practices. Hence, there is growing evidence to suggest that teachers' personal beliefs about assessment are as important as their assessment knowledge base, and such determinants can enable and empower them to implement high quality assessments.

In summary, the notion of assessment literacy has continually evolved over time in the literature. Within the general education field, the common notion of assessment literacy has been associated with teachers' knowledge base and/or skills with regard to key stages for conducting assessments (refer to Chapter 2 for a detailed discussion). However, within the language education field, there is a call for an expanding definition of the concept of assessment literacy in terms of including personal beliefs about assessment. The following section reports the body of exploratory studies in relation to assessment literacy in terms of its underlying knowledge base and belief structures.

3.3 Research on Assessment Literacy

This section presents the body of exploratory research focused on the assessment knowledge base and personal beliefs about assessment that is thought to underpin classroom assessment literacy of teachers. It also discusses the research, which is still in its infancy stage, with regard to the relationship between teachers' classroom assessment literacy and assessment implementation.

3.3.1 Assessment Knowledge Base

The first underpinning classroom assessment literacy of teachers to be considered is their assessment knowledge base. Individual teachers' knowledge base has been acknowledged as a vital aspect that contributes to the success and effectiveness of

assessment implementation (Bandura, 1997; Ajzen, 1991, 2005; Fishbein & Ajzen, 2010). It has also been widely reported that the quality of assessment has a relationship with the quality of instruction and students' learning (Tang, 1994; Boud, 2006; Biggs & Tang, 2007; Joughin, 2009; Earl, 2013). As such, there is a need for teachers to have a sound assessment knowledge base to enable them to implement high quality assessments. Assessment specialists have argued that the greater the assessment knowledge base teachers possess, the more capable they are of implementing quality assessments to enhance instruction and student learning (Stiggins, 1991a, 1995; Popham, 2006, 2009). Possessing an adequate assessment knowledge base can help teachers to have a better understanding of the process for conducting classroom assessment (refer to Chapter 2 for a detailed discussion about each key stage of the assessment process). Such a knowledge base will equip them with an appreciation of the assessment process. That is, this knowledge base can enable them to deeply engage with the assessment process, to make an informed choice about the skill and knowledge domains to be assessed, to design/select the appropriate task and be fully aware of the rationale behind it (Price et al., 2012). This knowledge base will also enable them to interpret assessment data accurately, and to know how to deal with any bias that may creep into the self-made or selected assessment tasks (Stiggins, 1991a, 1995; Popham, 2006, 2009). Unfortunately, research has repeatedly shown there is an insufficient level of the assessment knowledge base for teachers, spread across a range of schools within many countries around the world and this has been the case for over five decades (Mayo, 1967; Plake, 1993; Davidheiser, 2013; Gotch & French, 2013). The following sections will explore a range of studies employing self-reported measures and test instruments to examine the assessment knowledge base of school teachers.

3.3.1.1 Self-reported Measures

An individual's perceived capabilities to undertake a task has been proposed to influence his/her expended effort, persistence, motivation and confidence (Bandura, 1997) and as such, it is important for teachers to be able to accurately identify their strengths and weaknesses in assessment. Within the language education literature, there is

a dearth of studies that have been carried out to examine school teachers' self-reported measures of assessment expertise and professional development needs. Of the few studies that have been undertaken, all have reported the tendency for teachers' to identify the need for further training and development of expertise in assessment knowledge and understanding, as well as technical skills. For example, Hasselgreen, Carlsen, and Helness (2004) and Huhta, Hirvala, and Banerjee (2005) administered a questionnaire to European EFL teachers about their language testing and assessment needs. Such findings revealed that teachers indicated the need for more assessment training in various areas. They included portfolio assessment, preparing classroom tests, peer and self-assessments, interpreting test results, giving feedback on students' work, validity, reliability, item writing and statistical analyses. López Mendoza and Bernal Arandia (2009) administered an online qualitative survey using a series of open-ended questions to 82 EFL teachers in Columbia. They found that teachers reported their assessment knowledge base was limited in numerous areas including an understanding of assessment purposes (i.e., formative versus summative); knowledge of different types of assessment methods and what information each type provided; how to give more effective feedback to students; how to empower students to take charge of their learning; ethical issues with respect to assessment use and how results were used; and the concepts of validity and reliability. Similarly, Guerin (2010) surveyed 100 foreign language teachers about their assessment knowledge and perceived needs and found that the majority considered themselves to have limited assessment knowledge base. They indicated a range of assessment training needs regarding the knowledge of concepts and content associated with language testing and assessment. Consistent with such findings, Fulcher (2012), employed an online survey to elicit assessment training needs from 278 language teachers. Fulcher found that teachers who reported their assessment knowledge was poor tended to indicate their needs for training in test design and development slightly lower than in large-scale standardised testing, classroom testing, washback, validity and reliability. The researcher concluded that all teachers were aware of their current insufficient assessment knowledge base.

Within general education, there have been a number of studies that have examined teachers' perceptions of their assessment competence using self-reported measures. The broad term "competence" has been used here to capture the underpinning knowledge and skills required by teachers to conduct educational assessments (Gillis, 2003). Such studies have been summarised in Table 3.1.

Table 3.1 Summary of Studies Examining Teacher Assessment Competence Using Self-reported Measures

Study	Name of measure	Developed	Adapted	Sub-scale	No. of items	Item format	Target population	Sample size	Reliability estimates
Zhang (1996)	Assessment Practice Inventory (API)		☑	2	67	5-point rating scale (Skill scale: 1= not at all skilled to 5= very skilled; Use scale: 1= not at all skilled to 5= very skilled)	school teachers	311	.97
Schaff (2006)	Classroom Assessment Practices Inventory	☑			60	4-point rating scale (1= strongly disagree to 4= strongly agree)	school teachers	117	.89
Chapman (2008)	Assessment Efficacy	☑			14	6-point rating scale (1= not at all confident to 6= Very confident)	school teachers	61	.91
Alkharusi et al. (2011)	API		☑		50	5-point rating scale (Skill scale: 1= not at all skilled to 5= very skilled)	school teachers	233	.95
Alkharusi et al. (2012)	Self-confidence Scale in Educational Measurement		☑	6	54	5-point rating scale (1= very low competence to 5= very high competence)	school teachers	165	.93

Of the four self-reported measures presented in Table 3.1, two (i.e., Assessment Practice Inventory and Classroom Assessment Practices Inventory) were developed to align with the seven standards for Teacher Competence in Educational Assessment of Students jointly issued by the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) (AFT, NCME, & NEA, 1990). These seven standards included:

1. Choosing Appropriate Assessment Methods;
2. Developing Assessment Methods;
3. Administering, Scoring, and Interpreting Assessment Results;
4. Using Assessment Results for Decision Making;
5. Developing Valid Grading Procedures;
6. Communicating Assessment Results; and
7. Recognising Unethical Assessment Practices.

It should be noted that these seven standards had been criticised due to their narrow coverage in relation to classroom-based assessment aspects typically encountered by teachers in their daily instruction (Schafer, 1991; Stiggins, 1995, 1999; Arter, 1999; Brookkhart, 2011a). Despite these standards covering some vital stages in the assessment process encountered by classroom teachers, they did not address two crucial stages of the process, namely keeping accurate records of assessment data and managing quality assurance of the assessment process (refer to sections 2.2.7 and 2.2.9 in Chapter 2 for a detailed discussion). Each study presented in Table 3.1 will be considered next.

A previous study conducted by Zhang (1996) aligned its self-reported measure with the seven standards for Teacher Competence in Educational Assessment of Students. The researcher utilised and calibrated a self-reported survey (i.e., the Assessment Practice Inventory originally developed by Zhang & Burry-Stock, 1994) to examine school teachers' self-perceptions of their assessment expertise in Alabama (see Table 3.1). The findings from the Rasch analyses revealed that teachers found Using Assessment Results for Decision Making the most difficult competency to perform, whereas they found Communicating Assessment Results the easiest.

Another study that also utilised self-reported measures based on the seven standards for Teacher Competence in Educational Assessment of Students was conducted by Schaff (2006). Schaff (2006) designed and calibrated a self-reported survey of the assessment related practices of elementary school teachers from seven school districts in Illinois. The findings from the Rasch analyses showed that items addressing Administering, Scoring, and Interpreting Assessment Results and Using Assessment Results for Decision Making were the easiest items for teachers to agree with. In contrast, items addressing Choosing Appropriate Assessment Methods and Developing Valid Grading Procedures were the most difficult items for teachers to agree with. This result was contrary to Zhang's (1996) finding.

Several studies compared the results of teachers' self-reported measures of competence in educational assessment with their assessment knowledge test scores (see Table 3.1 for a summary of self-reported measures and Table 3.2 for a summary of test instruments). For example, Chapman (2008) reported that all teachers (61) perceived themselves to be confident and skilful in making appropriate educational decisions for assessment data, despite the fact that less than two-thirds obtained correct answers to 70% of the assessment knowledge test. Similar to Chapman's (2008) finding, Alkharusi, Kazem, and Al-Musawai (2011) and Alkharusi, Aldhafri, Alnabhani, and Alkalbani (2012) found that teachers perceived themselves to be skilful in educational assessment, whereas their assessment knowledge test scores demonstrated that their assessment knowledge base was quite low. These findings indicated that teachers' assessment knowledge base, as measured by the self-reported instrument was less accurate than the test instrument used.

As such, the findings of teachers' assessment knowledge base through the use of self-reported measures should be treated with caution, given research showed that such measures tended to be inaccurate. These findings were consistent with previous research analysing the accuracy of using self-reported measures. This indicated that respondents tended to over-report socially desirable behaviours and under-report socially undesirable behaviours in order to prevent them from any perceived embarrassment or possible consequences in future (Tourangeau & Yan, 2007; Holbrook & Krosnick, 2010). This issue of over-reporting socially desirable behaviours and under-reporting socially

undesirable behaviours in any self-reported measures has been referred to as social desirability response bias (Tourangeau & Yan, 2007; Marsden & Wright, 2010; Mitchell & Jolley, 2013). The theoretical perspectives from psychologists Barry Schlenker and Michael Weigold (1989) and Barry Schlenker (2012) and sociologist Erving Goffman (1959) have argued that individuals tend to influence the way in which they are perceived by others in order to pursue the goals of social interaction. Being perceived favourably by others promotes the individuals' views that they may experience an increase in rewards and a reduction in punishments. These perceptions therefore motivate individuals to convey images of themselves to look much better than they actually are.

3.3.1.2 Objective Measures

Given that research has shown that self-reported measures of teachers' assessment expertise provides inaccurate information of their actual assessment knowledge, there has been a push to develop objective test instruments (using multiple-choice items) to directly measure the levels of their assessment knowledge base within the general educational literature. To uncover the levels of assessment knowledge base of school teachers, a number of studies developed and/or adapted assessment knowledge tests to directly measure teachers' assessment knowledge. Table 3.2 provides a summary of studies that employed objective assessment knowledge tests.

Table 3.2 Summary of Studies that Used Assessment Knowledge Tests to Measure Teacher Assessment Knowledge Base

Study	Test name	Developed	Adapted	Sub-scales	No. of items	Item format	Target population	Sample size	Reliability estimates
King (2010)	Criterion-referenced Assessment Literacy Test	☑		3	24	MCQs	school teachers & administrators	352 (teachers) & 28 (administrators)	.73
Gotch & French (2013)	Measurement Knowledge Test	☑		3	20	MCQs	school teachers	650	.47
Plake (1993)	Teacher Assessment Literacy Questionnaire (TALQ)	☑		7	35	MCQs	school teachers	555	.54
Quilter & Gallini (2000)	TALQ		☑	7	21	MCQs	school teachers	117	.50
Chapman (2008)	TALQ		☑	7	16	MCQs	school teachers	61	.54
Alkharusi et al. (2012)	TALQ		☑	7	32	MCQs	school teachers	165	.62
Mertler (2003)	Renaming the TALQ instrument into Classroom Assessment Literacy Inventory (CALI)		☑	7	35	MCQs	in-service & pre-service teachers	197 (in-service teachers) & 67 (pre-service teachers)	.57 (in-service teachers) & .74 (pre-service teachers)
Mertler (2005)	CALI		☑	7	35	MCQs	in-service & pre-service teachers	101 (in-service teachers) & 67 (pre-service teachers)	.44 (in-service teachers) & .74 (pre-service teachers)
Alkharusi et al. (2011)	TALQ		☑	7	35	MCQs	in-service & pre-service teachers	233 (in-service teachers) & 279 (pre-service teachers)	.78 (both groups of teachers)
Mertler & Campbell (2005)	Assessment Literacy Inventory (ALI)	☑		7	35	MCQs	pre-service teachers	249	.74
Davidheiser (2013)	ALI		☑	7	35	MCQs	school teachers	102	.82

Table 3.2 above showed that several studies developed objective multiple-choice tests (MCQs) to measure the assessment knowledge base of school teachers, while others adapted existing tests in their investigations. Of these tests, the Criterion-referenced Assessment Literacy Test was designed to measure teachers' knowledge base in relation to norm-referenced and criterion-referenced tests, validity and reliability and misuses of assessment data. The Measurement Knowledge Test, however, was limited to norm-referencing tests and focused on interpretation of standardised scores, scores in relation to one another within a student, across students, and across schools and proficiency level interrelation. In contrast, the two remaining tests, namely the Teacher Assessment Literacy Questionnaire (TALQ) and Assessment Literacy Inventory (ALI), were developed to align with the seven standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990) and focused on criterion-referencing testing.

Overall, studies that developed and/or adapted tests to directly measure teachers' assessment knowledge highlighted that teachers demonstrated limited assessment knowledge for implementing high quality assessments. It should be noted, however, that two tests (i.e., Measurement Knowledge Test and Teacher Assessment Literacy Questionnaire) had rather low internal consistency reliabilities (see Table 3.2). Each study in Table 3.2 will be considered next.

There were two studies that designed multiple choice tests to measure school teachers' assessment knowledge base. For example, King (2010) developed a Criterion-referenced Assessment Literacy test to measure the assessment knowledge base of school teachers and administrators from the states of Alabama and Mississippi in the United States of America (USA). The results revealed that teachers and administrators correctly answered 47% of the questions regarding the theoretical differences between norm-referenced and criterion-referenced tests, 59% of the questions concerning the concepts of reliability and validity, and 67% of the questions in relation to the potential misuses of assessment data. The study concluded that teachers and administrators had an insufficient assessment knowledge base in the areas of norm-referenced and criterion-referenced tests, validity and reliability and misuses of assessment data. In line with King's (2010) study, Gotch and French (2013) developed a Measurement Knowledge Test and

administered it to elementary school teachers in Washington State. The findings indicated that teachers demonstrated their highest performance in using cut-scores, understanding the concept of the median as a measure of central tendency and interpreting percentile ranks. However, they showed their lowest performance in dealing with score reliability or evaluating properties of assessments to make informed decisions. The researchers concluded that teachers had a limited assessment knowledge base to implement their assessments.

In addition to the two tests developed in the abovementioned studies, an additional study focused on the design of a multiple choice test to measure the assessment knowledge base of school teachers identified in the seven standards issued by AFT, NCME, and NEA (AFT, NCME, & NEA, 1990) (see Table 3.2). Plake (1993) developed a multiple-choice test titled “Teacher Assessment Literacy Questionnaire (TALQ)” to measure the assessment knowledge base of school teachers. Through its content validation process, the instrument had been reviewed by 20 members of the National Council on Measurement in Education (NCME), and a pilot study with and feedback from 70 teachers and 900 educational professionals. The findings showed that teachers performed highest on Administering, Scoring, and Interpreting Assessment Results whereas they performed lowest on Communicating Assessment Results.

A range of studies adapted Plake’s (1993) Teacher Assessment Literacy Questionnaire to continue to measure school teachers’ assessment knowledge base in different settings (Quilter & Gallini, 2000; Mertler, 2003, 2005; Chapman, 2008; Alkharusi et al., 2011; Alkharusi et al., 2012). These studies reported both similar and different findings in terms of sub-scale score and average score to the original research. The inconsistencies of the findings amongst these studies appeared to be related to various factors such as the researchers adapting fewer items from the original multiple-choice test, using different target populations and sample size and low internal consistency reliability (see Table 3.2). For example, Mertler (2003, 2005) found that, on average, pre-service teachers answered slightly less than 19 of the 35 items correctly, while in-service teachers answered slightly less than 22 out of the 35 items correctly. Pre-service teachers’ best performance was on Choosing Appropriate Assessment Methods and their poorest performance was on Developing Valid Grading Procedures. In contrast,

in-service teachers performed highest on Administering, Scoring, and Interpreting Assessment Results, confirming Plake's (1993) finding, whereas their lowest performance was on Developing Valid Grading Procedures which was contrary to Quilter and Gallini's (2000) finding. Quilter and Gallini (2000) reported that teachers' lowest performance was on Developing Assessment Methods and their highest performance was on Developing Valid Grading Procedures. Alkharusi et al. (2011) examined the assessment knowledge base of both pre-service teachers enrolled at the College of Education at Sultan Qaboos University and in-service school teachers in Oman. The study found that in-service teachers had a lower level of assessment knowledge than pre-service teachers. On average, in-service teachers scored 12.55 whereas pre-service teachers scored 15.30 out of the 35 items, which was inconsistent with Mertler's (2003, 2005) findings.

However, two of the studies in Table 3.2 reported their findings differently from the other studies in terms of score range and technical aspects of assessment, such as reliability and percentile rank, leading to incompatibility of their findings with others. For example, Chapman (2008) measured the assessment knowledge base of school teachers in Western Massachusetts in the US and reported that teachers achieved correct answers ranging from 5 to 15. Teachers' lowest assessment knowledge was found in relation to reliability, standardised test, percentile rank, grade equivalent and criterion-referenced information. Alkharusi et al. (2012) measured the assessment knowledge base of school teachers who taught various subjects in the Sultanate of Oman and reported that teachers had a low level of assessment knowledge, as indicated by their assessment knowledge test score, varying from 3 to 21, with an average of 12.42.

It appeared that the 35 multiple-choice item test of the Teacher Assessment Literacy Questionnaire (Plake, 1993) and its various adaptations, had rather weak internal consistency reliabilities for in-service teachers, despite the measure achieving moderate internal consistency reliabilities for pre-service teachers (see Table 3.2). This suggests that the 35 multiple-choice item test of Teacher Assessment Literacy Questionnaire should be limited to pre-service teachers only and any conclusions drawn about differences between in-service teachers and pre-service teachers should be treated with caution.

Given the low internal consistency reliabilities (i.e., Cronbach's alpha) reported on the 35 multiple-choice item test of the Teacher Assessment Literacy Questionnaire (Plake, 1993) from a number of studies including Plake (1993), Quilter and Gallini (2000), Mertler (2003, 2005), Chapman (2008) and Alkharusi et al. (2012) (see Table 3.2), Campbell and Mertler (2004) and Mertler and Campbell (2005) developed a new "Assessment Literacy Inventory (ALI)". The ALI comprised 35 multiple-choice items that paralleled the seven standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990). Unlike the test instrument in Plake's (1993) study, the ALI was thought to be a user friendly format because the seven items related to a single scenario for a total of five scenarios. The use of these five scenarios was thought to reduce cognitive overload in relation to reading 35 unrelated items which appeared in Plake's (1993) test instrument. Mertler and Campbell (2005) administered the ALI to pre-service teachers and the researchers concluded that the ALI functioned reasonably well from a psychometric perspective. They called for a study to be carried out with in-service teachers completing the ALI in order to ascertain its appropriateness as a measure of assessment knowledge base.

In response to Mertler and Campbell's call, Davidheiser (2013) adapted the 35 multiple-choice item test from Mertler and Campbell's (2005) Assessment Literacy Inventory (ALI) and administered it via online to high school teachers in the Central Bucks School District in the US to measure the levels of their classroom assessment knowledge base. The results indicated that, on average, teachers scored 24 out of the 35 items correctly, which was similar to Mertler's (2003, 2005) findings. Their lowest performance was on Developing Assessment Methods, which was consistent with Quilter and Gallini's (2000) study. Their highest performance, however, was on Recognising Unethical Assessment Practices, which was contrary to previous studies of teachers' assessment knowledge base using the seven standards for Teacher Competence in Educational Assessment of Students. Such inconsistency may have been due to Davidheiser's (2013) study, which utilised a different test, namely Assessment Literacy Inventory, as opposed to previous studies that used the Teacher Assessment Literacy Questionnaire.

Hence, there were four tests developed to directly measure the assessment knowledge base of in-service and pre-service school teachers within the general education literature (e.g., Criterion-referenced Assessment Literacy Test, Measurement Knowledge Test, Teacher Assessment Literacy Questionnaire and Assessment Literacy Inventory). Whilst the “Criterion-referenced Assessment Literacy Test” and the “Measurement Knowledge Test” were the measures of the teacher assessment knowledge base in the areas of norm-referenced and criterion-reference tests, standardised tests and educational measurement concepts, the “Teacher Assessment Literacy Questionnaire” and “Assessment Literacy Inventory” were the measures of Teacher Competence in Educational Assessment of Students issued by the AFT, NCME, and NEA in 1990. Overall, although there were some contradicting findings in relation to teachers’ high and low performance identified for each of the seven standards and/or other areas, there was similarity in a key finding regarding the insufficient assessment knowledge base of the school teachers in implementing their assessments.

In summary, extensive research has been undertaken within the general education domain to examine school teachers’ assessment knowledge base through employing self-reported measures and test instruments. Despite the fact that studies using self-reported measures demonstrated high internal consistency reliabilities, these studies faced certain methodological limitations. Research showed that the self-reported measures of teachers’ assessment knowledge base were associated with more inaccurate results than the objective assessment knowledge tests, given that teachers tended to rate their knowledge base higher (Chapman, 2008; Alkharusi et al., 2011; Alkharusi et al., 2012). To avoid biased results with self-reported measures of teachers’ assessment knowledge base, other studies developed and/or adapted test instruments to directly measure the school teachers’ assessment knowledge base. For example, the:

- Criterion-referenced Assessment Literacy Test (King, 2010);
- Measurement Knowledge Test (Gotch & French, 2013);
- Teacher Assessment Literacy Questionnaire (Plake, 1993; Quilter & Gallini, 2000; Chapman, 2008; Alkharusi et al., 2011; Alkharusi et al., 2012); and
- Assessment Literacy Inventory (Mertler & Campbell, 2005; Davidheiser, 2013).

It should be noted that the studies using these tests provided limited information with regard to teachers' assessment knowledge progression because test results typically examined the overall raw score (e.g., average score and/or score range) (see Griffin et al., 2013 for a detailed discussion). In considering these tests, the two 35 multiple-choice item tests (i.e., Teacher Assessment Literacy Questionnaire and Assessment Literacy Inventory) appeared to be the most useful for teachers; however the tests would benefit from revision to be diagnostic for teachers within a classroom-based context. Thus, additional research is warranted to construct a developmental test to measure the assessment knowledge progression of teachers within the language and/or general education fields in order to provide more diagnostic information of their assessment knowledge areas that could be improved. To achieve this, the developmental test should be associated with the probabilistic interpretation of teachers' assessment knowledge progression (Griffin, 2007; Forster & Masters, 2010; Masters, 2013a). Within this probabilistic interpretation, each individual teacher's performance is interpreted in terms of developmental levels at which his/her score is simply used as the code for the levels of his/her assessment knowledge progression. As such, the developmental test can provide useful formative information in relation to teachers' assessment knowledge progression. Furthermore, future assessment knowledge test developers should also incorporate an additional two key stages of the assessment process, namely assessment record-keepings and managing quality assurance of the assessment process (see sections 2.2.7 and 2.2.9 in Chapter 2 for a detailed discussion), which are not covered in the seven standards for Teacher Competence in Educational Assessment of Students.

3.3.2 Assessment Beliefs

The second aspect of classroom assessment literacy of teachers has been associated with the personal beliefs about assessment. Within the perspectives of reasoned action theory (Fishbein & Ajzen, 1975, 2010) and planned behaviours theory (Ajzen, 1991, 2005), individuals' personal beliefs have been identified as the most powerful contributors to their behaviours/performances. Such beliefs have been indicated to affect their motivation, aspirations and outcome expectations (Zimmerman, 1995; Schunk & Pajares, 2004; Zimmerman & Cleary, 2006). The term "belief" has been

defined as the favourable or unfavourable evaluation/judgement of an individual towards the object, event, or person (Pajares, 1992; Ajzen, 1991, 2005). In the current study, the term “assessment belief” has been used to refer to an individual teacher’s perception of the worthiness/importance of different aspects of the assessment process, such as a specific method used to collect evidence. In the literature with respect to teachers’ beliefs, various terms such as beliefs, attitudes, perceptions and conceptions were used interchangeably (see Pajares, 1992; Rogers et al., 2007). For the purpose of clarity and consistency, in the studies reviewed next, the term “belief” was used to refer to “attitude,” “perception” and/or “conception”.

3.3.2.1 Stages of the Assessment Process: Teachers’ Beliefs

Substantial research has been undertaken to examine teachers’ personal beliefs about the purposes for conducting assessments (i.e., formative versus summative purposes). Teachers had been found to have differences in beliefs about the purposes for conducting assessments. For example, Brown, Lake, and Matters (2011) administered a questionnaire to examine 784 primary school teachers’ beliefs about the purposes for conducting their assessments in Queensland. The findings indicated that teachers had more endorsement towards assessments serving formative purposes (i.e., improving teaching and learning) than summative purposes (i.e., making students accountable through grades and/or certificates). Antoniou and James (2014), using classroom observations, document analyses and semi-structured interviews, also found that primary school teachers in Cyprus believed that formative assessments played an important role in promoting effective teaching and learning. Teachers had been found to have differences in beliefs about the usefulness of classroom developed assessments and large scale standardised tests. Large scale testing typically serves accountability purposes (i.e., summative functions) whereas classroom-based assessment serves both formative and summative purposes. For instance, Leighton, Gokiart, Cor, and Heffernan (2010), employing a questionnaire to explore secondary school teachers’ perceptions about assessments in Alberta, Canada, found that teachers believed their classroom assessment tasks generated more diagnostic information (i.e., in terms of the learning process,

consequences for meaningful learning and use of learning strategies) than that found within large scale tests. The study concluded that teachers endorsed classroom developed assessments more than large scale testing.

Previous research has also examined teachers' personal beliefs about the assessment methods (i.e., traditional versus innovative methods) to gather evidence of students' learning. For example, Gullickson (1984) surveyed 391 school teachers in rural Mid-western states in the US about their beliefs of the instructional use of tests. The findings indicated that teachers believed traditional assessment, such as tests, to be the best method to assess students' learning, as it was thought to enhance their instruction, increase students' effort, influence students' self-concepts and encourage competition amongst students. Similarly, Xu and Liu (2009) used an interview to examine one college EFL teacher's assessment belief, and they also reported that the teacher believed tests were the best assessment method to measure students' learning, as opposed to innovative assessment tasks. In contrast to the findings by Gullickson (1984) and Xu and Liu (2009), McMillan and Nash (2000), administering a questionnaire to examine school teachers' assessment beliefs, revealed that teachers believed that innovative assessment like construct-response tasks were the best assessment format to assess students' learning, as they provided more information about students' achievements. Similarly, Schwager and Carlson (1994), surveying school teachers about their assessment and instruction, found that teachers' beliefs about their assessments varied, depending on the way they viewed themselves as either traditional or innovative teachers. Cheng et al. (2004) and Rogers et al. (2007) found that ESL/EFL university teachers endorsed both innovative and traditional assessments. Canadian ESL/EFL university teachers employed innovative assessment methods more often than their Chinese and Hong Kong counterparts. In contrast, Chinese and Hong Kong teachers predominantly employed traditional assessment methods like tests in assessing students' learning. Inbar-Lourie and Donitsa-Schmidt (2009) surveyed 113 EFL school teachers in Israel about the factors underlying the usage of innovative assessment, and found that teachers' beliefs were the predictor of the use of innovative assessment to assess students' learning. The different findings amongst these studies in relation to teachers' beliefs about the use of traditional

and/or innovative assessment methods were relevant to the purposes for conducting assessments, and the values and culture of their institutions.

Teachers had been found to have differences in beliefs about the usefulness of quality assurance procedures in their assessment practices. For instance, Gullickson (1984) and Gullickson and Ellwein (1985), using a questionnaire to examine school teachers' assessment beliefs, reported that teachers believed that conducting statistical analyses of the test scores (i.e., calculating their test reliability or item analyses) were impractical. In his subsequent study, Gullickson (1986) administered a questionnaire to 24 professors, who were teaching an educational measurement course, and 360 school teachers about their perspectives of pre-service educational measurement courses. Findings indicated that the professors believed that undertaking statistical analyses of the test scores were important whereas the teachers believed such statistical analyses were unnecessary. Oescher and Kirby (1990) also reported that teachers believed their tests were reliable and valid, despite the fact that statistical analyses of the test scores were not carried out. In contrast, King (2010), employing a questionnaire to examine school teachers' and administrators' beliefs about assessments, reported that teachers and administrators had more favourable endorsements toward educational statistics and they believed that conducting statistical analyses of the test scores were useful to them.

In summary, all studies that examined teachers' personal beliefs about assessment predominantly employed questionnaires through using rating-scales, with the occasional use of classroom observations, document analyses and interview techniques. Studies that utilised rating-scales provided little information with respect to the developmental intensity or progression of teachers' personal beliefs about assessment, as they interpreted their results in terms of using an average score. More research is needed to construct developmental rating-scales to measure the levels of progression of teachers' personal beliefs about assessment. Such scales would provide better understanding of such beliefs about assessment.

3.3.3 Relationship between Assessment Knowledge and Assessment Practice

Within social cognitive (Bandura, 1989, 1997), reasoned action (Fishbein & Ajzen, 1975, 2010) and planned behaviours theories (Ajzen, 1991, 2005), a lack of requisite knowledge and/or skills of any individual teachers make them unable to carry out the intended tasks. In other words, if the teacher lacks knowledge and/or skills to implement the assessment, it is more likely that s/he fails to conduct the intended assessment and/or implements poor quality assessments.

There were several studies undertaken to examine the relationship between teachers' assessment knowledge base and their assessment practices (Mertler, 2000; Harlen, 2005b; Alkharusi et al., 2012). For example, Mertler (2000) administered a questionnaire to teachers across school levels in different subject areas about their assessment practices. Mertler found that teachers who demonstrated limited knowledge and understanding of the concepts of validity and reliability tended not to use statistical procedures to analyse their assessment data such as calculating test reliability or item analyses. Similarly, Black, Harrison, Hodgen, Marshall, and Serret (2010) conducted a qualitative study with 12 school teachers in Oxfordshire in the United Kingdom to examine their knowledge and understanding of the concept of validity in their summative assessments and how this impacted on their assessment practices. They reported that teachers' knowledge of the concepts of validity and reliability was limited, and such inadequacies had been shown to contribute to variations of their actual assessment practices. Such studies highlight that teachers' limited assessment knowledge base can have a negative impact on their assessment implementation.

3.3.4 Relationship between Assessment Belief and Assessment Practice

Within reasoned action (Fishbein & Ajzen, 1975, 2010) and planned behaviours theories (Ajzen, 1991, 2005), personal beliefs of individual teachers have been proposed as a good predictor of their behaviours/performances. In other words, these personal beliefs can influence the ways teachers carry out their assessments. Within the language education literature, there were very few studies that examined the relationship between

teachers' personal beliefs about assessment and actual classroom assessment practices. For example, Rogers et al., (2007) employed a questionnaire to investigate the ESL/EFL university teachers' beliefs about assessment in three ESL/EFL tertiary contexts. The findings indicated that university teachers' personal beliefs were attributable to variations in their assessment practices. The researchers concluded that these beliefs were mixed and contradictory. There was a positive relationship between what teachers perceived as the importance of assessments on instruction and student learning, and the actual uses and purposes of assessments. The teachers' beliefs about the way assessment should be implemented, the time needed for assessment, and their preparation for and understandings of assessments were directly linked to their actual practices.

Within the general education field, a greater number of studies have been undertaken to explore the relationship between school teachers' assessment beliefs and their actual practices. A study that examined the association between teachers' beliefs was completed by Brown et al. (2009). Brown et al. (2009) surveyed 288 Hong Kong school teachers' beliefs about assessment and found that there were statistically significant relationships between teachers' personal beliefs about the purposes of assessments and actual practices.

There are a small number of studies that have examined teachers' beliefs about the purposes and stakes of the assessment and how such beliefs influenced their actual assessment practices. Of these studies, all have found a positive relationship between beliefs and behaviour. For example, Gay (1990) surveyed 168 school teachers in North Carolina in the United States about their test administration and found that 35% of these teachers believed standardised achievement tests had accountability and high-stakes functions. Given such beliefs, teachers reported their engagement in unethical behaviours such as allowing students to talk during testing, leaving students unsupervised and making gestures to help students choose the correct answer to raise students' test scores during test administration. In alignment with Gay's (1990) findings, Herold (2011) interviewed one high school teacher about the administration of the Pennsylvania System of School Assessment (PSSA) exams to her students and found that because she believed these exams were associated with accountability and high-stakes functions, she gave her students definitions for unfamiliar words, discussed with them reading passages they

didn't understand, and commented on their writing responses at various times. On a few occasions she even pointed them to the correct answers on difficult test questions during test administration. Given the small sample size of Herold's study, such findings should be treated as tentative only. Thus, teachers' personal beliefs about accountability purposes and high stakes in assessment were related to their involvements in unethical behaviours during test administration.

Other studies have examined the relationship between school teachers' beliefs about the perceived importance of students' non-academic achievement factors and how such beliefs could impact on their actual assessment practices. For example, Stiggins, Frisbies, and Griswold (1989), Brookhart (1993) and McMillan and Nash (2000) reported that teachers believed that students' non-academic achievement factors, such as efforts and attitudes, were important because these factors played a crucial role in students' learning. They therefore incorporated these factors into their final course grades. Such findings reinforced teachers' personal beliefs about assessment, directly linked to their actual assessment practices.

There are, however, a number of studies that have revealed that teachers' assessment beliefs had no relation to their actual assessment practices. For example, Rieg (2007) found that although secondary school teachers recognised the effectiveness of providing students with opportunities to choose various forms of assessment methods to assess their learning, and assisting them in preparing for assessments by offering them study skills and non-graded tests and quizzes, they did not implement such strategies in their actual assessment practices. Similarly, Black et al. (2010) found that despite school teachers having positive endorsements toward the use of innovative assessment methods, they retained employing the test formats in their actual assessment practices. Such discrepancies between teachers' assessment beliefs and their actual practices had been attributed to teachers' insufficient assessment knowledge base, which led to their inability to implement the intended assessments, despite their positive endorsement of such assessments.

3.4 Summary

In summary, teachers' assessment knowledge base and their personal beliefs about assessment are important factors that underpin classroom assessment literacy and affect their actual assessment practices. Teachers' limited classroom assessment knowledge base was found to have a negative impact on their assessment implementation (Mertler, 2000; Harlen, 2005b; Black et al., 2010; Alkharusi et al., 2012). Teachers' limited assessment knowledge base was found to contribute to their predominant use of traditional assessment for summative functions, a lack of using statistical procedures to analyse their assessment data, bias in judging students' work and variations in their assessment practices. Equally, teachers' personal beliefs about assessment were found to result in variations in their assessment practices (Gullickson, 1984; Oescher & Kirby, 1990; Brookhart, 1993; Rogers et al., 2007; Xu & Liu, 2009; Leighton et al., 2010; Brown et al., 2011). Many teachers considered that traditional assessment was the best method to assess their students' learning and improving their instruction, whereas others viewed innovative assessment tasks as more effective assessment methods. Whilst some teachers believed conducting statistical analyses of their test data to be unnecessary, others believed it was important to undertake such statistical analyses. Some also believed their students' non-academic achievement factors, such as efforts and attitudes, played a role in students' learning and therefore included such factors in the determination of final course grades. Thus, teachers' assessment knowledge base and their personal beliefs about assessment were the underpinning factors that altered their current assessment practices.

As an outcome of the literature review, it appears there was a lack of measurement scales that provided diagnostic or formative information about the developmental assessment knowledge base and/or progression of teachers' personal beliefs about assessment. More research is warranted to construct these developmental scales in order to address such a gap in the literature. In the next chapter, the body of exploratory research regarding the influence of the background characteristics of teachers on their classroom assessment literacy will be explored.

Chapter 4: Background Characteristics Influencing Classroom Assessment Literacy

This chapter explores the collection of studies undertaken to examine the influence of teachers' background characteristics on their classroom assessment literacy (i.e., assessment knowledge base and assessment beliefs). This research is explored within both general and language education fields.

4.1 Background Characteristics Influencing Classroom Assessment Literacy

The background characteristics that have been identified to potentially influence levels of classroom assessment literacy (i.e., assessment knowledge base and assessment beliefs) of teachers include their pre-service assessment training, teaching experience, academic qualification, class size, professional development, teaching hours, gender and prior assessment experience as students. Each will be considered in this chapter.

4.1.1 Pre-service Assessment Training

Research showed that the quality of assessment training teachers received during their pre-service teacher education programmes impacted their assessment knowledge base. For example, several studies (Mayo, 1967; Mertler, 1999; King, 2010) revealed that teachers received insufficient assessment training in their teacher education programmes, while others reported that, in some instances, teachers did not receive any formal education in assessment during their pre-service teacher education programmes (Gullickon, 1984). Wise, Lukin, and Roos (1991) added that most teachers in their investigation reported studying assessment related principles and procedures for less than one semester during their pre-service teacher education programmes. Furthermore, less than half of the teachers surveyed viewed their pre-service training as insufficient and attributed their assessment knowledge to their teaching experience. In line with Wise, Lukin, and Roos (1991), Impara, Plake, and Fager (1993) indicated that less than 70% of

the teachers in their study reported having some assessment training during their pre-service teacher education programmes while nearly 30% of teachers reported having no assessment training. Brown (2008) also reported that one in seven teachers in his investigation indicated they had no assessment training, whereas a third had received some assessment training as part of their pre-service teacher education programmes. This finding was disturbing, given that research has shown that the assessment knowledge base of in-service teachers who had received specific training in assessment during their pre-service education programmes was higher than those who did not receive any pre-service assessment training (Plake, 1993; Plake, Impara, & Fager, 1993; Plake & Impara, 1997; Zhang & Burry-Stock, 2003; Schaff, 2006). It can therefore be seen that pre-service teacher education programmes play an important role in ensuring that all teachers are provided with the specific training in educational assessment required for effective classroom-based assessments.

This raises curriculum issues associated with pre-service programmes. For example, there has been a concern associated with embedding educational assessment training into other teacher education courses (e.g., Teaching Methodology). This was largely due to the fact that assessment was often overlooked (Schafer & Lissitz, 1987) or that it was taught by instructors who lacked assessment expertise (Schafer & Lissitz, 1987). As such, there is general agreement among assessment and measurement experts that it is important for teacher education programmes to provide pre-service teachers with a stand-alone course, focusing on educational assessment knowledge and skills, in order to enable them to acquire assessment-related principles and procedures required within their professional practice (see Stiggins, 1991b, 1999; Griffin et al., 2012; Leung, 2014).

There had also been reports that in a number of pre-service education courses, although there was a component of assessment in the pre-service education curriculum, much of this was irrelevant to their work as classroom teachers. And as such, the assessment knowledge and skills required tended to be acquired through their teaching experience (Gullickson, 1984, 1986). Such irrelevant assessment knowledge appeared to be related to the use of dated assessment textbooks in pre-service teacher education programmes. As Masters (2013a) pointed out, educational assessment tended to be treated at a superficial level in pre-service teacher education programmes and the

assessment textbooks employed were usually based on 20th century educational assessment principles and procedures, most of which hampered rather than developed student teachers' explicit understandings about assessments. Campbell and Collins (2007) found that educational assessment textbooks mainly emphasised traditional assessment aspects, whereas other crucial topics such as assigning grades, selecting and constructing response items, developing rubrics and assessing performance assessments, and interpreting assessment results were not consistently identified as important, or missing.

Limited pre-service training in educational assessment does not appear to be restricted to the general education sector. For example, Tsagari (2008) found that EFL school teachers in her study had inadequate assessment training. The extent to which there is a direct relationship between pre-service training in assessment and assessment expertise has yet to be examined in the language education field.

4.1.2 Teaching Experience

Previous research revealed the impact of years of teaching experience on teachers' perceived assessment competence. For instance, Zhang and Burry-Stock (1997) found that teaching experience were the main predictors of teachers' perceived assessment competence in developing performance assessments and using informal observations. Similar findings were reported by Chapman (2008) and Alkharusi (2011), who both found that teaching experience contributed to a greater self-perceived assessment competence. Teaching experience has also been found to be related to self-confidence. For example, Bol, Stephenson, O'Connell, and Nunnery (1998) reported that teachers with more teaching experience demonstrated more confidence in employing innovative assessment tasks than those with less teaching experience. Such findings suggest that teaching experience has an association with teachers' perceived assessment competence.

Despite such an association between perceptions and teaching experience, numerous studies have found no relationship between the *actual* level of teachers' assessment knowledge base (i.e., indicated by the test score) and teaching experience (Schaff, 2006; Chapman, 2008; King, 2010; Alkharusi et al., 2011; Gotch & French,

2013). This set of findings typically indicated that over time, and without supplementing teachers with further in-service assessment training, the level of assessment knowledge base decreased.

It appeared there was a discrepancy between the findings reported by studies employing self-reported measures and those using test instruments. Whilst the self-reported measures indicated the influence of years of teaching experience on teachers' perceived competence in educational assessment, test instruments showed that years of teaching experience had no impact on teachers' assessment knowledge base. Given there are few, if any studies, that have examined the relationship between teaching experience, assessment knowledge expertise and perceived assessment competence, further research in this area is required.

4.1.3 Academic Qualification

A range of studies showed mixed results with respect to the level of academic qualification of teachers and how this influences the level of their classroom assessment knowledge base. For example, Chapman (2008) found that teachers' academic qualifications (i.e., bachelor versus masters degree) had no impact on their classroom assessment knowledge. In contrast, King (2010) reported that teachers and administrators with advanced qualifications comprising specialist and doctorate degrees had significantly higher assessment knowledge than those who possessed a bachelor or masters degree. Such mixed results may be due to these studies looking at the qualification level only, and not necessarily what the course specialised in.

4.1.4 Gender

There is limited research that has specifically focused on gender and assessment literacy. Of the few studies that have been undertaken, the focus has typically been on the relationship between gender and self-perceived assessment competence. For example, Alkharusi (2011) reported there were statistically significant differences in teachers' self-perceived assessment competence with respect to their gender. Female teachers perceived themselves to be more skilful in writing test items and communicating assessment results

than did male teachers. This finding, however, should be treated with caution, given the research indicated that self-reported measures had an association with social desirability response bias (Chapman, 2008; Alkharusi et al., 2011; Alkharusi et al., 2012). As such, it was possible that female teachers in this study may have overreported their assessment competence, yet further research in this area is required before any definitive conclusions can be made on the relationship between gender and assessment competence.

4.1.5 Professional Development

Considerable research has shown that providing school teachers with in-service professional development focusing on educational assessment promotes assessment knowledge (Borko, Mayfield, Marion, Flexer & Cumbo, 1997; Sato, Wei, & Darling-Hammond, 2008; O’Leary, 2008; Mertler, 2009; Black et al., 2010; Towndrow, Tan, Yung, & Cohen, 2010; Koh, 2011; Griffin, Care, Robertson, Crigan, Awwal, & Pavlovic, 2013). Such findings highlighted Stiggins’ (2010) argument that the lack of in-service assessment training delivery worldwide contributed to a limited assessment knowledge base of teachers.

4.1.6 Class Size

A body of research has shown that class size influences classroom assessment practices. For example, Gibbs and Lucas (1997) found that large class sizes impacted assessment implementation of teachers at Oxford Brooks University in the UK by means of substituting assessment tasks, such as essays, with tests/examinations in assessing students’ learning. The researchers also reported that teachers with large class sizes used less coursework assessments than teachers with small class sizes. Consistent with Gibbs and Lucas’s (1997) findings, Nakabugo et al. (2007) reported that large class sizes impacted assessment practices of primary school teachers in Uganda. The researchers revealed that large class sizes induced teachers to employ fewer performance-based assessment tasks in order to reduce heavy marking loads. The teachers also used ticks, crosses and marks in assessing students’ work without giving descriptive feedback on the strengths and weaknesses of their work. In line with Gibbs and Lucas’s (1997) and

Nakabugo et al.'s (2007) findings, Sun and Cheng (2013) reported that class sizes influenced Chinese EFL school teachers' grading practices (i.e., the extent to which the teachers took into consideration students' efforts and study habits in their grading decision making). Hence, there appears to be strong evidence to suggest that as class size increases, there is less likelihood for teachers to implement innovative assessments, provide constructive and timely feedback to students and to not be influenced by extraneous factors in their assessment decision-making processes. While studies have looked at the impact of class size on assessment implementation, there does not appear to be any research focusing on an association between class size and classroom assessment literacy.

4.1.7 Teaching Hours

Research also highlighted that the number of teaching hours influenced teachers' classroom assessment practices. For instance, Rath (2010) found that some teachers never returned the marked assignments to their students and did not provide feedback on students' work, due to the fact that they had so many teaching hours per week. This assessment pattern was in contrast to their self-perception of the vital role of feedback in helping students diagnose the areas needing to be improved. Consistent with Rath's (2010) findings, Haing (2012) reported that the number of teaching hours per week (i.e., 30 hours) negatively impacted on the way in which Cambodian EFL university teachers designed their assessment tasks and the way they marked students' work. Although research has looked at teaching hours in relation to assessment implementation, there does not appear to be any research that has examined the relationship between teaching hours and classroom assessment literacy.

4.1.8 Assessment Experience as Students

Research further showed that the assessment experience teachers had as students influenced their personal beliefs toward undertaking assessments. For example, Green and Stager (1986) reported that school teachers' prior learning experience (as students) influenced their current beliefs about classroom tests. Given that teachers' prior learning

was assessed by tests, teachers employed tests to assess students' learning. Xu and Liu (2009) also found that one EFL college teacher's previous assessment experience (as a student) impacted her current assessment practice and her future plans for conducting assessments. Specifically, this teacher experienced traditional assessment methods, such as tests, as the major form of assessment during her undergraduate study, and her predominant test-taking experience led her to trust summative assessment. This finding highlighted the influence of the type of assessment experience encountered by teachers as students in relation to enactive learning or learning by doing within the social cognitive theory (Bandura, 1997). Bandura (1997) contends that this type of learning is the most powerful way of influencing individuals' future performances. As such, the assessment experience (as students) with traditional assessment (e.g., tests) caused them to value such methods and therefore led them to employ tests to assess their students' learning.

4.2 Summary

In summary, there were similar findings with respect to the impact of teachers' background characteristics on their classroom assessment literacy (i.e., pre-service assessment training, professional development participation and assessment experience as students) and practice (i.e., class size and teaching hours). There were, however, mixed findings with regard to the influence of teachers' teaching experience and academic qualifications on their level of classroom assessment knowledge. Given teachers' background characteristics have been shown to impact such levels, further research needs to explore the interplay between background characteristics of teachers and their classroom assessment literacy in order to obtain a greater understanding of the nature of classroom assessment literacy.

In the next chapter, the methodology that underpinned the current study will be presented.

Chapter 5: Methodology

This chapter is divided into three parts. Part one presents the research design of the study and justifies the use of a mixed methods approach. Part two presents an overview of the quantitative phase of the study, including a description of the sampling framework, the data collection procedures, scales development processes and the data analysis procedures employed. Part three details the qualitative phase of the study, in which the selection and characteristics of the interview sample are described, the process employed for designing the interview questions is documented, and data collection as well as data analysis procedures employed to examine the documents collected and interview data presented.

5.1 Part One: Mixed Methods Approach

5.1.1 Rationale and Key Characteristics of the Mixed Methods Approach

The present study employed a mixed methods approach through integrating quantitative and qualitative methods within a single study to obtain a better understanding of the complexity of research inquiry for classroom assessment literacy (Johnson & Onwuegbuzie, 2004; Bryman, 2006; Greene, 2007; Tashakkori, Teddlie, & Sines, 2013; Creswell, 2014). A mixed methods approach was chosen, as either quantitative or qualitative methods were not adequate in providing an in-depth understanding into the complexities of classroom assessment literacy of EFL instructors (Creswell, 2012). The use of mixed methods complements the strengths and weaknesses of quantitative and qualitative methods (Johnson & Onwuegbuzie, 2004; Greene, 2007; Teddlie & Tashakkori, 2009; Creswell, 2014).

Quantitative and qualitative methods have different strengths and weaknesses. Quantitative methods are associated with deductive process involving a large number of participants and data. The main features of quantitative methods are relevant to deduction, confirmation, hypothesis testing, standardised data collection and statistical analysis (Silverman, 2013; Creswell, 2014). Despite the fact that quantitative results have

the potential to be generalised with regard to the whole population, they usually attract criticism due to the difficulty in deriving meanings brought to social life, due to the use of pure quantitative logic, namely numbers, that exclude the study of many interesting activities people actually do in their day-to-day lives (Creswell, 2012; Silverman, 2013). While hypotheses can be tested through self-reports of knowledge, attitudes or behaviours for quantitative methods, it has been argued that the data have limitations in the depth of information received, since self-reported measures cannot be probed directly (Edmonds & Kennedy, 2013; Creswell, 2014).

Alternatively, the use of qualitative methods typically involves a small number of participants. It is thought to provide a richer description of the genuine picture and complexities of reality, as it explores things in their natural setting. Qualitative methods have been associated with “how” and “why” things occur the way they do, and numerous researchers have argued that one can gain an in-depth understanding and focus on meanings (Nunan, 1992; Creswell, 2012; Silverman, 2013). This is relevant to induction and exploration of the underlying issues of a research inquiry. However, the sole use of qualitative data collection has been criticised for limited generalisation, as it does not entail sampling techniques and it is a time-consuming process (Teddlie & Tashakkori, 2009; Creswell & Plano Clark, 2011; Edmonds & Kennedy, 2013).

Within a mixed methods approach, the researcher typically constructs knowledge on pragmatic grounds (Teddlie & Tashakkori, 2009; Creswell & Plano Clark, 2011; Tashakkori et al., 2013; Creswell, 2014). Quantitative and qualitative data can be collected concurrently or sequentially, which can assist the researcher to better comprehend the research inquiry. In designing a mixed methods study, the researcher needs to take into consideration three main issues: priority, implementation and integration (Creswell & Plano Clark, 2011; Creswell, 2014). Priority is associated with whether quantitative or qualitative methods are given greater focus in the study, while implementation is relevant to the collection and analysis of quantitative and qualitative data concurrently or sequentially. Integration refers to the phase in the research process where the researcher combines or mixes both quantitative and qualitative data.

5.1.2 Mixed Methods Sequential Explanatory Design

The current study employed a sequential explanatory design comprising two distinct phases (Creswell & Plano Clark, 2011; Creswell, 2014). Phase one comprised both test and survey design methodology in which measuring instruments were developed and administered to the EFL university instructors. The goals of the quantitative phase were to develop and validate a set of scales to measure classroom assessment literacy development of instructors including their classroom assessment knowledge and personal beliefs about assessment. The quantitative phase was also used to evaluate whether the set of scales was measuring a single, underlying construct referred to as “classroom assessment literacy” (see section 5.2.4.2 for a detailed explanation of the hypothesised one-factor congeneric measurement model being tested), and to examine the impact of the instructors’ background characteristics on such measures. Phase two included collecting documents of the learning goals and assessment-related policies from the two recruited departments. Phase two also included semi-structured interviews with a small number of selected participants from the first phase of this study. The goals of the qualitative phase were to obtain an in-depth understanding of instructors’ classroom assessment literacy levels and the impact it may have had on their actual assessment practices. Furthermore, it aimed to explore the influence of instructors’ background characteristics (e.g., assessment experience as students) and the departments’ assessment-related policies (e.g., assessment purposes) on instructors’ classroom assessment literacy and implementation. Thus, the priority in the present study was given to the first quantitative phase, as it was mainly used for collecting data to address the proposed research questions. Although these two phases were administered separately, the results from both quantitative and qualitative phases were integrated through the interpretation of the findings for the whole study (see Figure 5.1).

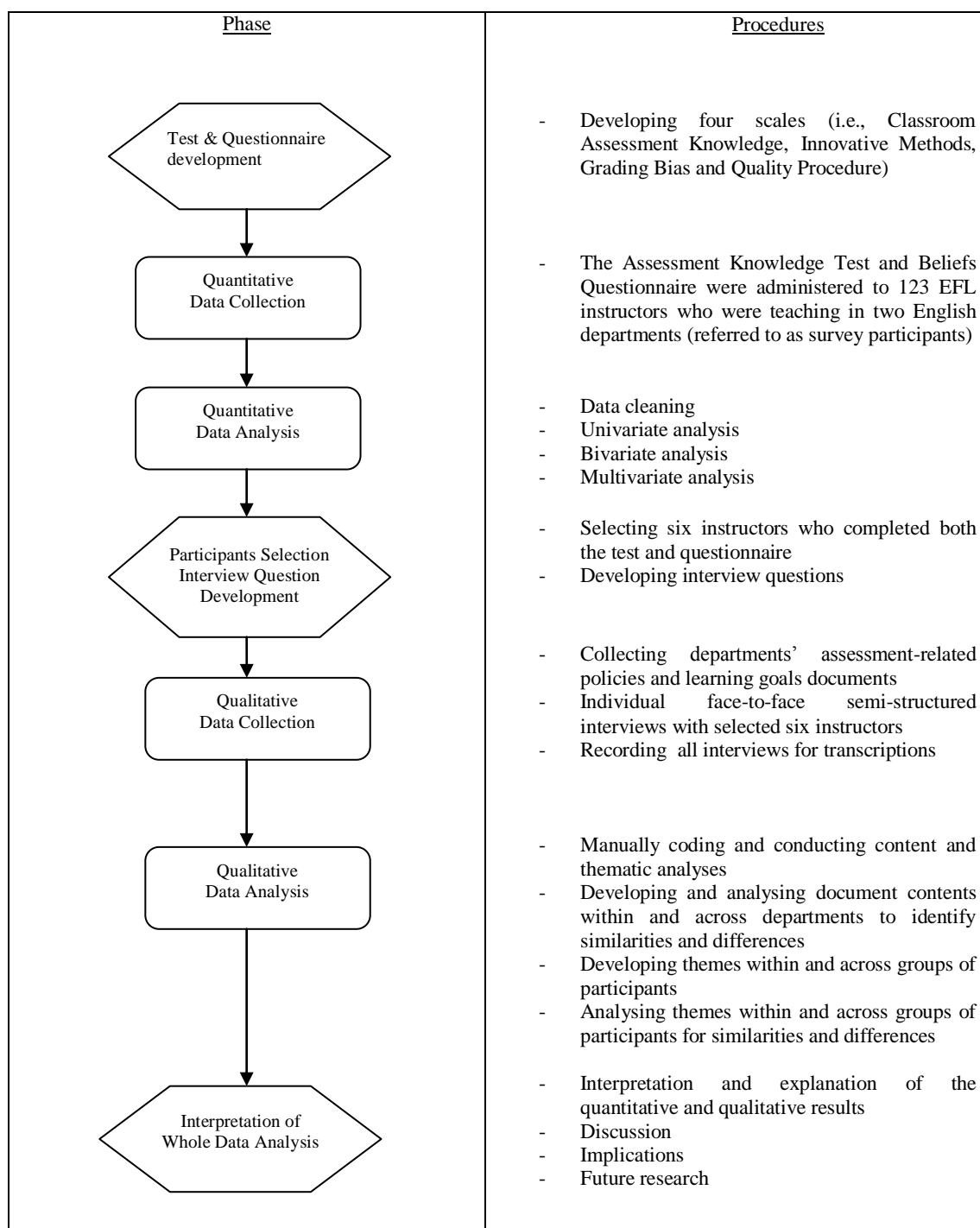


Figure 5.1 Diagram for the mixed methods sequential explanatory design procedures

5.1.3 Advantages and Challenges of the Sequential Explanatory Design

Educational researchers have frequently discussed the strengths and challenges of using a mixed methods approach in the literature (Johnson & Onwuegbuzie, 2004; Bryman, 2006; Greene, 2007; Teddlie & Tashakkori, 2009; Creswell, 2014). Specifically, the advantages of employing the sequential explanatory design (Teddlie & Tashakkori, 2009; Creswell, 2012) for the present study were:

1. It provided complementary data in terms of capturing both quantitative and qualitative data, by exploring the quantitative results in greater depth through follow-up qualitative data.
2. It was logical due to its sequential phases: one stage leading to the next stage.
3. It was manageable for a single researcher to conduct the entire study, as it comprised two distinct phases.

However, the challenges and/or difficulties of employing the sequential explanatory design for the present study were:

1. It required the researcher to make decisions as to which quantitative results needed to be explained in the qualitative phase.
2. It required more time to complete the study.
3. It demanded the researcher to make decisions concerning who would be sampled in the qualitative phase and what criteria would be employed for selecting these participants.

(Creswell, 2012)

The following sections detail the research methods employed in the current study.

5.2 Part Two: Quantitative Phase

5.2.1 The Target Sample

The current study was limited to classroom assessment literacy. The target sample (Ross, 2005) of the study was defined as all classroom-based instructors who were

responsible for developing their own assessment tasks for teaching and learning within EFL programmes in Cambodian Higher Education institutions. According to the list obtained from the Ministry of Education, there were 1893 EFL instructors who were teaching in 60 universities in Cambodia (The Department of Cambodian Higher Education, 2009). This target sample was representative of the national group of classroom-based instructors within EFL programmes in Cambodian universities.

5.2.1.1 The Sampling Framework

The present study employed a purposive sampling within one Cambodian city-based university comprising two English departments. The university is the oldest and largest institution that offers EFL programmes. It is the only university in Cambodia that provides a Bachelor of Education in teaching English as a Foreign Language (TEFL) to English-major scholarship and fee-paying students. Hence, this university was selected on the basis that participant course enrolment characteristics matched the study aims.

The recruited departments were an English-major department and an English non-major department. The former department comprised 70 EFL instructors and 3200 students. In contrast, the latter department consisted of 53 EFL instructors and 5074 students. All 123 instructors within both departments were invited to participate in the current study.

The English-major department offered the Bachelor of Education in TEFL and the Bachelor of Arts in English for Work Skills through a four-year course designed for both scholarship and fee-paying English major students. From year 1 to year 3, students were taught general English courses to enrich their English language proficiency. In contrast, in year 4, students were trained to become EFL school teachers or tertiary instructors if they enrolled in the Bachelor of Education programme. Within this programme, they studied Teaching Methodology, Applied Linguistics and Foundations of Education subjects. However, if they enrolled in the Bachelor of Arts in English for Work Skills programme, they studied Report Writing, Communication Skills, Translation and Interpreting and English for International Business subjects. In contrast, the English non-major department provided a non-degree programme that offered general English

language courses for three years' duration. The programme is available to both scholarship and fee-paying non-major English students at various departments within this selected university as an academic platform to enrich their level of English language proficiency that corresponds to the requirements of their major subjects.

Given the study selected the sample from two English departments within the same university, this can be reflected in a higher intra-class correlation (Izard, 2004a). The effect of this (i.e., design effect) is to overestimate the statistical significance in most cases (Ross, 1993). The design effect is associated with the ratio of the variance of an estimate under the complex sample design (i.e., clustering) to the variance of the same estimate that would apply with a simple random sample (Ross, 2005). There was, however, no adjustment for design effects in the current study given its results were not used to infer to other universities in Cambodia and beyond. That is, its results were related to only the selected university, which was known as purposive sampling.

5.2.2 Data Collection Procedures

Prior to approaching the targeted participants, the permission letters for undertaking the study had been obtained from the rector of the university and the heads of the two English departments. All 123 instructors were approached to voluntarily participate in the study by completing the test and questionnaire. Both instruments were designed to measure the classroom assessment literacy development of instructors.

5.2.2.1 Response Rate

Of the total 123 instructors, 108 (i.e., 59 English-major and 49 English non-major instructors) completed both the test and questionnaire, yielding a response rate of 88%. The remaining 15 instructors did not participate in this study, as they did not complete both instruments. Given the study received a high response rate of 88%, the data obtained was judged representative of the selected university.

5.2.2.2 Test and Questionnaire Administration

Prior to completing the test and questionnaire, each instructor was provided with a written statement explaining the purposes of the study and a consent form outlining the voluntary nature of the participation. Both the test and questionnaire were written in English, given respondents were the instructors of English language subjects. In addition to completing the test and questionnaire, all instructors were also invited to participate in the second phase of the study (refer to section 5.3.1 regarding detailed procedures employed to select six instructors). If the participants decided to take part in phase two, they were asked to complete and sign the consent form provided.

The test in the present study titled “Classroom Assessment Knowledge” comprised 27 multiple-choice items with each item having four options: the correct answer and the other three options being the distracters (see Appendix A). Approximately 45-60 minutes were required for each instructor to complete the test.

The questionnaire comprised two main parts: instructors’ background characteristics and their personal beliefs about assessment. Part one required participants to provide their demographic information including age, gender, departmental status, academic qualification and its discipline, teaching experience, number of teaching hours, number of students per class, formal study in educational assessment, duration of assessment course run and their perceived assessment preparedness. Part two focused on instructors’ personal beliefs about assessment. The items were organised around three scales: Innovative Methods (9 items), Grading Bias (7 items) and Quality Procedure (6 items). For each of the assessment beliefs, participants were asked to rate the items on a four-point rating scale including “not useful at all, a little useful, useful and very useful,” “never, sometimes, often and always,” and/or “strongly disagree, disagree, agree and strongly agree” (see Appendix B). Approximately 15-25 minutes were required for each instructor to complete the questionnaire.

5.2.3 Test and Questionnaire Development Processes

This section details the development of a set of scales employed in the study. All scales in the test and questionnaire had been developed based on an extensive review of

the existing literature from both general education and language education fields. The aim of developing these four scales was to measure classroom assessment literacy development of instructors.

All items were reviewed by two groups of six specialists in the areas of EFL/ESL and/or general classroom-based assessment. The items were reviewed in both Australia and in Cambodia with such specialists. These two groups of panellists were initially asked to complete the “Classroom Assessment Knowledge” test and compare their answers to identify any disagreement with the answer key. Secondly, the panellists were requested to check the appropriateness of the items and underlying constructs and the match of items to research questions (i.e., content validity). Finally, the panellists were asked to check for gender bias, racial bias, stereo-typing roles, language issues, mechanics (spelling abbreviations, acronyms, punctuation and capitalisation), grammar (sentence structure, pronouns, verb forms, uses and tenses) and clarity (conciseness and consistency) of the items (i.e., to enhance the construct validity).

Most of the comments received from these two groups of panellists were associated with ambiguity of five items in the “Classroom Assessment Knowledge” test. Similarly, the main comments received for the items in the questionnaire were in relation to ambiguity and expression of some items such as clarity of instructions, verb forms, uses and tenses and sentence structure. Overall, the specialists agreed that the “Classroom Assessment Knowledge” test was a suitable instrument for measuring the instructor assessment knowledge within a classroom-based assessment context. They also considered that the questionnaire was an appropriate measure in examining the instructors’ personal beliefs about assessment in both the language and general education contexts. Based on the feedback from these panellists, the test and questionnaire were revised prior to the pilot. These revisions enhanced the face and content validity of the instruments.

Prior to the main data collection, both the “Classroom Assessment Knowledge” test and questionnaire had been piloted with 78 instructors in another Cambodian institution delivering EFL courses. The pilot data was analysed to show how each item performed individually and as a subscale set. Preliminary investigation of the reliability

of each of the measures (i.e., Cronbach's alpha) was also carried out using the pilot data. Minor modifications were made to the items in light of the pilot findings.

5.2.3.1 The Measures

The Background Characteristics

The participants were asked to provide their background characteristics in the questionnaire. The questions and associated scale of measurement with response options are displayed in Table 5.1 below. Four interval variables (i.e., age, years of teaching experience, number of teaching hours per week and class size) were recoded into ordinal variables to undertake cross-tabulation analyses.

Table 5.1 Instructor Background Information

Background Characteristics	Variable Name	Data Sought	Measure Level	Recode Variable Name	Values
Age	AGE	Years	Interval ordinal \Rightarrow	R_AGE	1= 22-25 2= 26-30 3= 31-39 4= 40-55
Gender	GENDER	Male/Female	Nominal	GENDER (same as previously)	1= male 2= female
Current teaching department	DEPART	English-major/Non-major departments	Nominal	DEPART (same as previously)	1=English-major department 2= English Non-major department
Highest academic qualification held	HAQ	Bachelor/Master/Doctoral	Nominal	R_HAQ	1= Bachelor 2= Master & Doctoral
Qualification discipline	QD	Education/Law/Business/Politics/Others	Nominal	QD (same as previously)	1= Education 2= Law 3= Business 4= Politics 5= Others
Years teaching English at university level	TEXPER	Number of years	Interval ordinal \Rightarrow	R_TEXPER	1= 1 2= 2 3= 3-5 4= 6-20
Current hours taught per week	THOUR	Number of hours	Interval ordinal \Rightarrow	R_THOUR	1= 5-12 2= 13-21 3= 22-50
Average number of students taught per class	CSN	Number of students	Interval ordinal \Rightarrow	R_CSN	1= 25-28 2= 29-35 3= 36-55
Previous formal studies in assessment during undergraduate teacher preparation programme	FSA	Yes/No	Nominal	FSA (same as previously)	1= Yes 2= No
Length of assessment course	LAC	Select one among the given four options	Ordinal	R_LAC	1= Less than 1 semester 2= 1 or more than 1 semester (s)
Level of preparedness for designing and conducting classroom-based assessment	LPA	Select one among the given four options	Ordinal	R_LPA	1= Unprepared 2= Prepared 3= Very Prepared

The next section presents the four scales constructed to measure the classroom assessment literacy development of instructors including their classroom assessment knowledge base and their personal beliefs about assessment. The details with regard to the development of these four scales are reported in Chapter 6. The names of the constructs (expressed by the scales) are shown in lower case, whilst the names of the

variables employed in the analysis are presented in capitals. The constructs and associated variables were:

1. Classroom Assessment Knowledge scale, labelled CAK.
2. Innovative Methods scale, labelled IM.
3. Grading Bias scale, labelled GB.
4. Quality Procedure scale, labelled QP.

Each of the constructs and associated variables has been described below.

The Classroom Assessment Knowledge Scale

The Classroom Assessment Knowledge (CAK) scale measured the level of the instructors' assessment knowledge base in a classroom-based assessment context. It comprised 27 multiple-choice items with three items related to each of the nine subscales (see Appendix A). The instrument was designed to cover the seven standards for Teacher Competence in Educational Assessment of Students issued by the American Federation of Teachers (AFT), the National Council on Measurement in Education (NCME), and the National Education Association (NEA) (AFT, NCME, & NEA, 1990); with two new standards, totalling nine standards. The expanded standards for the present study took into consideration criticisms associated with the narrow aspects of the original standards, particularly in relation to classroom-based assessment activities required by instructors in their daily instruction (Schafer, 1991; Stiggins, 1995, 1999; Arter, 1999; Brookkhart, 2011a). That is, although the seven standards covered some vital stages in the assessment process, they did not address two crucial stages: keeping accurate records of assessment data and managing quality assurance of the assessment process. As such, the new standards have been incorporated to address keeping accurate records (Griffin & Nix, 1991; Airasian, 2000; Bachman & Palmer, 2010) and managing quality assurance (Dunbar et al., 1991; Gipps, 1994b; Harlen, 1994, 2007; Gillis et al., 2009). The latter can lead to improvements in the accuracy, appropriateness, fairness and transparency of assessment outcomes in order to ensure comparability of standards in undertaking assessments across and between classes and universities (Dunbar et al., 1991; Gipps,

1994b; Harlen, 1994, 2007; Gillis et al., 2009). As Table 5.2 displays, standards 1, 2, 3, 4, 5, 6 and 9 were the original seven standards issued in 1990 by the AFT, NCME and NEA, whilst standards 7 and 8 are the new ones. Eleven items (2, 3, 4, 12, 13, 14, 16, 17, 21, 23 & 25) within the “Classroom Assessment Knowledge” scale (see Appendix A) were adapted from Mertler and Campbell’s (2005) Assessment Literacy Inventory (ALI). ALI had been reported to have a reasonable internal consistency reliability ($\alpha = .74$). The current study adapted these 11 items out of 35 multiple-choice items from Mertler and Campbell’s (2005) Assessment Literacy Inventory. Such adaptation was due to these items related to a classroom-based assessment context and matched the study’s purpose. The other new 16 multiple-choice items were developed based on an extensive review of literature in both the general education and language education fields. Table 5.2 displays the nine standards and associated items within the Classroom Assessment Knowledge scale.

Table 5.2 Nine Standards and Associated Items within the Classroom Assessment Knowledge Scale

Standard	Source			Item number	
	AFT, NCME, & NEA (1990)	Expanded	Mertler & Campbell’s (2005) ALI	Adapted	Developed
1. Choosing Appropriate Assessment Methods	☑		☑		1, 10 & 19
2. Developing Assessment Methods	☑		☑	2	11 & 20
3. Administering, Scoring, and Interpreting Assessment Results	☑		☑	3, 12 & 21	
4. Developing Valid Grading Procedures	☑		☑	14 & 23	5
5. Using Assessment Results for Decision Making	☑		☑	4 & 13	22
6. Recognising Unethical Assessment Practices	☑		☑	17	8 & 26
7. Keeping Accurate Records of Assessment Information		☑			6, 15 & 24
8. Ensuring Quality Management of Assessment Practices		☑			9, 18 & 27
9. Communicating Assessment Results	☑		☑	16 25	7

The Innovative Methods Scale

The Innovative Methods (IM) scale consisting of nine items was designed to address the extent to which the instructors endorsed the use of innovative methods in assessing students' learning. Respondents were asked to indicate their agreement to a four-point rating scale, varying from "Not Useful at all" to "Very Useful" with regard to the use of nine different assessment methods (oral presentation to self-assessment method). Figure 5.2 displays the items within the Innovative Methods scale.

To what extent is each of the following assessment types/methods useful in assessing students' learning?

1. Self-assessment
 2. Portfolio
 3. Peer assessment
 4. Individual conference
 5. Reflective journal
 6. Individual assignment/ project work
 7. Assessments that resemble the English language use in your students' real life situations
 8. Assessments that provide regular feedback indicating the ways to improve your students' future performance
 9. Oral presentation
-

Figure 5.2 Items within the IM scale

The Grading Bias Scale

The Grading Bias (GB) scale was designed to address the extent to which the instructors believed they were influenced by their students' personal characteristics such as age, gender, appearance, behaviour, attitude, effort and general abilities when marking students' work/performance. Respondents were asked to indicate their agreement to a four-point rating scale ranging from "Never" to "Always" with regard to students' personal characteristics they perceived frequently influenced their marking of the students' work/performance. Figure 5.3 displays the items within the Grading Bias scale.

Which of the following characteristics of your students influence you when marking their work (i.e., essays/assignments/presentations)?

1. Age
 2. Gender
 3. Appearance
 4. Behaviour
 5. Attitude
 6. General abilities
 7. Effort
-

Figure 5.3 Items within the GB scale

The Quality Procedure Scale

The Quality Procedure (QP) scale addressed the extent to which the instructors endorsed the use of procedures designed to enhance the quality of the assessment process. It comprised six items covering various issues associated with maintaining assessment records, implementing quality assurance mechanisms and communicating feedback to students in a timely and effective way. Respondents were asked to indicate their agreement to a four-point rating scale varying from “Strongly Disagree” to “Strongly Agree”. Figure 5.4 displays the items within the Quality Procedure scale.

To what extent do you agree with each of the following assessment quality procedures?

1. It is just as important to maintain detailed records of the assessment process as it is to maintain records of students’ results.
 2. It is important to construct accurate reports about students’ achievement for communicating to both students and administrators.
 3. It is important to employ various methods to record students’ achievement.
 4. It is important for instructors to gather together regularly to design and check the quality of the assessment process and results.
 5. It is my responsibility to ensure that my assessments are valid and reliable before using them.
 6. It is important to provide students with their assessment results in a timely and effective way.
-

Figure 5.4 Items within the QP scale

5.2.4 Quantitative Data Analysis

5.2.4.1 Item Response Modelling Procedure

Each scale developed to measure the classroom assessment literacy of instructors was calibrated using item response modelling (IRM) (Rasch, 1960). Item response modelling procedures can generate both item difficulty/parameter and person/respondent ability/perception on the same measurement scale. Thus, the same measurement scale can be employed to refer to item difficulty/parameter and person ability/perception.

The Rasch Simple Logistic Model (Rasch, 1960) for dichotomous items and the Partial Credit Model (Masters, 1982) for polytomous items were employed to analyse each of the scales using ConQuest software version 2.0 (Wu, Adams, Wilson, & Haldane, 2007). The main underlying assumption for the Rasch Simple Logistic Model is a probabilistic model of correct/incorrect answer whilst the Partial Credit Model (PCM) models the degree of correctness/endorsement in answering a question. Under the Rasch models, it is assumed that people with high ability are expected to score higher than those with low ability for an item. It is noted, however, that the Rasch Simple Logistic Model is a special case of the PCM. As such, both Rasch Simple Logistic Model and PCM can be carried out in one analysis (Wu & Adams, 2007).

The Fit of the Model

Two measures employed to assess how well the test/scale was constructed were the standard error of measurement and the fit of the data/items to the Rasch model. The standard error of measurement provides the information on the precision of the item difficulty/parameter estimates. The standard error of measurement for each item was calculated by estimating the difference between the true item difficulty/parameter and the estimated item difficulty/parameter using responses of all respondents to that particular item (Wright & Stone, 1979). The fit of the data/items refers to the extent to which the data/items fit the Rasch model. Thus, fit statistics are useful for examining the degree to which the model can predict the responses.

The most common suggestion for examining how the data/items fit the Rasch model has been the OUTFIT and INFIT mean squares statistics (Bond & Fox, 2007; de Ayala, 2009; Boone, Staver, & Yale, 2014). These statistics were developed by Wright and Masters (1982) based on the work of Wright and Panchapakesan (1969). OUTFIT and INFIT are also known as UNWEIGHTED and WEIGHTED statistics. The INFIT/WEIGHTED mean square has been referred to as the square of the standardised residuals that are weighted by the variance of the item response (Wu & Adams, 2007).

The OUTFIT/UNWEIGHTED and INFIT/WEIGHTED statistical value varies from zero to infinity. When the item/data fits the Rasch model, both OUTFIT/UNWEIGHTED and INFIT/WEIGHTED statistics have the expectation of 1.00. That is, the expectation of these statistics is 1.00 when the items/data fit the Rasch model (Wu & Adams, 2007). As such, recommendations in relation to the acceptable ranges of item mean-square fit statistics have been proposed by several researchers in the literature. For instance, Wright and his colleagues (1994) recommend the reasonable ranges for INFIT and OUTFIT of multiple-choice test (high stakes) as between 0.8 to 1.2 and the rating scale (Likert/questionnaire) as between 0.6 to 1.4. Adams and Khoo (1996) also suggest a rule of thumb for acceptable ranges of INFIT and OUTFIT between 0.7 and 1.3. de Ayala (2009) and Boone, Staver, and Yale (2014) further suggest the reasonable values for INFIT and OUTFIT, ranging between 0.5 to 1.5. As fit statistics are dependent on the sample size, it is suggested that fit statistics should not be employed solely for accepting or rejecting items (Bond & Fox, 2007; Wu & Adams, 2007). The discrimination and reliability indices from the Classical Test Theory should be used in conjunction with fit statistics to make an assessment of the items. The discrimination index refers to the correlation between the person's score on an item with the person's total score on the test/scale (Wu & Adams, 2007). The discrimination and reliability indices from Classical Test Theory can be obtained from the ConQuest software version 2.0 (Wu et al., 2007), employed for the analysis of the current study.

Establishing Reliability and Validity Employing Item Response Modelling

Under the Classical Test Theory, the most common reliability measure is the use of indices of internal consistency coefficient alpha named after its developer (Cronbach, 1951). The internal consistency coefficient alpha has been defined as the interrelatedness of a set of items within a test/scale (Cortina, 1993; Schmitt, 1996). Internal consistency reliability, however, is not an independent measure of homogeneity/uni-dimensionality, despite it being a necessary condition for homogeneity (Cortina, 1993; Clark & Watson, 1995; Schmitt, 1996). Generally, the internal consistency reliability can be affected by the length of the test (Lord & Novick, 1968; Cortina, 1993; Clark & Watson, 1995; Schmitt, 1996; Schmidt & Embretson, 2013). That is, when the test is short, it is more likely that its internal consistency reliability is low. Furthermore, the reliability estimation (i.e., Cronbach's alpha) depends on the interpretation for the entire test/scale score (Cortina, 1993; Wu & Adams, 2007). It does not specifically provide information on how well the individual items are measured within the test/scale.

In contrast, under the Item Response Theory, the reliability estimation is undertaken through the Rasch reliability estimate or person separation index (Wright & Masters, 1982). Similar to the Classical Test Theory measures of reliability (i.e., Cronbach's alpha), the Rasch reliability estimate or person separation index varies from 0 to 1. A separation index of 0 suggests that the persons/respondents cannot be separated, whereas the value of 1 for a separation index indicates that the persons/respondents can be well separated along the variable. The reliability estimates for both Cronbach's alpha and person separation indices can be obtained from the ConQuest software version 2.0 (Wu et al., 2007).

5.2.4.2 Structural Equation Modelling Procedure

A confirmatory factory analysis (CFA) using a structural equation modelling (SEM) approach was employed to evaluate the fit of a hypothesised one-factor congeneric measurement model to the data. The one-factor congeneric measurement model comprises one Classroom Assessment Knowledge variable and three variables of personal assessment beliefs that have been hypothesised to be the underpinnings of

instructors' classroom assessment literacy. The current study employed IBM SPSS Amos software version 20 for the CFA analysis.

Steps for Undertaking CFA Employing SEM

Researchers (Bollen & Long, 1993; Byrne, 2010; Schumacker & Lomax, 2010) have generally agreed there are seven common steps in conducting a CFA using a SEM approach as follows:

1. Model conceptualisation;
2. Path diagram construction;
3. Model specification;
4. Model identification;
5. Parameter estimation;
6. Assessment of model fit; and
7. Model re-specification.

Each step will be considered next.

Model Conceptualisation

Model conceptualisation involves two distinct components: a measurement model and a structural model (Jöreskog & Sörbom, 1989; Bollen, 1989; Diamantopoulos & Siguaw, 2000; Byrne, 2010; Blunch, 2013). The measurement model, which is also known as a confirmatory factor analysis, has an affinity with factor analysis in that it illustrates indicator variables, measured with error, as influencing the underpinning latent variable. However, the structural model has an affinity with path analysis in that it shows linear relationships amongst the latent variables, which can be presented in the form of path diagrams and associated path coefficients. Hence, the measurement model is concerned with how well the indicator variables measure the latent variable, whilst the structural model emphasises relationships amongst latent variables and their explanatory power.

It has been suggested that the measurement model needs to be tested before the structural model being tested in order to certify that the indicator variables are valid and reliable (Jöreskog & Sörbom, 1993; Kline, 2011). It is also recommended that when the indicator variables produce a poor fit for a construct, the modification of the proposed theory must be done prior to testing it. It is further suggested that the researcher should test the measurement models separately for each construct involved in a structural model. Then the researcher should test two constructs at a time and eventually test all constructs simultaneously. Moreover, it is advised that the constructs themselves should be permitted to freely correlate (Jöreskog & Sörbom, 1993). That is, the covariance matrix of the constructs should not be constrained.

Path Diagram Construction and Model Specification

Path diagram construction refers to drawing a pictorial representation to depict the underlying variables in a model and the relationships amongst them. Model specification, however, refers to writing coded instructions for a SEM programme so that the model represented in the path diagram can be estimated correctly by the SEM programme being employed (Diamantopoulos & Siguaw, 2000). Model specification in AMOS is conducted by employing either syntax input (using Programme Editor) or by drawing a path diagram as input (using AMOS Graphics). Hence, path diagram construction and model specification are synonymous in AMOS (Byrne, 2010; Blunch, 2013).

To carry out path diagram construction and model specification in SEM, it requires a priori specification of the model comprising propositions, which have stemmed from past research or theory (Diamantopoulos & Siguaw, 2000; Byrne, 2010). The propositions are expressed as mathematical equations and illustrated in path diagrams. The equations define the set of relationships amongst the variables of interest expected to be found in the data. Thus, it is a confirmatory rather than exploratory technique of modelling that evaluates the fit of the theoretically based propositions to the data.

In the present study, one measurement model had been hypothesised to explain the interrelationships amongst the four variables that tapped into a latent construct of classroom assessment literacy. This measurement model was formulated as a one-factor congeneric measurement model. This model was considered to be defined by the

interrelationships amongst the four variables of classroom assessment literacy. A CFA using SEM approach was employed to evaluate whether or not this hypothesised model fitted the data.

One-factor Congeneric Measurement Model: Classroom Assessment Literacy

A one-factor congeneric measurement model was formulated for the current study. There is a growing body of research that suggests the classroom assessment literacy is underpinned by two distinct domains (i.e., Classroom Assessment Knowledge base and Assessment Beliefs). This study sets out to empirically determine whether these two constructs (i.e., knowledge and beliefs) load on a common Classroom Assessment Literacy factor. Hence, a parsimonious one latent construct was hypothesised to be responsible for all observed indicator variables of classroom assessment literacy. The variable names have been abbreviated as follows: IM= Innovative Methods, GB= Grading Bias, QP= Quality Procedure and CAK= Classroom Assessment Knowledge. The name of the construct of classroom assessment literacy has also been abbreviated: CAL= Classroom Assessment Literacy. A one directional arrow suggests that the variation in the variable pointed to is explained by the variation of the second order variable from which the arrow is linked (see Figure 5.5).

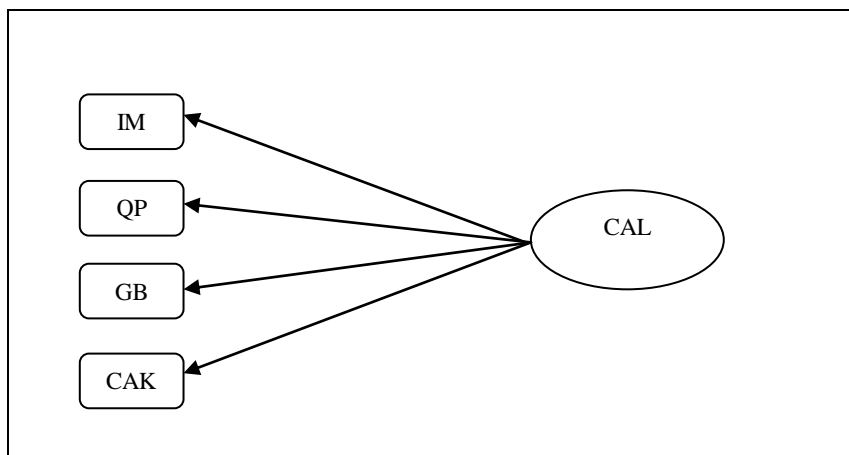


Figure 5.5 One-factor congeneric measurement model: Classroom Assessment Literacy

Model Identification

In SEM, the model identification requires the estimation of unknown parameters (i.e., factor loadings) based on the observed covariances/correlations. This identification is relevant to whether a unique solution for the model can be received (Diamantopoulos & Siguaw, 2000). As such, model identification relies on the number of parameters to be estimated. Models can be categorised as just identified, over identified, or under identified (Mueller, 1996; Kelloway, 1998; Byrne, 2010; Schumacker & Lomax, 2010; Kline, 2011; Ullman & Bentler, 2013). A just identified model, also known as a saturated model, exists when the number of parameters to be estimated is exactly equal to the number of equations. And hence, there is just one unique solution that can reproduce the covariance or correlation matrix. An over identified model, however, occurs when the number of equations is greater than the number of unknown parameters to be estimated, and it therefore generates a number of unique solutions that can provide the best fit to the data. Conversely, in an under identified model, the number of unknown parameters to be estimated exceeds the number of equations, and hence, a unique solution cannot be determined. In SEM, the requirement of a model specification is either just identified or over identified. The over identified model, however, is a preferred specification as it can generate fit indices.

Parameter Estimation

As the model has been specified and checked for identification, the next step is to calculate the estimates for the model parameters. A number of estimation procedures can be employed to solve the equations specified in the model including Maximum Likelihood (ML), Ordinary Least Squares (OLS), Generalised Least Squares (GLS), Weighted Least Squares (WLS) and Unweighted Least Squares (ULS) (Bollen, 1989; Jöreskog & Sörbom, 1989). The most commonly employed method in the literature is the ML estimation procedure (Schumacker & Lomax, 2010; Blunch, 2013).

The underlying assumption for the ML method of estimation is multivariate normal distributions of observed variables. Furthermore, ML requires adequate sample size and interval scales of measurement. It is recommended that product-moment

correlations be computed and ML estimation be employed to solve the mathematical equations employed to define the model when the sample size is sufficient, interval scales are used and multivariate normal distribution is met (Schumacker & Lomax, 2010; Kline, 2011). Kline (2011) and Blunch (2013) recommend that sample size should have a minimum of 100 for parameter estimations. In the current study, the sample size was 108 and the observed variables (i.e., IM, QP, GB and CAK) were constructed using item response modelling procedures in which all variables were measured on the Rasch scale. Hence, these variables were interval measures. Given each variable was continuous and had a multivariate normal distribution (refer to section 7.1.2 in Chapter 7), the ML method was thought to be the most appropriate estimation technique to be employed.

Assessment of Model Fit

Once a model has been specified, checked for identification and the parameters have been estimated, the next step is to evaluate whether or not the model fits the data. SEM does not have a single test of significance that identifies an adequate model fit to the data. There is a number of goodness of fit indices reported by SEM computer packages that can be used to assess model fit. Goodness of fit indices are employed to estimate how closely the model can explain variance in the data. The fit statistics can be divided into two types: fit statistics and comparative fit statistics (Kelloway, 1998; Schumacker & Lomax, 2010; Kline, 2011). Most fit indices compare the observed covariance/correlation matrix to the modelled matrix. A number of goodness of fit indices (Hu & Bentler, 1999; Schumacker & Lomax, 2010; Arbuckle, 1983-2011) can be selected for evaluation of model fit and each is displayed in Table 5.3.

Table 5.3 Goodness of Fit Criteria and Acceptable Level and Interpretation

Goodness of Fit Criteria	Feature	Acceptable Level and Interpretation
Fit Statistics Chi-square (χ^2)	χ^2 is a statistical test of significance that is employed for an assessment of the null hypothesis (i.e., there is no difference between the theoretical specified model and the data). It assesses the overall fit of the model to the data and it depends on sample size.	A non-significant chi square value ($p > .05$) indicates that the model fits the data well.
χ^2/df ratio	The χ^2/df ratio is employed as alternative indices of χ^2 as it does not depend on sample size.	χ^2/df ratio should not exceed 2
Goodness of Fit (GFI)	The GFI is based on the ratio of the sum of squared discrepancies of the observed variables. It depends on sample size. It assesses how well the model fits in comparison to no model at all.	0 (no fit) to 1 (perfect fit) Value close to .95 indicates a good fit.
Adjusted Goodness of Fit (AGFI)	The AGFI adjusts the GFI for degree of freedom in the model. The AGFI does not rely on sample size. It assesses how well the model fits in comparison to no model at all.	0 (no fit) to 1 (perfect fit) Value close to .95 indicates a good fit.
Root Mean-Square Error of Approximation (RMSEA)	The RMSEA is a measure of the discrepancy per degree of freedom having first diminished the discrepancy function somewhat as a function of sample size.	Value of .05 to .08 indicates a good fit.
PCLOSE	The PCLOSE is a measure for testing the null hypothesis that RMSEA (in the population) is less than .05	Value PCLOSE > .05 indicates the close fit hypothesis
Standardised Root Mean-square Residual (SRMR)	The SRMR is a residual that is divided by its estimated standard error. A residual is an observed minus a fitted covariance.	Value SRMR < .05 indicates a good fit.
Comparative Fit Index Comparative Fit Index (CFI)	The CFI is based on the non-central χ^2 distribution (χ^2/df). It is constrained to fall between 0 and 1	Value close to .95 indicates a good fit.
Tucker-Lewis Index (TLI)	The TLI is another comparative fit indices that can exceed a value of one	Value close to .95 indicates a good fit. Value TLI > 1 indicates lack of parsimony.
Normed Fit Index (NFI)	NFI indicates the percentage improvement in fit over the baseline independence model	0 (no fit) to 1 (perfect fit) Value close to .95 indicates a good fit.
Parsimonious Fit Parsimonious Normed Fit Index (PNFI) and the Parsimonious Goodness of Fit Index (PGFI)	Both PNFI and PGFI emphasise the cost-benefit of fit and degree of freedom. They are both a modification of the NFI but take into consideration the number of degrees of freedom of fit. They are employed to compare two rival theoretical models with different degrees of freedom.	0 (no fit) to 1 (perfect fit)

Source: Hu & Bentler (1999); Schumacker & Lomax (2010); and Arbuckle (1983-2011)

Model Re-specification

The last step in undertaking a CFA using a SEM approach is the model re-specification if the model is not supported by the data. To determine how the model should be modified to fit the data better, Schumacher and Lomax (2010) recommend three main procedures for undertaking a specification search. They include an examination of the critical ratios, standardised residuals and modification indices. The critical ratio (also known as t-value) refers to a comparison of the ratio of the parameter estimate to its estimated standard error. The critical ratio should be larger than ± 1.96 at the $\alpha = .05$ significance level. Standardised residuals are the residual covariances that are divided by an estimate of their respective standard errors. Generally, large standardised residuals provide an indication of a poorly fitting model. For instance, standardised residuals that are larger than 1.96 or less than -1.96 suggest that a particular covariance is not well reproduced by the hypothesised model at the $\alpha = .05$ significance level. The modification index is relevant to the decrease in the value of the chi-square that would result if the suggested parameter was allowed to freely estimate in a revised model. The modification index is the most useful way to re-specify the hypothesised model (Byrne, 2010). It identifies not only the degree to which the hypothesised model can be improved, but also suggests where the point of most likely improvement. Generally, the parameter associated with the largest modification index should be estimated first as it can improve the fit most when permitted to be estimated. However, if the parameter with the largest modification index cannot be justified by the theoretical rationale, the second largest modification index should be considered and so forth.

In summary, the procedures used for re-specification are the critical ratio (t-value), the standardised residuals and modification indices. Each of these results can be used to identify the source of misspecification as well as to indicate how the model should be respecified to fit the data better. Once the model has an acceptable fit to the data, it cannot be concluded that it is the best model, since there could be other equivalent models. The covariance structure techniques, however, could be employed to explicitly differentiate between alternative models that fit the data badly and those that fit the data

well. Furthermore, the respecified model needs to be based on both theoretical rationale and data driven.

5.3 Part Three: Qualitative Phase

5.3.1 The Sample

The purpose of phase two was to undertake follow-up interviews with a small number of participants from phase one to shed further light on phase one findings. Given this purpose, six instructors were selected from the 33 instructors who volunteered to participate in phase one and phase two. As the number of instructors that volunteered for the semi-structured interviews was more than required, a stratified random sample was employed to select three English-major instructors and three English non-major instructors for interviews. These six instructors were selected based upon the following criteria: their self-reported measure of departmental status, gender, the level of their academic qualification, the number of years of their teaching experiences (i.e., less experienced, more experienced and most experienced instructors) as well as the level of their classroom assessment knowledge (i.e., Band levels 1, 2, 3 and 4 on the Classroom Assessment Knowledge Test obtained from phase one of the study) (refer to Table 6.2 in Chapter 6). Table 5.4 shows the table of specifications regarding the number of individuals who were within each of the characteristics specified in the criteria.

Table 5.4 Table of Specifications for Selecting Six Participants

Background Variable	Value	English-major Participant	English Non-major Participant
		Frequency	Frequency
Gender	Male	16	10
	Female	3	4
Academic qualification level	Bachelor	10	14
	Master	9	0
Years of teaching experience	Less experienced	0	0
	More experienced	9	9
	Most experienced	10	5
Classroom Assessment Knowledge level	Band level 1 (Novice)	0	4
	Band level 2 (Competent)	5	0
	Band level 3 (Proficient)	10	10
	Band level 4 (Expert)	4	0

It should be noted that on the condition that the individual instructor matched all the criteria, his/her code was put into the two boxes for random selection (i.e., boxes one and two were for the English-major and English non-major departments respectively). There were nine English-major and seven English non-major participants whose codes were used for a random selection. Three English-major participants (i.e., two males and one female) and three English non-major participants (i.e., one male and two females) were randomly selected from the two boxes for the interviews (refer to Table 8.2 in Chapter 8 for detailed background characteristics of these selected participants).

5.3.2 Data Collection Procedures

5.3.2.1 Departmental Learning Goals and Assessment-related Policies

Prior to undertaking the semi-structured interviews with the selected instructors, documents of the learning goals and assessment-related policies of the two departments were collected for analyses. The aim of these document analyses was to examine the relationship between departmental learning goals and assessment-related policies and the instructors' actual assessment practices.

5.3.2.2 Interview Administration

Semi-structured one-on-one interviews of 45 to 60 minutes were conducted with each of the six instructors. The interviews were conducted six months after the test and questionnaire administration at the university during January to February 2012. The interview schedules were based on the availability of each interviewee to suit his/her schedule. An informal conversational style of interview was used to permit in-depth interactions between the interviewee and the interviewer in a relaxed atmosphere (Kvale, 1996). Each interviewee was encouraged to share his/her thoughts on classroom assessment knowledge and assessment beliefs. Despite the fact that questions were prepared in advance, additional probing questions were also employed to follow up with what each interviewee had just said in order to gain in-depth understandings of the issues.

Brief notes were also used during the interview, comprising short phrases. With the permission of interviewees, all interviews were digitally recorded using a voice recorder. Each interview was then transcribed and reviewed by the interviewer for accuracy and content (Creswell, 2012). All interviews were conducted in English, given that interviewees were instructors of English language subjects.

A number of open-ended questions had also been developed for the semi-structured interviews in order to obtain an in-depth understanding of the instructor's classroom assessment literacy development (as previously measured in phase one). Similar to the phase one instruments (i.e., the test and questionnaire), all open-ended questions were panelled by the same two groups of six specialists in the areas of EFL/ESL and/or general classroom-based assessment. As mentioned earlier, the questions were panelled in both Australia and in Cambodia with such specialists. The panellists were requested to check the appropriateness of the questions and underpinning constructs and their match to the research questions (i.e., content validity).

Prior to the main interview data collection, a pilot study with two instructors was conducted to ensure the research process and the data collection procedures were appropriate and could achieve the desired outcomes (Glesne, 2011; Creswell, 2012). That is, the interview questions pilot was to gain insights with regard to the clarity of the questions and other unforeseen issues during the interview process. The most common feedback received was associated with word-choice of questions. This feedback was used to revise the questions for the main semi-structured interviews. These revisions helped to enhance the face and content validity of the questions employed.

5.3.3 Interview Questions Development Processes

5.3.3.1 Interview Questions

The interview questions were developed based on the results obtained in the quantitative phase. The issues to be explored in the qualitative phase were relevant to the instructors' prior assessment experience as students, their assessment training during pre-service teacher education programme, perceived level of classroom assessment competence, perceived ideal assessment, knowledge and understanding of the concepts of

“validity” and “reliability”, methods they employed to enhance validity and reliability of their assessments, as well as the decision-making processes employed for grading students. Figure 5.6 displays the questions employed for interviews.

-
1. Describe the way in which you were typically assessed when you were studying to be a teacher. To what extent do you think your experience as a student has influenced the way in which you assess your students as an instructor?
 2. Tell me about your previous assessment training? To what extent has this training sufficiently prepared you for assessing your students?
 3. If you had to score yourself out of 10 on your knowledge and skills in assessment, how would you score yourself? On this scale, a 10 would mean that you had mastered all the knowledge and skills required to conduct good quality assessments, whilst a score of 1 would mean that you still have an enormous amount to learn. Please explain why you gave yourself this score. To what extent do you think your current assessment knowledge influences your assessment practices?
 4. What are the main factors (e.g., organizational, personal and student related) that influence the way in which you design and conduct your classroom assessments? These may include access to learning and teaching resources, opportunities for professional development workshops, workload issues and student motivation. What suggestions would you make to try to address some of these issues?
 5. In the ideal world, where you had the time and resources to design and administer a good quality assessment, what would this assessment look like? Why do you think this would be the ideal assessment? How does this ideal assessment compare to how you currently conduct assessment?
 6. What do the terms “validity” and “reliability” mean to you? How do you try to check that your assessments are valid and reliable? How could you do things differently to improve the validity and reliability of your assessments? Why is it important to enhance the validity and reliability of your assessments?
 7. How do you determine your course grades? To what extent do you think your course grades reflect your students’ learning achievements? How are your course grades used by yourself and others?
-

Figure 5.6 Interview questions

5.3.4 Qualitative Data Analysis

Content analysis was employed to analyse departmental learning goals and assessment-related policies, whilst a thematic analysis was used to analyse the interview transcripts in order to obtain in-depth meaning of participants’ viewpoints (Miles & Huberman, 1994; Patton, 2002; Gibbs, 2007; Yin, 2009). Content and thematic coding involving a combination of inductive coding and deductive coding techniques was used

for the analyses. Inductive coding technique refers to the process that permits content/themes to emerge directly from the data. However, deductive coding technique is relevant to the process in which predetermined codes are derived from the theoretical framework used to generate content/themes from the data (Miles & Huberman, 1994; Fereday & Muir-Cochrane, 2006). Prior to undertaking coding, there was a preliminary reading of departmental learning goals and assessment related policy documents, interview transcripts and written memos to better understand the data. Inductive coding techniques were then employed for the analysis, in which each of the departmental learning goals and assessment related policy documents and interview transcripts were read and reread. This was undertaken to identify and label recurring content/themes and/or patterns through focusing on key aspects of the learning goals and assessment-related policies and interviewees' words, phrases, sentences and/or paragraphs. In addition to conducting inductive coding analysis, deductive coding analysis was employed to identify and label content/themes based on the predetermined factors derived from the theoretical framework. Great care was also taken to avoid forcing data into a priori thematic categories, given codes existed for them (Miles & Huberman, 1994). Both inductive and deductive analyses were conducted manually, given the small number of documents and transcripts. Despite manual analysis being time-consuming, it enabled the researcher to gain an intimate understanding of the content/themes emerging (Creswell, 2012). Based on the codes, the descriptions and content/themes for each of the learning goals and assessment-related policies and participants' interviews were generated. As an outcome of the comparison of departmental assessment-related policies, a table was created. To further compare themes within and across two groups of instructors, tables of matrices were created for English-major and English non-major instructors separately (Glesne, 2011). Each table covered participant codes, themes, sub-themes and some quotations relevant to themes/sub-themes (see Appendix C).

As with the quantitative data analysis, the reliability and validity issues needed to be addressed in analysing the qualitative data in order to ensure the accuracy and consistency of the results obtained (Burns, 2000; Silverman, 2013; Creswell, 2014). Some qualitative researchers tend to use the term "trustworthiness" to refer to both "reliability" and "validity" aspects (Dick, 1990; Gillham, 2000; Glesne, 2011). Other

qualitative researchers, however, have argued that validity plays a more important role than reliability in ensuring trustworthiness and credibility in qualitative research (Creswell & Plano Clark, 2011). Reliability is associated with multiple coders as a team to reach agreement on codes used, whilst validity is relevant to accurate information obtained through the data collection processes and analyses.

To address the reliability issue, with the participants' permission, all interviews were digitally recorded to capture all conversations, and notes were taken by the interviewer to allow for analyses and further probing as the interview progressed. Thus, these considerations increased the reliability of the data collection (Glesne, 2011; Creswell, 2014). In ensuring consistency of the analyses, intercoder reliability was used. Intercoder reliability has been defined as the process in which two or more coders code the documents and/or transcripts independently and compare the agreement on coding used (Miles & Huberman, 1994). To analyse the qualitative data, two independent coders first coded and analysed all learning goals and assessment related policy documents and interview transcripts for content/themes individually using inductive and deductive techniques. When the coding procedure was completed, the two coders met to review discrepancies and resolve differences through discussion and negotiated consensus. These two coders then compared the content/themes. In both comparisons, 85% coding agreement was initially achieved between the two coders, and following their discussion and negotiated consensus, the coding agreement increased to 90%, suggesting reasonable intercoder reliability. The 90% coding agreement was above the rule of thumb for reasonable intercoder reliability of 80% agreement proposed by Miles and Huberman (1994).

In addressing validity issues, the bias the researcher was likely to bring to the study (Creswell, 2013) needed to be clarified, as the researcher played a participatory role through her engagement and involvement with participants during data collection and analyses. As the researcher had previously worked in the two selected English departments, the participants could feel uncomfortable or refuse to participate in the study. Furthermore, they could feel uncomfortable talking about their current classroom assessment competence as well as expressing their actual assessment beliefs. Some could feel their career opportunities were at risk through their participation, especially if they

made comments about departmental administrators and/or heads of department. However, by explicitly explaining the voluntary nature, confidentiality, approval for the study from the Human Research Ethics Committee, and the purpose of the research to the participants, as well as the fact that the researcher was on PhD study leave, participants' concerns about relationships and confidentiality should have been minimised.

It is also widely acknowledged that the researcher's active participatory role in the research process has the potential to allow her own personal views or presuppositions to influence the results of the study (Burns, 2000). This potential for subjectivity, however, could be minimised through the integration of inductive and deductive coding analyses. These coding analyses are intended to ensure the trustworthiness of the results obtained, as they complemented each other and could be used to verify the results obtained from each analysis (Fereday & Muir-Cochrane, 2006; Bradley, Curry, & Devers, 2007). This combination of inductive and deductive techniques is also known as triangulation (Miles & Huberman, 1994) of analysis methods. In addition to the integration of inductive and deductive coding, comparisons across departmental learning goals, assessment related policy documents and participants' themes were conducted to provide valid evidence of the content/themes generated (Burns, 2000). Given the researcher developed and created the data through her engagement and involvement with the documents and participants during the interview process, she would remain honest to the voices of the interviewees and documents when analysing data (Patton, 2002). Moreover, a constant and careful auditing (Creswell, 2014) carried out by the researcher's academic supervisors on all research processes and data analyses in the study could further enhance validity. To further strengthen the validity and credibility of the results, the themes were endorsed by quotations from the raw data to ensure that data interpretation remained directly linked to the actual words of participants (Patton, 2002; Fereday & Muir-Cochrane, 2006).

In summary, the trustworthiness and credibility of the qualitative findings could be established by the use of several extensive procedures. One of the procedures was collecting departmental documents and employing a digital recorder to capture all conversations from interviews for transcription. The next procedure was using two independent coders to code and analyse the learning goals and assessment related policy documents and interview transcripts for content/themes, and the triangulation of

inductive and deductive coding analysis methods. The last procedure was employing the illustrations of quotations from the raw data to ensure data interpretation remained directly linked to the actual words of participants, minimising the researcher's bias brought to the study. Careful auditing was conducted by the researcher's academic supervisors on all research processes and data analyses in the study. The next chapter documents the development and validation of a set of scales to measure classroom assessment literacy progression of instructors.

Chapter 6: Scale Development Processes

This chapter presents the processes employed for developing the four scales in the current study. All scales were calibrated using the ConQuest software version 2.0 (Wu et al., 2007). The purposes of analyses were to evaluate the measurement properties of these scales; examine the item statistics; and compute psychometric estimates of the instructors' classroom assessment knowledge and assessment beliefs (i.e., innovative methods, grading bias and quality procedure). Section 6.1 describes the process for developing the four scales and discusses the properties of each scale calibration. Section 6.2 provides the summary statistics of scales.

6.1 Development of the Scales

The four scales developed for the present study comprised "Classroom Assessment Knowledge (CAK)," "Innovative Methods (IM)," "Grading Bias (GB)" and "Quality Procedure (QP)". The underlying assumption for the development of these scales was that responses to the subset of items were determined by the respondents' positions on a series of latent continua. The following sections describe the development process of each item set for the four scales.

6.1.1 Development of the Classroom Assessment Knowledge Scale

The Classroom Assessment Knowledge (CAK) scale comprised 27 multiple-choice items with three items related to each of the nine standards (refer to section 5.2.3.1 in Chapter 5). The test was constructed to measure the assessment knowledge of instructors in assessing the student's learning. Figure 6.1 illustrates the nine standards and associated items within the CAK scale.

-
1. Choosing Appropriate Assessment Methods (1, 10, & 19)
 2. Developing Assessment Methods (2, 11, & 20)
 3. Administering, Scoring, and Interpreting Assessment Results (3, 12, & 21)
 4. Developing Valid Grading Procedures (5, 14, & 23)
 5. Using Assessment Results for Decision Making (4, 13, & 22)
 6. Recognising Unethical Assessment Practices (8, 17, & 26)
 7. Keeping Accurate Records of Assessment Information (6, 15, & 24)
 8. Ensuring Quality Management of Assessment Practices (9, 18, & 27)
 9. Communicating Assessment Results (7, 16, & 25)
-

Figure 6.1 Nine standards and associated items within the Classroom Assessment

Knowledge scale

Based on the results from the item analyses of these 27 MCQ items for the Classroom Assessment Knowledge scale, three items (Q1, Q3 and Q27) showed poor discrimination index (see Figure 6.2 for the detail of these three item analyses). The discrimination index refers to the correlation between the respondent's score on the item and his/her total scores on the scale. If the discrimination value is 0, it suggests that there is no correlation between the item score and the total scores. Hence, the higher the discrimination value, the better the item is able to separate respondents according to their level located on the measurement scale (Wu & Adams, 2007). These three items did not have discriminating power in separating respondents positioned between low and high on the Classroom Assessment Knowledge scale. As such, they were deleted from the Classroom Assessment Knowledge scale (see Figure 6.2). Item Q1 was associated with standard 1 (i.e., Choosing Appropriate Assessment Methods) whereas item Q3 related to standard 3 (i.e., Administering, Scoring, and Interpreting Assessment Results). Unlike its counterparts, item Q27 was associated with standard 8 (i.e., Ensuring Quality Management of Assessment Practices).

Item:1 (Q1)

Cases for this item 108 Discrimination 0.09
 Item Threshold(s): -1.14 Weighted MNSQ 1.06
 Item Delta(s): -1.14

Label	Score	Count	% of total	Point-biserial	T value (probability)
A	0.00	18	16.67	0.04	0.37(.709)
C	0.00	8	7.41	-0.20	-2.06(.042)
D	1.00	82	75.93	0.09	0.91(.364)

Item:3 (Q3)

Cases for this item 108 Discrimination 0.09
 Item Threshold(s): 0.73 Weighted MNSQ 1.10
 Item Delta(s): 0.73

Label	Score	Count	% of total	Point-biserial	T value (probability)
A	0.00	34	31.48	0.03	0.29(.769)
B	0.00	8	7.41	-0.17	-1.80(.074)
C	0.00	26	24.07	-0.02	-0.24(.807)
D	1.00	40	37.04	0.09	0.90(.370)

Item:27 (Q27)

Cases for this item 108 Discrimination 0.08
 Item Threshold(s): 1.47 Weighted MNSQ 1.10
 Item Delta(s): 1.47

Label	Score	Count	% of total	Point-biserial	T value (probability)
A	0.00	14	12.96	-0.05	-0.50(.620)
B	1.00	25	23.15	0.08	0.81(.418)
C	0.00	50	46.30	0.12	1.20(.231)
D	0.00	19	17.59	-0.20	-2.07(.041)

Figure 6.2 Detail of three item analyses

Separate from the deletion of items 1, 3 and 27 from the Classroom Assessment Knowledge scale, two options within items 7 and 8 were accepted as the correct answers, while only one option for each of the other items was regarded as the correct answer. As Figure 6.3 shows, within item 7, options B and C were collapsed together as these two options contained key information in providing a reasonable explanation “*for assigning course grade by Mr. Chan Sambath*”. Similarly, options C and D of item 8 were collapsed into one option as the information given in the question did not mention whether it was the open book examination or the closed book examination. Therefore, option C was also a possible answer for this question.

-
7. In a routine conference with his students, Mr. Chan Sambath is asked to explain the **basis** for assigning his course grade. Mr. Chan Sambath should
 - A. explain that the grading system was imposed by the school administrators.
 - B. refer to the information that he presented to his students at the beginning of the course on the assessment process.
 - C. re-explain the students the way in which the grade was determined and show them samples of their work.
 - D. indicate that the grading system is imposed by the Ministry of Education.

 8. Mr. Chan Sambath was worried that his students would not perform well on the semester examination. He did all of the following to help increase his students' scores. Which was **unethical**?
 - A. He instructed his students in strategies for taking tests.
 - B. He planned his instruction so that it focused on concepts and skills to be covered on the test.
 - C. He allowed his students to bring in their coursebooks/materials to refer to during the test
 - D. He allowed students to practice with a small number of items from the actual test.
-

Figure 6.3 Items 7 & 8

The results of the calibration for the finalised 24-item CAK scale obtained from both Classical and Rasch analyses are illustrated in Table 6.1. For each item, the summary statistics are displayed as follows. The facility refers to the proportion of the respondents who obtained the correct answer for the item. Next, the point-biserial refers to the correlation between the item score and the scale total scores. The item difficulty is presented next, which is reported as the Logit value. The standard error in relation to each item difficulty estimates (SE) is also reported. Finally, the INFIT mean square (INFIT MNSQ) is presented, showing the extent to which the pattern of item responses fit the Rasch model. The 24 items within this scale are presented in Table 6.1 according to decreasing item difficulty.

Table 6.1 Calibration Estimates for the Classroom Assessment Knowledge Scale

Item No.	Standard	Facility	Point-biserial	Logit	SE	INFIT MNSQ
Q6	Keeping Accurate Records of Assessment Information	0.22	0.21	1.41	0.24	1.02
Q15	Keeping Accurate Records of Assessment Information	0.24	0.16	1.30	0.24	1.05
Q24	Keeping Accurate Records of Assessment Information	0.30	0.26	0.98	0.22	1.02
Q22	Using Assessment Results for Decision Making	0.33	0.35	0.78	0.22	0.96
Q16	Communicating Assessment Results	0.34	0.29	0.74	0.22	1.01
Q20	Developing Assessment Methods	0.35	0.22	0.69	0.22	1.03
Q13	Using Assessment Results for Decision Making	0.38	0.31	0.56	0.21	0.98
Q11	Developing Assessment Methods	0.40	0.27	0.47	0.21	1.02
Q9	Ensuring Quality Management of Assessment Practices	0.44	0.21	0.25	0.21	1.05
Q19	Choosing Appropriate Assessment Methods	0.45	0.16	0.21	0.21	1.08
Q21	Administering, Scoring, and Interpreting Assessment Results	0.48	0.21	0.08	0.21	1.04
Q25	Communicating Assessment Results	0.48	0.37	0.08	0.21	0.96
Q14	Developing Valid Grading Procedures	0.56	0.50	-0.30	0.21	0.87
Q23	Developing Valid Grading Procedures	0.58	0.38	-0.39	0.21	0.94
Q4	Using Assessment Results for Decision Making	0.59	0.34	-0.43	0.21	0.97
Q18	Ensuring Quality Management of Assessment Practices	0.59	0.16	-0.43	0.21	1.10
Q10	Choosing Appropriate Assessment Methods	0.66	0.14	-0.74	0.22	1.10
Q17	Recognising Unethical Assessment Practices	0.67	0.24	-0.79	0.22	1.03
Q12	Administering, Scoring, and Interpreting Assessment Results	0.68	0.22	-0.84	0.22	1.01
Q5	Developing Valid Grading Procedures	0.73	0.31	-1.14	0.23	0.99
Q2	Developing Assessment Methods	0.74	0.26	-1.19	0.23	0.98
Q7	Communicating Assessment Results	0.82	0.37	-1.73	0.26	0.91
Q8	Recognising Unethical Assessment Practices	0.84	0.45	-1.88	0.28	0.84
Q26	Recognising Unethical Assessment Practices	0.87	0.42	-2.12	0.30	0.89

Table 6.1 illustrates that the item difficulty estimates varied from -2.12 to +1.41 logits, with a range of 3.53 logits. The item difficulty estimates are an indication of the difficulty/demand of items. The range of item difficulty estimates demonstrates the levels of respondents' assessment knowledge that the CAK scale is suitable to measure. The mean item difficulty was constrained to zero. The standard deviation of the item difficulty level was 0.98 and the person separation reliability was .74. This measure of CAK scale reliability determines how sufficiently well separated the respondents were in terms of their levels within the latent construct.

The instructors' assessment knowledge estimates varied from -2.35 to +2.36 logits, with a range of 4.71 logits. Item 6 "*Mr. Chan Sambath is planning to keep assessment records as a part of his assessment and reporting process. Which of the following is the least important assessment information to be recorded?*" was the most difficult item within the CAK scale with 22% of the sample obtaining the correct answer.

This item related to standard 7 (i.e., Keeping Accurate Records of Assessment Information). In contrast, item 26 “*Prior to the semester examination, Mr. Keo Ratana reveals some information to his students. Which of Mr. Keo Ratana’s action was unethical?*” was the easiest item within the CAK scale with 87% of the sample obtaining the correct answer. This item related to standard 6 (i.e., Recognising Unethical Assessment Practices).

Item INFIT MNSQ values vary from 0.84 to 1.10. Adams and Khoo (1996) suggest a rule of thumb for the acceptable ranges of INFIT and OUTFIT between 0.7 and 1.3. Similarly, de Ayala (2009) and Boone, Staver, and Yale (2014) propose the reasonable ranges for INFIT and OUTFIT as between 0.5 and 1.5. Wright and his colleagues (1994) also recommend the reasonable ranges for INFIT and OUTFIT of multiple-choice test (high stakes) as between 0.8 to 1.2 and a rating scale (Likert/questionnaire) as between 0.6 to 1.4. Thus, these recommendations demonstrate that all 24 items within the CAK scale had an acceptable fit to the Rasch model. Furthermore, the standard errors of estimates for each of the items are acceptable, ranging from 0.21 to 0.30. The point biserial coefficient estimates for each item ranged from 0.14 to 0.50. The fit statistics and the point biserial coefficient estimates indicate that the set of items is measuring a dominant construct. There were no items with zero scores or perfect scores. Overall, the items formed a cohesive scale representing “Classroom Assessment Knowledge”.

The estimates of the person proficiency parameter and the item difficulty parameter were placed on a chart called a variable map. This variable map comprising three different sections has been displayed in Figure 6.4.

The person-item variable map shows the range of person proficiencies and item difficulties on the same scale. On the left of the map is the person proficiency distribution (shown with x). The person proficiency estimates in Figure 6.4 varied from -2.35 to +2.36, with a range of 4.71 logits. This broad range suggests that variation in the assessment knowledge of the instructors in the sample was quite large. On the right of the map is the item difficulty distribution, which shows each item located on the scale according to its difficulty measure. The item difficulty varied from -2.12 to +1.41, with a range of 3.53 logits. The person-item map shows there is a good match between the

instructors' proficiency levels and the range of item difficulty values, indicating that the test instrument has the appropriate level of difficulty to assess this group of instructors.

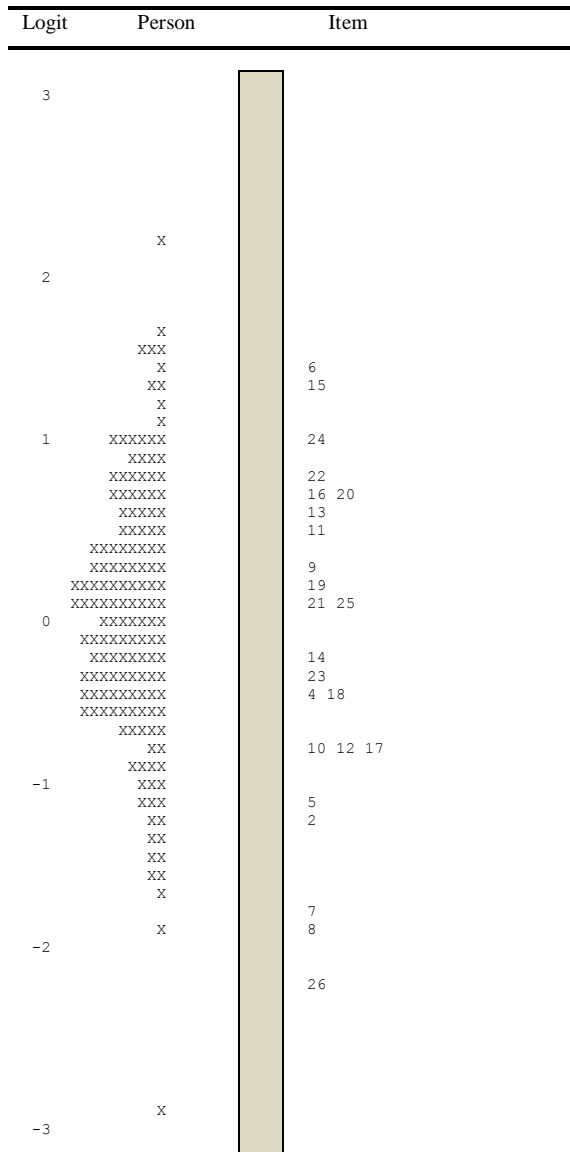


Figure 6.4 Variable Map of the CAK scale

Figure 6.4 shows that the item difficulty estimates for the CAK scale were plotted in decreasing order at varying levels on the measurement. These clustered items were further examined for their common themes. Through the interpretation of these themes, four band levels were identified in terms of levels of difficulty (see Table 6.2).

Table 6.2 Interpretation of the Instructor Classroom Assessment Knowledge Levels from Analyses of the CAK Scale

Item	Logit Range	Description of Instructor Classroom Assessment Knowledge Levels	Percentage
Level 4: Expert			
Q6, Q15, Q24, Q22, Q16, & Q20	> 0.43	On average, the instructor is able to justify the different purposes of classroom assessment such as serving either summative or formative functions prior to using it; verify an appropriate interpretation of complex statistical concept like percentile when communicating it to the key stakeholders; construct and justify assessment records to gather a variety of students' assessment information appropriately for decision-making; judge a range of students' learning behaviours to be observed and recorded in the class in an appropriate and consistent way; and evaluate the appropriate types of assessment information to be recorded for making an accurate inference of the students' learning achievements.	24%
Level 3: Proficient			
Q13, Q11, Q9, Q19, Q21, & Q25	$> -0.17 \text{ to } \leq 0.43$	On average, the instructor is able to construct various assessment methods including multiple-choice questions, essay questions, and authentic assessment tasks to assess student learning outcomes; distinguish the two types of interpretative frameworks such as criterion-referenced and norm-referenced frames of referencing with respect to interpreting students' performance; combine the basic statistical information such as mean and standard deviation for interpreting the students' raw scores appropriately; analyse the relationship between test raw scores and its percentile rank for appropriate interpretation of the scores; and ensure the accuracy and consistency of the assessment procedure.	33%
Level 2: Competent			
Q14, Q23, Q4, Q18, Q10, Q17, & Q12	$> -1.20 \text{ to } \leq -0.17$	On average, the instructor is aware of using appropriate assessment methods to assess students' learning and elicit information from them for formative purposes; demonstrate an understanding about the importance of having consultations with colleagues who had assessment expertise and piloting the new developed test questions to ensure their validity and reliability prior to using them; aware of the validity issue associated with grading procedure implemented; explain the students' raw scores on a 100-scale to the relevant assessment key stakeholders; and show caution on his/her own engagement in assessment unethical behaviours like the issue of adding marks to students' test scores.	33%
Level 1: Novice			
Q5, Q2, Q7, Q8, & Q26	≤ -1.20	On average, the instructor is aware of the way in which grades could be used to accurately reflect the students' learning achievements; aware of his/her own engagement in unethical behaviours like revealing exam information to some students and/or allowing students bring course material to the test room to refer to during the test; and describe the ways in which s/he determined his/her course grades to the assessment key stakeholders.	10%

Table 6.2 illustrates that of the 108 instructors who took the CAK, 10% were placed in level 1 (novice). Another 33% of the instructors were identified as competent. A cohort of 33% were grouped as proficient. A further 24% were classified as expert. Given only 24% of the instructors were able to master the Expert level, it could be concluded that the assessment knowledge level of this sample of instructors was limited. The highest assessment knowledge level for this group of instructors was relevant to standard 6 (i.e., Recognising Unethical Assessment Practices) whereas, their lowest assessment knowledge level was associated with standard 7 (i.e., Keeping Accurate Records of Assessment Information).

6.1.2 Development of the Innovative Methods scale

The Innovative Methods (IM) scale consisted of nine items and was constructed to measure the extent to which the instructors endorsed the use of innovative assessment methods in assessing their students' learning. Respondents were asked to indicate their agreement to the four-point rating scale varying from "Not useful at all" to "Very useful". The response patterns for items 1, 4, 5, 6, and 8 were coded in the following way: Not useful at all= 0, A little useful= 1, Useful= 2, and Very useful= 3. Within this scale, items 2, 3, 7, and 9 were recoded as either "A little useful" or "Useful" through collapsing "Not useful at all" with "A little useful," and "Useful" with "Very useful" using a dichotomous score of zero and one respectively. These items were collapsed given the expected score on the item did not increase according to the level of the respondents' endorsements toward innovative methods used. The results from the Classical and Rasch analyses of both the Rasch Simple Logistic Model and the Partial Credit Model are displayed in Table 6.3. The summary of the findings follows the same pattern as for the Classroom Assessment Knowledge scale.

Table 6.3 Calibration Estimates for the Innovative Methods Scale

Item	Facility	Point-biserial	Logit	SE	INFIT MNSQ
1. Self-assessment	0.51	0.37	-0.15	0.17	1.05
2. Portfolio	0.62	0.33	-0.60	0.22	1.00
3. Peer assessment	0.66	0.37	-0.80	0.22	1.10
4. Individual conference	0.68	0.55	-1.10	0.14	0.97
5. Reflective journal	0.67	0.50	-1.52	0.16	0.96
6. Individual assignment/ project work	0.75	0.34	-1.60	0.16	1.12
7. Assessments that resemble the English language use in your students' real life situations	0.82	0.32	-1.86	0.27	0.98
8. Assessments that provide regular feedback indicating the ways to improve your students' future performance	0.83	0.47	-1.90	0.20	0.93
9. Oral presentation	0.87	0.49	-2.28	0.31	0.83

The mean item INFIT was 0.99, with a standard deviation of 0.89. The INFIT MNSQ values illustrated in Table 6.3 ranged from 0.83 to 1.12 indicating that all items had an acceptable fit to the Rasch model. The standard deviation of the item parameter estimates was 0.70. The point biserial coefficient estimates for each item ranged from 0.32 to 0.55. This indicates evidence of the dominant construct validity of the scale. There were no items with perfect scores or zero scores.

Whilst the Table 6.3 shows average item parameter estimates across categories, the variable map in Figure 6.5 illustrates the item step parameter estimates. For example, 1.3 is item parameter estimates for item 1 category 3. The item parameter estimates in Figure 6.5 vary from -4.47 to +2.78 logits, a range of 7.25 logits. The range of item parameter estimates suggests the levels of instructors' endorsements toward innovative assessment methods that the instrument is suitable to measure. The instructors' endorsement toward innovative methods used estimates ranged from -3.48 to +3.96, a range of 7.44 logits. Thus, it appears that there is a good match between the instructors' endorsement towards innovative methods used and the range of item parameter estimates. Of the 108 instructors surveyed, there was one instructor with perfect scores (i.e., s/he responded to "very useful" for each of the nine items). The standard errors of estimates for the items were acceptable, ranging from 0.14 to 0.31. Item 1 was relevant to the endorsements toward the use of "*Self-assessment*". It was the item that instructors

considered the least useful on the Innovative Methods scale. However, item 9, concerned with their endorsements toward the employment of “*Oral presentation*,” was the item that instructors considered the most useful in assessing their students’ learning.

The estimates of instructor endorsement toward innovative methods and the item parameter were then plotted on the variable map (see Figure 6.5).

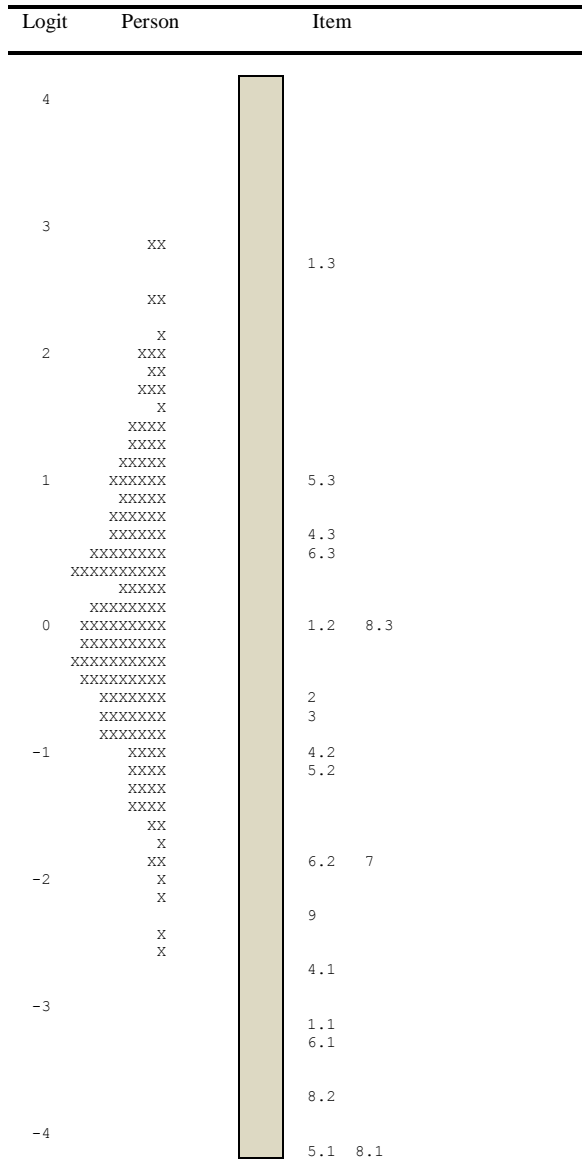


Figure 6.5 Variable Map of the IM scale

Figure 6.5 shows that the item parameter estimates for the IM scale were plotted in decreasing order. These clustered items were further examined for their common themes. Through the interpretation of these themes, two band levels were identified based on item parameter estimates (see Table 6.4).

Table 6.4 Interpretation of the Instructor Innovative Methods Levels from Analyses of the IM Scale

Item	Logit Range	Description of Instructor Innovative Methods Levels	Percentage
1. Self-assessment 2. Portfolio 3. Peer assessment 4. Individual conference 5. Reflective journal	≥ -1.94	Level 2: Advanced Innovative Methods On average, the instructor highly endorses the usefulness of using advanced innovative assessment methods in his/her assessment practice. That is, s/he seems to have a positive endorsement toward employing reflective journal, individual conference, portfolio, peer and self-assessments in assessing the students' learning.	93%
6. Individual assignment/ project work 7. Assessments that resemble the English language use in your students' real life situations 8. Assessments that provide regular feedback indicating the ways to improve your students' future performance 9. Oral presentation	< -1.94	Level 1: Basic Innovative Methods On average, the instructor perceives the usefulness of employing basic innovative methods in assessing the students' learning. That is, s/he tends to have a positive endorsement towards using oral presentation, authentic assessment and assignments in his/her assessment practice.	7%

As Table 6.4 illustrates, of the 108 instructors surveyed, 7% of instructors thought that oral presentation, authentic assessment and individual assignments/projects were useful methods to assess their students' learning. These instructors had not yet progressed to the endorsement of the advanced innovative methods. However, the remaining 93% of instructors surveyed were of the opinion that reflective journal, individual conference, portfolio, peer assessment, and self-assessments were useful methods to assess their students' learning. Given the scale is cumulative and developmental in nature, this group of instructors also endorsed the use of basic innovative methods comprising oral presentation, authentic assessment and individual assignments/projects in assessing their

students' learning. Hence, this sample of instructors appeared to have positive endorsements toward employing both basic and advanced innovative methods in their assessment practices.

6.1.3 Development of the Grading Bias Scale

The Grading Bias (GB) scale was designed to examine the extent to which the instructors believed they were influenced by their students' non-academic achievement factors, such as age, gender, appearance, behaviour, attitude, effort and general abilities, when marking their work and/or performance. Respondents were asked to indicate how often their students' non-academic achievement factors influenced their marking. The items have a four-point rating scale ranging from "Never= 0" to "Always= 3".

The results from both the Classical and Rasch analyses of the Partial Credit Model are displayed in Table 6.5. The summary of the findings follows the same pattern as for the Innovative Methods scale.

Table 6.5 Calibration Estimates for the Grading Bias Scale

	Item	Facility	Point-biserial	Logit	SE	INFIT MNSQ
1.	Age	0.07	0.52	3.27	0.24	0.88
2.	Gender	0.08	0.58	3.21	0.23	0.87
3.	Appearance	0.10	0.60	2.84	0.20	0.86
4.	Behaviour	0.35	0.77	0.97	0.14	0.75
5.	Attitude	0.36	0.72	0.94	0.15	0.91
6.	General abilities	0.59	0.52	-0.52	0.13	1.23
7.	Effort	0.70	0.51	-1.40	0.16	1.13

Table 6.5 shows the INFIT MNSQ values ranged from 0.75 to 1.23 suggesting that all items had an acceptable fit to the model. The mean item INFIT was 0.95, with a standard deviation of 0.17. The point biserial coefficient estimates for each item ranged from 0.51 to 0.77. The fit statistics and the point biserial coefficient estimates provide evidence that the set of items is measuring a dominant construct of the scale. There were no items with perfect scores or zero scores.

The standard deviation of the item parameter level was 1.86. The item parameter estimates displayed in Figure 6.6 varied from -3.38 to +4.16, a range of 7.54 logits. The

instructors' beliefs about their grading bias estimates ranged from -4.36 to +5.59, a range of 9.95 logits. It is apparent that the range of item parameter estimates was much narrower than the range of instructors' beliefs about their grading bias. As such, it appears that the upper end of the scale may not have been as well matched to the upper end of the instructors' range of grading bias. Of the 108 instructors surveyed, there was one instructor with perfect scores (i.e., s/he responded to "always being influenced" for each of the seven items) and six instructors with zero scores (i.e., they indicated for each of the seven items that they were "never influenced").

The standard errors of estimates for the items were acceptable, with the largest value of 0.24 for item 1. Item 1 was about students' "Age". It has the least influence on the Grading Bias scale, with only 7% of the sample indicating they were often or always being influenced by age. In contrast, item 7, which was concerned with students' "Effort," was the item that most instructors (70%) reported as often or always influenced them, when marking their students' work/performance. Consideration of the fit indices and low standard errors in association with the high person separation reliability of .87 indicates that the scale had reliable measurement of the Grading Bias construct.

The estimated levels of the instructors' grading bias and the item parameter were then plotted on the variable map (see Figure 6.6).

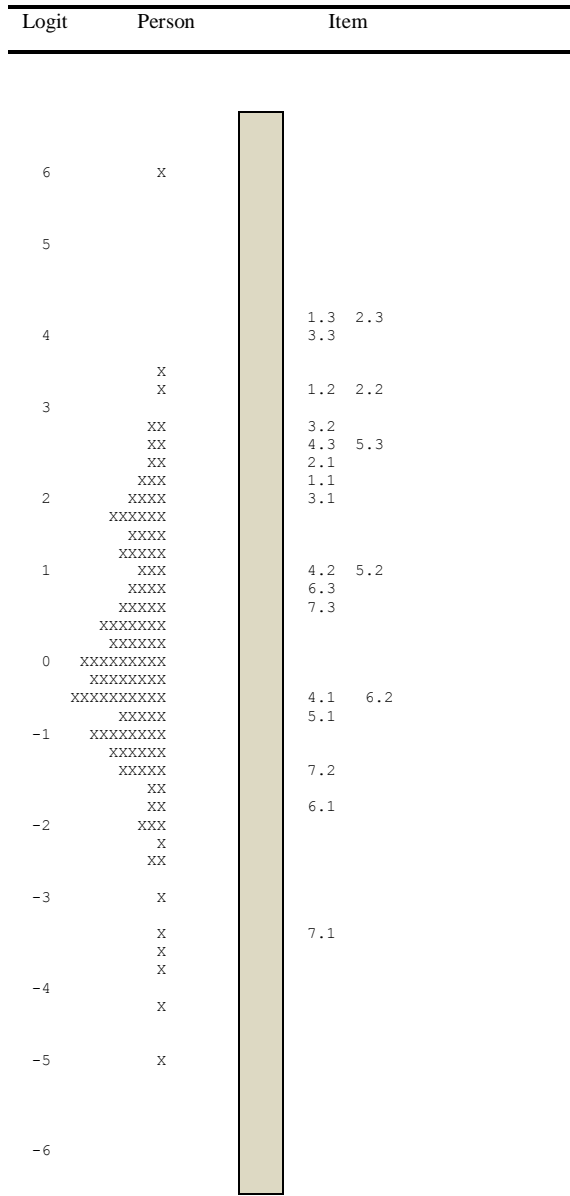


Figure 6.6 Variable Map of the GB scale

Figure 6.6 shows that the item parameter estimates for the GB scale were plotted in decreasing order. These clustered items were further examined for their common themes. Through the interpretation of these themes, three band levels were identified (see Table 6.6).

Table 6.6 Interpretation of the Instructor Grading Bias Levels from Analyses of the GB Scale

Item	Logit Range	Description of Instructor Grading Bias Levels	Percentage
1. Age 2. Gender 3. Appearance	≥ 1.08	Level 3: Seniority and Physical Appearance Influence On average, the instructor is influenced by the students' age, gender and physical appearance when s/he marks their work/performance.	31%
4. Behaviour 5. Attitude	≥ 0.05 to < 1.08	Level 2: Attitude and Behaviour Influence On average, the instructor is influenced by the students' attitude and behaviours when s/he marks students' work/performance.	13%
6. General abilities 7. Effort	< -0.05	Level 1: General Ability and Effort Influence On average, the instructor is influenced by the students' general ability as well as their effort when s/he marks students' work/performance.	56%

As displayed in Table 6.6, of the 108 instructors surveyed, 56% reported their students' general ability and effort often or always influenced them when marking their work/performance. Another 13% indicated they were often or always influenced by their students' attitudes and behaviours when judging their work/performance. A further 31% reported their students' seniority and physical appearance often or always influenced the ways in which they marked their work/performance.

6.1.4 Development of the Quality Procedure Scale

The Quality Procedure (QP) scale was designed to examine the extent to which the instructors had positive endorsements toward the use of procedures designed to enhance the quality of the assessment process. It comprised six items covering various issues associated with maintaining assessment records, implementing quality assurance mechanisms and communicating feedback to students in a timely and effective way. Within this scale, respondents were asked to indicate their agreement to the four-point rating scale ranging from "Strongly Disagree" to "Strongly Agree". However, the response patterns were recoded as either "Disagree" or "Agree" by collapsing "Strongly Disagree," and "Disagree," together with "Agree", and keeping "Strongly Agree" as a separate category using a dichotomous score of zero and one respectively. These items

were collapsed, given the expected score on the item did not increase according to the level of respondents' endorsements toward quality procedure used. The results from both the Classical and Rasch analyses of the Rasch Simple Logistic Model are displayed in Table 6.7. The summary of the findings follows the same pattern as for the Grading Bias scale.

Table 6.7 Calibration Estimates for the Quality Procedure Scale

Item	Facility	Point-biserial	Logit	SE	INFIT MNSQ
1. It is just as important to maintain detailed records of the assessment process as it is to maintain records of students' results	0.29	0.37	1.19	0.24	0.97
2. It is important to construct accurate reports about students' achievement for communicating to both students and administrators	0.35	0.43	0.80	0.23	0.97
3. It is important to employ various methods to record students' achievement	0.36	0.29	0.75	0.23	1.05
4. It is important for instructors to gather together regularly to design and check the quality of the assessment process and results	0.41	0.36	0.50	0.22	0.97
5. It is my responsibility to ensure that my assessments are valid and reliable before using them	0.48	0.29	0.10	0.22	1.10
6. It is important to provide students with their assessment results in a timely and effective way	0.55	0.46	-0.24	0.22	0.90

The standard deviation of the item parameter estimates was 0.52. The INFIT MNSQ values ranged from 0.90 to 1.10 suggesting that all items had an acceptable fit to the Rasch model. The mean item INFIT was 0.99, with a standard deviation of 0.71. The point biserial coefficient estimates for each item ranged from 0.29 to 0.46. This provides evidence of the dominant construct validity of the scale.

The item parameter estimates displayed in Table 6.7 varied from -0.24 to +1.19, a range of 1.43 logits. The instructors' perceptions of the importance of implementing quality procedures also varied from -2.17 to +3.20, a range of 5.37 logits. Hence, the range of the item parameter estimates was much narrower than the range of the instructors' perceptions. As such, it appears that the upper end of the scale may not have been as well matched to the upper end of the instructors' range of perceptions of the

importance of quality procedures. There were four instructors with perfect scores (i.e., they agreed to all items) and 18 instructors with zero scores (i.e., they disagreed to all items). The standard errors of estimates for each of the items were acceptable, with the largest value of 0.24 for item 1. Item 1 was concerned with the use of assessment records *“It is just as important to maintain detailed records of the assessment process as it is to maintain records of students’ results”*. It was found to be the most difficult item for instructors to agree with on the Quality Procedure scale, with only 29% of the sample indicating their agreement to this item. On the other hand, item 6, which was concerned with the communication of feedback to the students *“It is important to provide students with their assessment results in a timely and effective way,”* was the item that the instructors found the easiest to agree to, with 55% of the sample indicated their agreement to this item.

The estimates of the level of the instructors’ perceptions of the importance of each quality procedure and the item parameter were plotted on the variable map (see Figure 6.7).

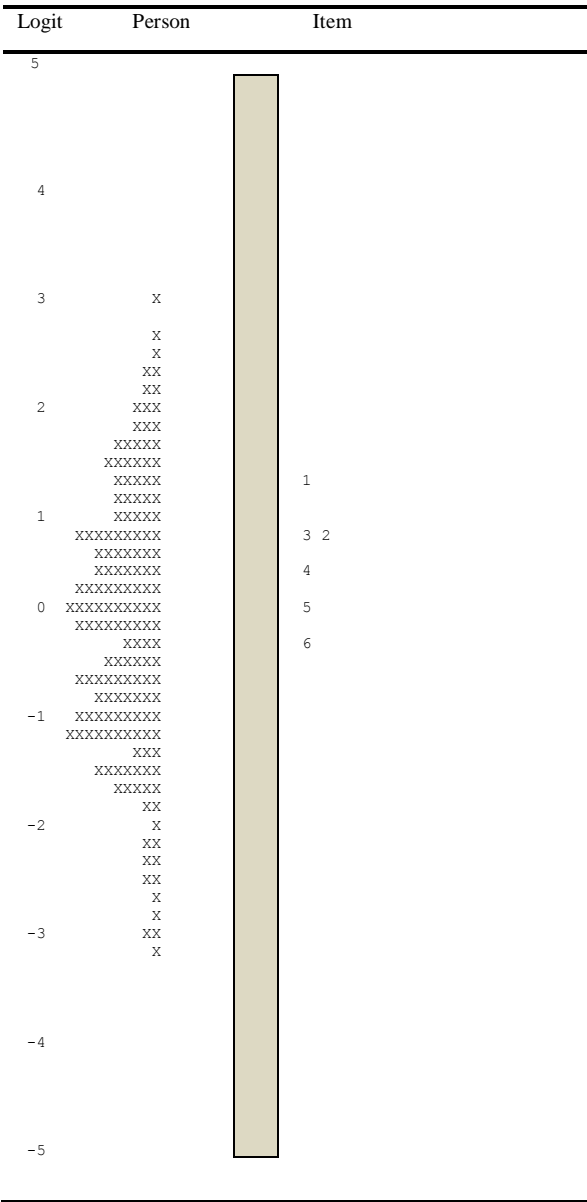


Figure 6.7 Variable Map of the QP scale

Figure 6.7 illustrates that the item parameter estimates for the Quality Procedure scale were plotted in decreasing order on the scale. These clustered items were further examined for their common themes and three band levels were identified (see Table 6.8).

Table 6.8 Interpretation of the Instructor Quality Procedure Levels from Analyses of the QP Scale

Item	Logit Range	Description of Instructor Quality Procedure Levels	Percentage
1. It is just as important to maintain detailed records of the assessment process as it is to maintain records of students' results 2. It is important to construct accurate reports about students' achievement for communicating to both students and administrators 3. It is important to employ various methods to record students' achievement	> 0.52	Level 3: Record-keeping On average, the instructor views the importance of using various methods to keep accurate written records of the assessment process and students' results for communicating to key assessment stakeholders including students, parents, administrators, and potential employers.	29%
4. It is important for instructors to gather together regularly to design and check the quality of the assessment process and results	≥ -0.11 to ≤ 0.52	Level 2: Ensuring Quality Assurance On average, the instructor perceives the vital role of constructing assessment tasks as a team and the quality assurance plays in the assessment process. That is, s/he holds a positive endorsement toward having a team developing assessment activities and having quality assurance to monitor the assessment process.	19%
5. It is my responsibility to ensure that my assessments are valid and reliable before using them 6. It is important to provide students with their assessment results in a timely and effective way	< -0.11	Level 1: Own Responsibility On average, the instructor shows a positive endorsement toward his/her own responsibility to ensure the validity and reliability of the assessments prior to using them as well as providing the students' assessment results in a timely and effective way.	52%

Table 6.8 shows that, of the 108 instructors surveyed, 52% reported they considered it was important they ensured the validity and reliability of their assessments prior to using them, as well as providing their students with feedback in a timely and effective way. Another 19% indicated that they endorsed quality assurance procedures in their assessment practices. A further 29% reported that they had positive endorsements toward using record-keeping in their assessment practices. As the scale is cumulative and developmental in nature, the result suggested that this sample of instructors positively

endorsed good quality assessment procedures in their assessment practices, given they typically had favourable attitudes toward such items.

6.2 Summary Statistics

The following Table 6.9 illustrates the summary statistics for each of the four scales.

Table 6.9 Summary Estimates of the Classical and Rasch Analyses for each Scale

Variable Name	No. of Items	No. of Scores	Scale Logit			Person Logit			Item INFIT (MNSQ)			Person Separation Reliability	Cronbach α
			Min	Max	Range	Min	Max	Range	Mean	Min	Max		
CAK	24	24	-2.12	1.41	3.53	-2.35	2.36	4.71	0.99 (0.68)	0.84	1.10	.74	.74
IM	9	19	-4.47	2.78	7.25	-3.48	3.96	7.44	0.99 (0.89)	0.83	1.12	.73	.71
GB	7	21	-3.38	4.16	7.54	-4.36	5.59	9.95	0.95 (0.17)	0.75	1.23	.87	.84
QP	6	6	-0.24	1.20	1.44	-2.17	3.20	5.37	0.99 (0.71)	0.90	1.10	.63	.64

As can be seen from Table 6.9, the range of the instructors' parameter estimates were adequately matched with the range of the item difficulty/parameter estimates within the Classroom Assessment Knowledge (CAK) and Innovative Methods (IM) scales. With regard to the Grading Bias (GB) and Quality Procedure (QP) scales, it appears that the range of the instructors' parameter estimates were much broader than the range of the item parameter estimates. This indicates that future revisions of these two scales would need to comprise more items at the scale ends (i.e., items which are hard to agree with) to better match the range of the instructors' parameter.

In examining the internal consistency reliability (i.e., Cronbach's alpha) of the scales, classical item analyses procedures were employed to measure the intercorrelation of the individual items on each scale. It is widely acknowledged that there is no clear standard with regard to acceptable or unacceptable levels of Cronbach's alpha (Clark & Watson, 1995; Schmitt, 1996). Previous literature details a variation in the values that represent the acceptable level of Cronbach's alpha. For instance, Nunnally (1978) has

proposed the level of Cronbach's alpha .70 as acceptable while others (see Deković, Janssens, & Gerris, 1991; Holden, Fekken, & Cotton, 1991; Myers & Oetzel, 2003) characterise the level of Cronbach's alpha as greater than .60 as good or adequate. As Table 6.9 shows, the level of Cronbach's alpha for all scales varies from .64 to .84 indicating moderate to high internal consistency reliabilities. Hence, all scales employed in the current study demonstrated acceptable internal consistency reliabilities.

Nuttall and Skurnik (1969) also investigated the issue of errors associated with traditional measures of test reliability such as Cronbach's alpha. Given the study adopts the Nuttall and Skurnik's (1969) approach by taking account of the errors of classification implied by the test reliability, all scales have adequate test reliability to support their categories when the margin of error associated with each category is \pm one category. For example, the:

- Classroom Assessment Knowledge scale (CAK) has adequate test reliability to support four categories [c.f. Table 6.2: 4];
- Innovative Methods scale (IM) has adequate test reliability to support four categories [c.f. Table 6.4: 2];
- Grading Bias scale (GB) has adequate test reliability to support six categories [c.f. Table 6.6: 3]; and
- Quality Procedure scale (QP) has adequate test reliability to support four categories [c.f. Table 6.8: 3].

In addition to the classical analyses of internal consistency reliabilities for all scales, the Rasch reliabilities or person separation reliabilities (Wright & Masters, 1982) for these scales were undertaken to further examine the reliability of each scale. Table 6.9 illustrates that the person separation reliability estimates range from .63 for the Quality Procedure variable to .87 for the Grading Bias variable demonstrating moderate to high internal consistency reliabilities. Hence, the results obtained from both Classical and Rasch analyses show that all scales had satisfactory measurement properties. The next chapter reports the results from the quantitative phase of the study.

Chapter 7: Quantitative Results

This chapter reports the results obtained from the quantitative phase of the study. The results are organised into three parts. Part one presents the univariate results. Part two reports the bivariate results. Part three presents the multivariate results in relation to the evaluation of a one-factor congeneric model using the confirmatory factor analysis (CFA).

7.1 Univariate Results

7.1.1 The Sample

The background characteristics of the sample are displayed in Table 7.1. Within the table, the coded “SD” indicates the standard deviation.

Table 7.1 Background Characteristics of the Sample

Background Variable	Value	Frequency	Range	Mean and SD
Departmental status	English-major	55% (59)		
	English non-major	45% (49)		
Gender	Male	74% (80)		
	Female	26% (28)		
Age			22-55	30.55 (SD= 6.44)
Years of teaching experience			1-20	5.32 (SD= 4.09)
Number of teaching hours per week			5-50	18.52 (SD= 9.06)
Number of students per class			25-55	34.19 (SD= 7.43)
Level of academic qualification	Bachelor's degree	100%		
	Master's degree	59%		
	Doctoral degree	2%		
With pre-service assessment training		59%		
Duration of assessment training	Less than one semester	25%		
	One semester	30%		
	Two semesters	3%		
	Over two semesters	1%		
Preparedness of assessment training	Unprepared	4%		
	Prepared	43%		
	Very prepared	12%		
Without assessment training		41%		

The sample consisted of 80 male and 28 female EFL instructors who taught at one Cambodian city-based university.

Table 7.1 details that 59% had undertaken assessment training during their pre-service teacher education programme, whereas 41% had not received such assessment training. Of those who received pre-service assessment training (59%), nearly two-quarters (25%) reported that the total duration of such training was less than one semester, about half (30%) indicated receiving an equivalent of one semester in training, and fewer received two semesters (3%) or over two semesters (1%) of pre-service training in assessment. Furthermore, a large proportion (43%) indicated that their pre-service training prepared them for conducting classroom assessments, whilst a sizeable

group said they were very prepared (12%) and a smaller number reported they were unprepared (4%) for conducting classroom assessments.

The respondents' ages ranged from 22 to 55 years, with a mean age of 30.55 (SD= 6.44 years). They had varied years of teaching experience, class sizes and number of teaching hours per week. Although all respondents obtained at least a Bachelor of Arts degree in teaching English as a Foreign Language (EFL), over two-quarters received master's degrees (59%) and only a few had a doctoral degree (2%) majoring in Education. With regard to the respondents who obtained master's degrees (59%), most majored in Education (43%), whereas fewer majored in Business (3%), Politics (2%), and other unclassified fields (11%).

Of the 108 respondents, 55% were classified as English-major instructors and the remaining 45% were classified as English non-major instructors.

The English-major instructors comprised 47 males and 12 females and their ages ranged between 22 and 46, with a mean age of 29.7 (SD= 5.49). More than half (68%) reported receiving formal studies in classroom assessment from their pre-service teacher education programme, whilst a sizeable group (32%) indicated they had no such assessment training. The class size they taught ranged between 25 and 35, with a mean class size of 30.4 (SD= 2.15). The teaching hours they taught per week ranged between 5 and 33, with a mean teaching hour of 18.1 (SD= 7.20).

The English non-major instructors consisted of 33 males and 16 females, aged between 22 and 55, with a mean age of 31.5 (SD= 7.36). Nearly half (49%) indicated they had formal studies in classroom assessment from their pre-service teacher education programme, although just over half (51%) reported having no such assessment training. The class size they taught ranged between 25 and 55, with a mean class size of 38.8 (SD= 8.84). The teaching hours they taught per week ranged between 6 and 50, with a mean teaching hour of 19.1 (SD= 10.93).

7.1.2 Tests of Normality

To undertake univariate and multivariate tests of normality of the scales, IBM SPSS Amos software version 20 was employed. The results are displayed in Table 7.2 in terms of each scale, mean, standard deviation (SD), skewness, kurtosis and

critical ratio (CR, also known as t-value) estimates. The scores of each scale have been reported in terms of the logit values, which were derived from the Rasch analysis (refer to Chapter 6).

Table 7.2 Mean, Standard Deviation, Skewness and Kurtosis Estimates

Scale	Variable Name	Mean	SD	Min	Max	Skewness		Kurtosis	
						Estimate	CR	Estimate	CR
Grading Bias	GB	0.03	1.84	-4.36	5.59	-0.13	-0.55	0.81	1.72
Innovative Methods	IM	-0.01	1.24	-3.48	3.95	0.11	0.45	0.34	0.73
Quality Procedure	QP	0.05	1.39	-2.17	3.20	0.03	0.14	-0.49	-1.05
Classroom Assessment Knowledge	CAK	0.01	0.93	-2.35	2.37	0.31	1.32	0.07	0.14

The normality of each scale was examined through assessment of the kurtosis and skewness estimates illustrated in Table 7.2 above. It has been recommended that the value of skewness or kurtosis that has exceeded the range between +2 and -2 would be considered to be significantly different from zero, indicating that a transformation of the scale might be needed (Tabachnick & Fidell, 2013). As Table 7.2 shows, all scales had the value of skewness or kurtosis within the range between +2 and -2, indicating that each satisfied the assumption of univariate normality. To further test the distribution of normality of all scales, Mardia's (1970, 1974) estimate of multivariate kurtosis was performed. According to kurtosis and skewness functions, the multivariate test results should be less than 3 to fulfil the normality condition. The test of multivariate normality of all four scales revealed a multivariate kurtosis of 2.041 and a critical ratio (or t-value) of 1.531. Hence, the assumption of multivariate normality of all scales was met.

7.2 Bivariate Results

7.2.1 Interrelationships among the Classroom Assessment Literacy Constructs

To examine the interrelationships among classroom assessment literacy constructs, comprising Grading Bias (GB), Innovative Methods (IM), Quality Procedure (QP) and Classroom Assessment Knowledge (CAK), Pearson product-moment correlations were calculated. To further examine the relationships of these classroom assessment literacy constructs with the instructors' age (AGE) (measured in years), their teaching experience (TEXPER) (measured in years), teaching hours (THOUR) (measured by the number of hours taught per week), and class size (CSN) (measured by the number of students per class), Pearson product-moment correlations were also computed. These analyses used IBM SPSS Statistics software version 20. The results of these analyses are displayed in Table 7.3.

Table 7.3 Pearson Product-Moment Correlations for the Relationships among the Classroom Assessment Literacy Constructs, Age, Teaching Experience, Teaching Hours, and Class Size

	GB	IM	QP	CAK	AGE	TEXPER	THOUR	CSN
GB	1.00	-.37**	-.32**	-.38**	-.01	-.05	.04	.34**
IM	-.37**	1.00	.26**	.32**	-.03	-.04	-.19*	-.16
QP	-.32**	.26**	1.00	.29**	.16	.18	.00	-.37**
CAK	-.38**	.32**	.29**	1.00	-.14	-.08	-.04	-.48**
AGE	-.01	-.03	.16	-.14	1.00	.85**	.13	-.92
TEXPER	-.05	-.04	.18	-.08	.85**	1.00	.09	-.17
THOUR	.04	-.19*	.00	-.04	.13	.09	1.00	-.14
CSN	.34**	-.16	-.37**	-.48**	-.92	-.17	-.14	1.00

Note: **p<.01
*p<.05

Given the GB, IM and QP variables were thought to underpin the instructors' personal beliefs of their actual classroom assessment implementation, it was expected they would be correlated. As the Grading Bias was considered to measure the extent to which the instructors were influenced by students' personal characteristics (e.g., effort

and attitude) when they marked their work/performance, it was therefore expected to be negatively correlated with Innovative Methods and Quality Procedure variables. However, as the two variables (i.e., IM and QP) were thought to measure the instructors' beliefs about the usefulness of innovative methods and the importance of quality assessment procedures, it was expected they had a positive relationship. Furthermore, given the Classroom Assessment Knowledge variable was thought to be the measurement of the instructors' classroom assessment knowledge base, it was expected to positively correlate with Innovative methods and Quality Procedure variables, and negatively correlate with the Grading Bias variable.

As can be seen from Table 7.3, the correlations highlighted 12 significant relationships among the four constructs of classroom assessment literacy. As expected, the extent to which the instructors believed their marking was influenced by students' personal characteristics (i.e., GB variable) was found to be negatively related to the extent to which the instructors believed in the usefulness of the innovative methods used (GB-IM, $r = -0.37$, $p < .01$), suggesting that the more the instructors were influenced by students' personal characteristics, the less they believed in the usefulness of the innovative methods employed. It was also found that the extent to which the instructors perceived their marking was influenced by students' personal characteristics (i.e., GB variable) was negatively correlated with the extent to which the instructors believed in the importance of quality procedures employed (GB-QP, $r = -0.32$, $p < .01$), indicating that the more the instructors were influenced by students' personal characteristics, the less they believed in the importance of quality procedures used. Furthermore, it was found that the extent to which the instructors believed their marking was influenced by students' personal characteristics (i.e., GB variable) had a negative relationship to the level of classroom assessment knowledge of the instructors (GB-CAK, $r = -0.38$, $p < .01$). This result indicated that the higher the classroom assessment knowledge instructors had, the less they believed they were influenced by students' personal characteristics when marking their work/performance. These results are consistent with what would be expected.

Table 7.3 also revealed that there was a significant positive relationship between instructors' beliefs of the usefulness of the innovative methods used and their beliefs of

the importance of quality procedures employed (IM-QP, $r = 0.26$, $p < .01$), indicating that the more the instructors believed in the usefulness of the innovative methods employed, the more they favoured the use of quality procedures. Moreover, the level of instructor classroom assessment knowledge (i.e., CAK variable) was found to be significantly related to the extent to which the instructors believed in the usefulness of the innovative methods used (CAK-IM, $r = 0.32$, $p < .01$), suggesting that the higher the classroom assessment knowledge instructors possessed, the more they believed in the usefulness of innovative methods employed. Furthermore, it was found that the level of instructor classroom assessment knowledge (i.e., CAK variable) had a positive relationship with the extent to which the instructors believed in the importance of quality procedures employed (CAK-QP, $r = 0.29$, $p < .01$). This result suggested that the higher the classroom assessment knowledge instructors had, the more they believed in the importance of quality procedures used in their classroom assessment practices. In the following sections, an examination of the influence of each of the instructors' background characteristics on classroom assessment literacy constructs is further investigated.

7.2.2 Classroom Assessment Literacy Variables as a Function of Age

To examine the influence of the instructors' age (AGE) on each of the classroom assessment literacy constructs, Pearson product-moment correlations were computed. The correlations presented in Table 7.3 (refer to section 7.2.1) showed there were no significant relationships among the four constructs of classroom assessment literacy with the instructors' age. Hence, there is no evidence that age of the instructors had an impact on the four constructs (i.e., CAK, GB, IM and QP) of classroom assessment literacy.

To further investigate the influence of the instructors' age on the band level positioning of each classroom assessment literacy variable, cross-tabulations were carried out. Prior to undertaking cross-tabulations, the instructor's age was recoded into four groups as presented in Table 5.1 (see section 5.2.3.1 in Chapter 5). Each instructor was located at a specific band level through estimates of the level of the assessment knowledge and assessment beliefs on the classroom assessment literacy variable (i.e., the logit values estimated through item response modelling and reported in Chapter 6). For each classroom assessment literacy variable (i.e., CAK, GB, IM and QP), the band levels

were cross tabulated with the recoded age groups of the instructors. The cross tabulations assisted the examination of the association of instructor age groups with band levels. The results of the cross tabulations are illustrated as a series of bar charts in Figure 7.1. In each bar chart, the band level is shown on the vertical axis (i.e., y axis) and the classification of the instructor age groups is displayed on the horizontal axis (i.e., x axis). There is one bar chart displayed for each scale. Each bar chart comprises instructor age groups variable in which the sum of each category in the age groups variable is equal to 100% across all band levels.

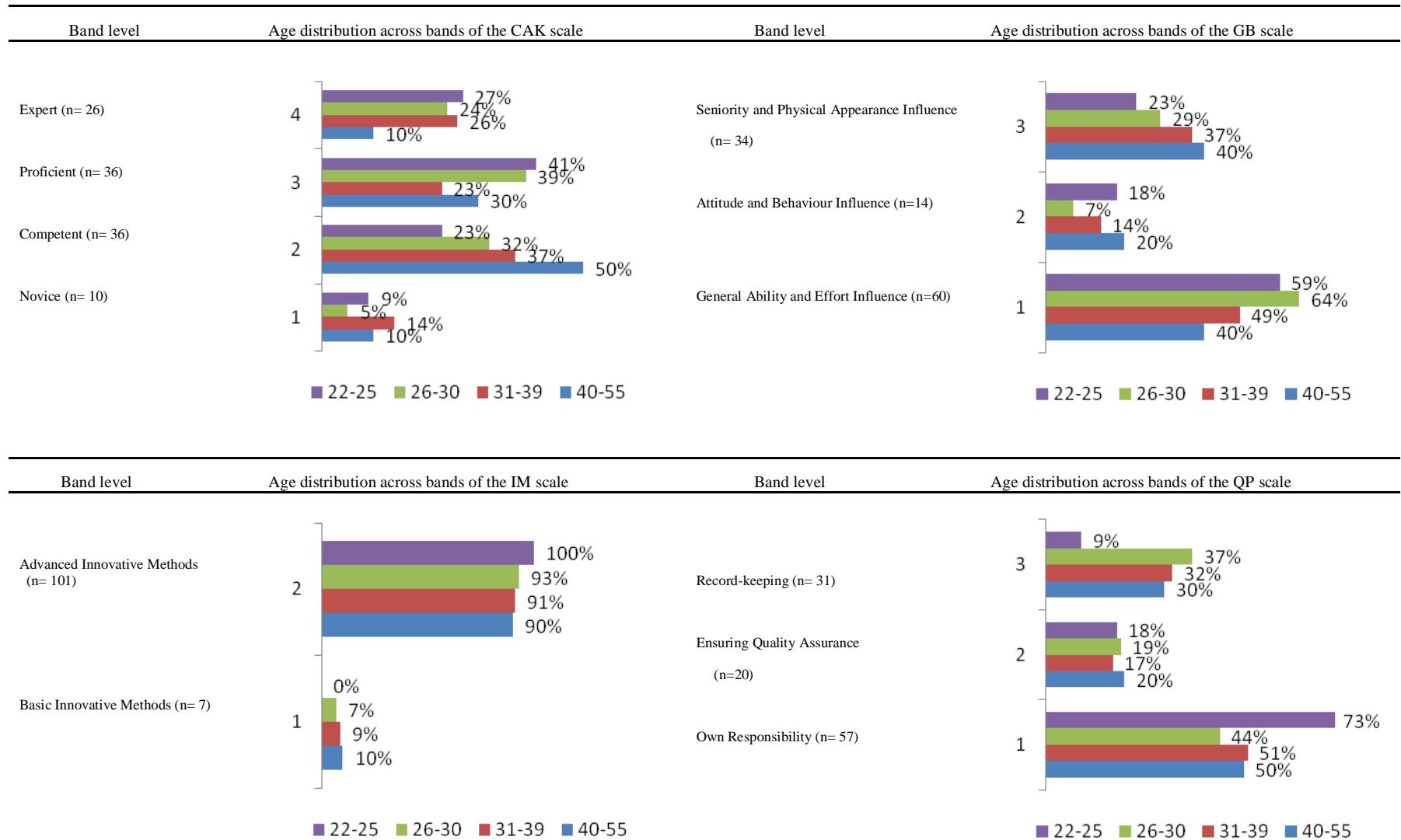


Figure 7.1 Recoded instructor age variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

An inspection on the influence of the age groups on the level of instructor progression along the Classroom Assessment Knowledge (CAK) scale revealed instructors aged between 31-39 and 40-55 years tended to predominantly locate at Band Level 2 (Competent). However, instructors aged between 22-25 and 26-30 years tended to predominantly operate at Band Level 3 (Proficient). This result suggested that the four age groups of the instructors tended to predominantly locate at the middle of the scale (i.e., Band Levels 2 and 3) while a small number tended to be at the lower and upper ends of the scale (i.e., Band Levels 1 (Novice) and 4 (Expert)). With regard to the Quality Procedure (QP) scale, Figure 7.1 showed that the majority of instructors seemed to be located at Band Level 1 (Own Responsibility). While very few instructors aged between 22-25 years (i.e., 9%) were operating at Band Level 3 (Record-keeping), more instructors from older age groups were performing at this particular band level (i.e., 37%, 32%, and 30% respectively). In relation to the Grading Bias (GB) scale, Figure 7.1 revealed that for all age groups, many instructors were at Band Level 1 (i.e., they were influenced by students' general ability and effort when marking their work). However, for older instructors (31-39 and 40-55 years old), a sizeable group (37% and 40% respectively) tended to be located at Band Level 3 (i.e., they were influenced by students' age, gender, and appearance when marking their work) as opposed to those aged between 22-25 and 26-30 years (23% and 29% respectively). There were a small number of these four age groups of instructors at Band Level 2 (i.e., they were influenced by students' attitudes and behaviours when marking their work), suggesting that the instructor age groups were quite extreme at the upper and lower ends of this scale. The percentage distribution patterns for the Innovative Methods (IM) scale showed that the majority of instructors tended to be represented at Band Level 2 (Advanced Innovative Methods), irrespective of their age. This indicated that the instructor age groups were mostly all at the upper end of this scale.

Hence, despite the weak correlation results, there is a suggestion that younger instructors (i.e., 22-25 and 26-30 years of age) may have higher classroom assessment knowledge than older instructors (i.e., 31-39 and 40-50 years of age). Moreover, there is an indication that older instructors (i.e., 26-30, 31-39 and 40-55 years of age) seem to have stronger preferences toward using record-keeping and lesser preferences toward

their own responsibility in ensuring quality assessment procedures than younger instructors (i.e., 22-25 years of age).

7.2.3 Classroom Assessment Literacy Variables as a Function of Teaching Experience

To investigate the influence of the number of years of teaching experience (TEXPER) of the instructors on each of the classroom assessment literacy constructs, Pearson product-moment correlations presented in Table 7.3 were examined. It was revealed that there were no statistically significant relationships amongst the four constructs of classroom assessment literacy (i.e., CAK, GB, IM and QP) with the number of years teaching.

To further examine the influence of the number of years of teaching experience on the band level positioning of each classroom assessment literacy variable, a number of cross-tabulations were undertaken. Prior to undertaking such analyses, the number of years of teaching experience for the instructors was recoded into four groups as presented in Table 5.1 (see section 5.2.3.1 in Chapter 5). The cross-tabulation analyses and their results depicted in each bar chart in Figure 7.2 followed a similar pattern as to that previously reported for age groups.

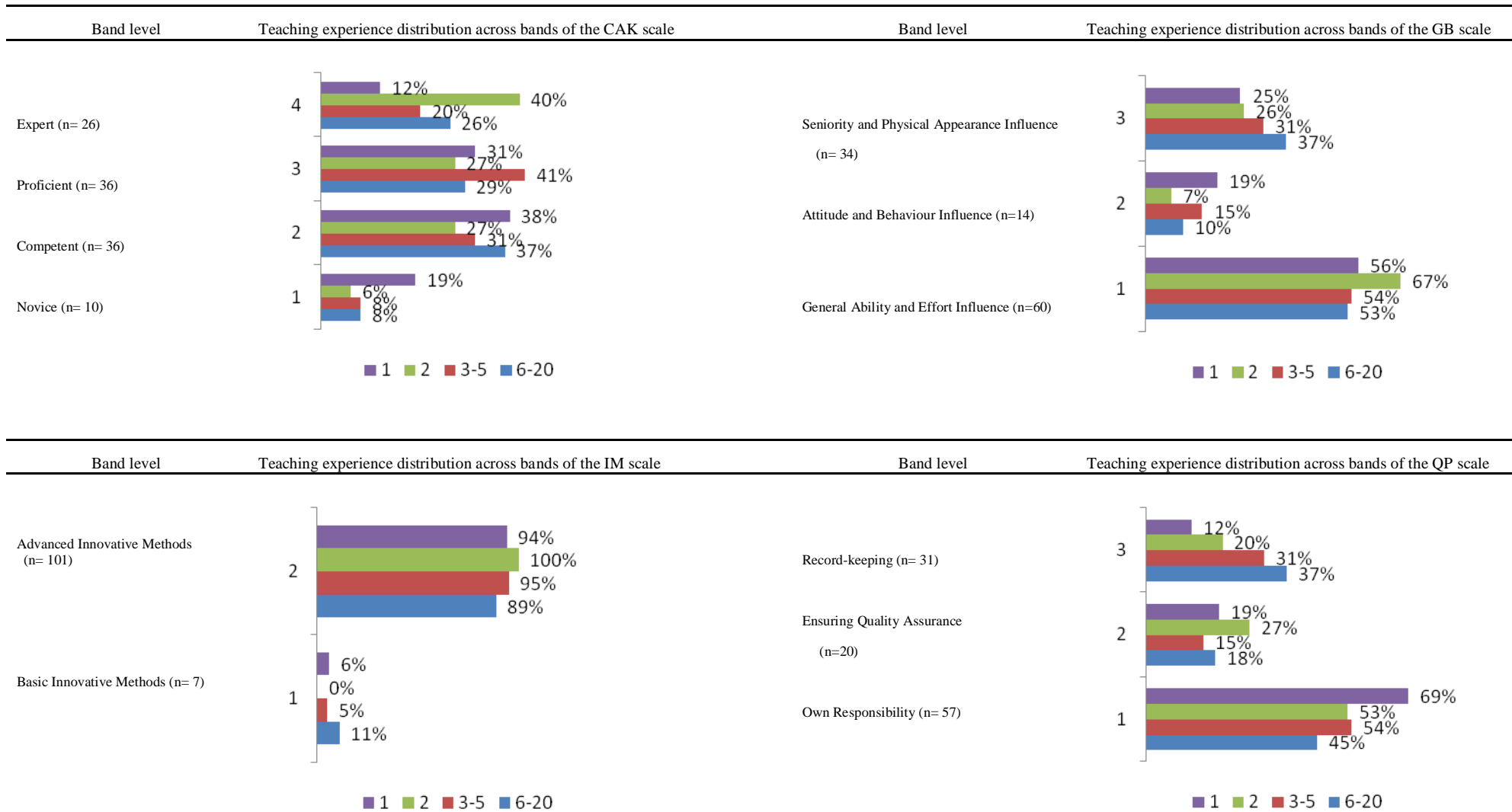


Figure 7.2 Recoded instructor teaching experience variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

In relation to the Classroom Assessment Knowledge (CAK) scale, Figure 7.2 illustrated that irrespective of the number of years of teaching experience, instructors tended to predominantly operate at Band Levels 2 (Competent) and 3 (Proficient). This result suggested that the instructors tended to mostly locate at the middle of this scale (i.e., Band Levels 2 and 3). The percentage distribution patterns for the Grading Bias (GB) scale showed that, irrespective of the number of years of teaching experience, approximately half of the instructors were at Band Level 1 (i.e., they were influenced by students' general ability and effort when marking their work) and about one-third were at Band Level 3 (i.e., they were influenced by students' age, gender, and appearance when marking their work). This suggested that instructors tended to be located mostly at the lower and the upper ends of this scale. In relation to the Innovative Methods (IM) scale, the majority of instructors tended to be represented at Band Level 2 (Advanced Innovative Methods) irrespective of the number of years of teaching experience, indicating these groups of instructors tended to operate mainly at the upper end of this scale. With respect to the Quality Procedure (QP) scale, Figure 7.2 showed that about half of the instructors were performing at Band Level 1 (Own Responsibility) whereas only a small number were at Band Level 2 (Ensuring Quality Assurance) irrespective of the number of years of teaching experience. For Band Level 3 (Record-keeping), there were more experienced instructors than inexperienced instructors (1 year (12%), 2 years (20%), 3-5 years (31%) and 6-20 years (37%). This result indicates that teaching experience may be associated with instructors' levels on the Quality Procedure scale.

7.2.4 Classroom Assessment Literacy Variables as a Function of Teaching Hours

To examine the influence of the number of teaching hours (THOUR) of the instructors on each of the classroom assessment literacy constructs, Pearson product-moment correlations were determined. The correlations shown in Table 7.3 (see section 7.2.1) revealed there was one significant relationship amongst the four constructs of classroom assessment literacy. The number of teaching hours (THOUR) variable was found to be negatively correlated with the innovative methods used (i.e., IM) variable

(THOUR-IM, $r = -.19$, $p < .05$). This result indicated that the more teaching hours the instructors taught per week, the less the instructors believed in the usefulness of innovative methods used.

To further investigate the influence of the number of teaching hours per week on the band level positioning of each classroom assessment literacy construct, cross-tabulations were computed. Prior to undertaking such analyses, the number of teaching hours per week was recoded into three groups as presented in Table 5.1 (see section 5.2.3.1 in Chapter 5). As with the previous background characteristics groups, the cross tabulation analyses and their results depicted in each bar chart are presented in Figure 7.3.

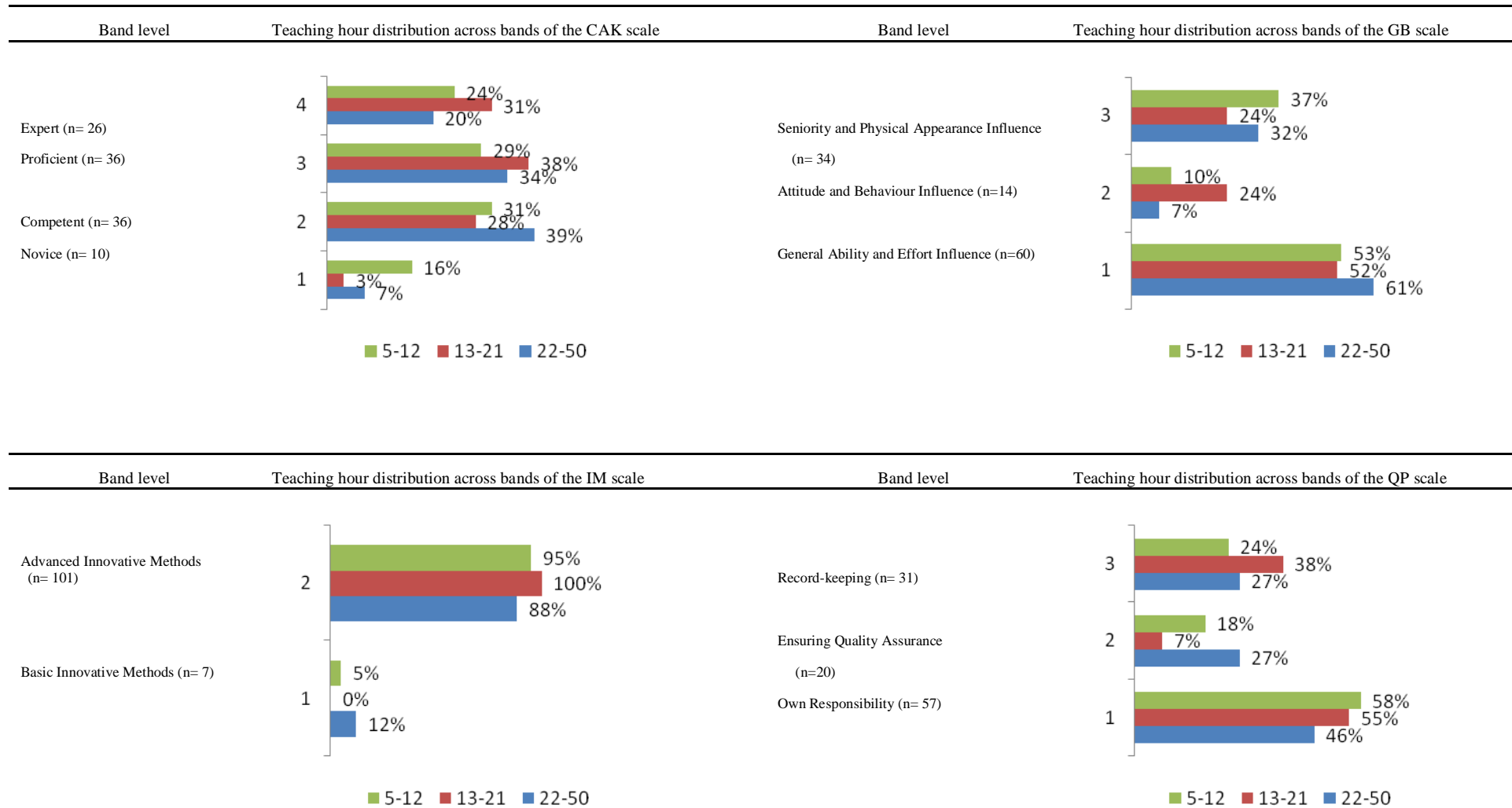


Figure 7.3 Recoded instructor teaching hour variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

Figure 7.3 illustrated that, on the Classroom Assessment Knowledge (CAK) scale, irrespective of the number of teaching hours per week, instructors tended to mostly locate at Band Levels 2 (Competent), 3 (Proficient), and 4 (Expert). In contrast, a very few instructors tended to be at Band Level 1 (Novice). With regard to the Innovative Methods (IM) scale, Figure 7.3 showed that the majority of instructors, irrespective of the amounts teaching hours they had per week, tended to be represented at Band Level 2 (Advanced Innovative Methods). However, more instructors who taught between 5-12 and 13-21 hours per week appeared to endorse the use of advanced innovative methods (i.e., Band Level 2) as opposed to those who taught between 22-50 hours per week. In relation to the Grading Bias scale, Figure 7.3 showed that most instructors tended to be operating at Band Level 1 (i.e., they were influenced by students' general ability and effort when marking their work) and Band Level 3 (i.e., they were influenced by students' age, gender, and appearance when marking their work) irrespective of the number of teaching hours per week. Similar patterns were found for the Quality Procedure (QP) scale, with most instructors who taught between 5-12 hours and 13-21 hours tended to be performing at Band Level 1 (i.e., they endorsed their own responsibility to ensure the quality assessment procedures) and Band Level 3 (i.e., they endorsed the use of record-keeping to ensure the quality assessment procedures), as were instructors who taught between 22-50 hours.

Hence, a closer examination of the band level attainment of instructors suggests that instructors who teach less hours (between 5-12 and 13-21 hours per week respectively) tend to have a more positive endorsement towards using advanced innovative assessment methods, such as self-assessment, peer assessment, portfolio, individual conference and reflective journal, in assessing student learning than those who teach between 22-50 hours per week. This result reinforces the correlation result that indicates that the less number of teaching hours instructors have per week, the more the instructors believe in the usefulness of innovative methods used in assessing student learning.

7.2.5 Classroom Assessment Literacy Variables as a Function of Class Size

Pearson product-moment correlations were also computed to examine the influence of class size (CSN) on each of the classroom assessment literacy constructs. The correlations reported in Table 7.3 indicated there were three significant relationships amongst the four constructs of classroom assessment literacy with the CSN variable. The CSN variable was found to have a positive relationship with the grading bias (i.e., GB) variable (CSN-GB, $r = .34$, $p < .01$), indicating the larger the class size, the more the instructors believed their marking was influenced by students' personal characteristics (i.e., effort and/or attitude). The CSN variable were also found to be negatively correlated with the extent to which the instructors believed in the importance of quality procedures employed (CSN-QP, $r = -.37$, $p < .01$), suggesting that the larger the class size, the less the instructors believed in the importance of the use of the quality procedures. In addition, it was found that the CSN variable had a negative relationship with the level of classroom assessment knowledge of the instructors (CSN-CAK, $r = -.48$, $p < .01$). This result indicated that the larger the class size, the lower the classroom assessment knowledge the instructors had in implementing their assessments (refer to section 9.2.4.2 in Chapter 9 for a detailed explanation).

To further investigate the influence of the class size on the band level positioning of each classroom assessment literacy variable, a number of cross-tabulations were undertaken. Prior to undertaking cross-tabulations, the class size was recoded into three groups as presented in Table 5.1 (see section 5.2.3.1 in Chapter 5). As with the previous background characteristics groups, the cross-tabulation analyses and their results are illustrated in each bar chart in Figure 7.4.

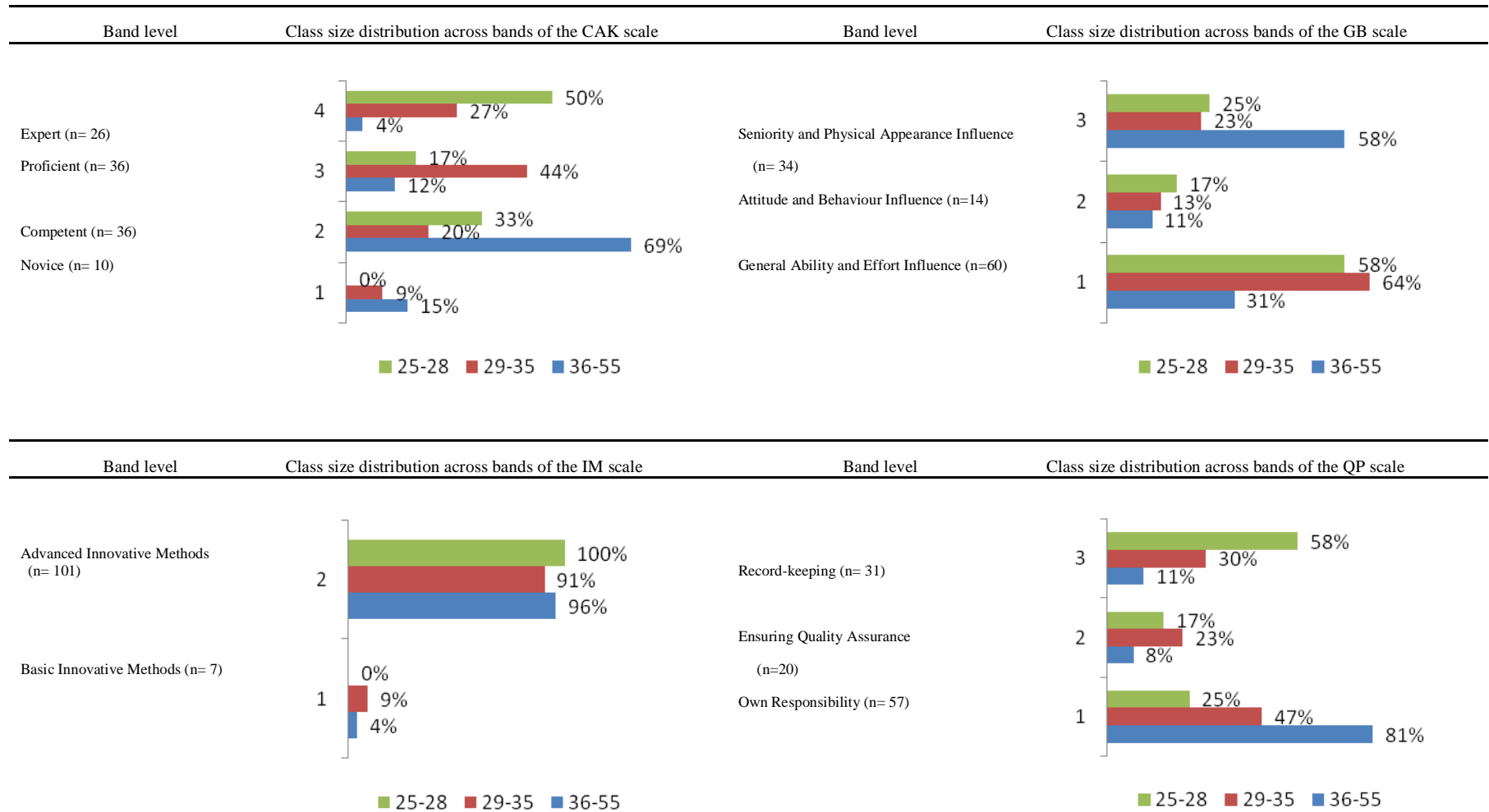


Figure 7.4 Recoded instructor class size variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

An examination of the level of instructor progression along the Classroom Assessment Knowledge scale revealed that approximately half of the instructors who taught small classes (between 25-28 students) had progressed to Band Level 4 (Expert), while the majority of those who taught between 29-35 students per class (44%) had progressed to Band Level 3 (Proficient). However, the majority of instructors (69%) who had large classes (between 36-55 students) were located at Band Level 2 (Competent). With regard to the Grading Bias (GB) scale, more than half of the instructors with small to medium class sizes tended to be located at Band Level 1 (i.e., they were influenced by students' general ability and effort when marking their work). In contrast, more than half of the instructors with large class sizes tended to be at Band Level 3 (i.e., they were influenced by students' age, gender and appearance when marking their work). In relation to the Quality Procedure (QP) scale, the majority who taught large classes (between 36-55 students) seemed to be operating at Band Level 1 (81%), where instructors considered it was their responsibility to implement quality assessment procedures. In contrast, more than half of the instructors who taught small classes (between 25-28 students) had progressed to Band Level 3 (58%), where they valued the importance of accurate record keeping and reporting in the assessment process. The percentage distribution patterns for the Innovative Methods (IM) scale showed that most instructors tended to operate at the upper end of this scale (i.e., Band Level 2). At Band Level 2, they endorsed using reflective journal, individual conference, portfolio, peer assessment and self-assessment in assessing students' learning.

Hence, a closer examination of the band level attainment of instructors indicates that class size influenced their classroom assessment literacy levels, which supported the correlation results that showed that class size impacts on three measures (i.e., CAK, GB and QP) of classroom assessment literacy. That is, the smaller the class size the instructors had, the higher the classroom assessment knowledge they demonstrated, and the higher endorsement they had in relation to the use of quality assessment procedures in their assessment practices. Furthermore, the smaller the class size, the less likely instructors were influenced by students' personal characteristics such as effort and attitude when marking their work.

7.2.6 Classroom Assessment Literacy Variables as a Function of Gender

To examine whether the gender of instructors influences each of the four constructs of classroom assessment literacy, a series of independent t-tests were conducted. Prior to performing t-tests, Levene's tests for Equality of Variance were computed to test the homogeneity of variance amongst the four variables. Levene's tests were not significant for all variables, suggesting that the homogeneity assumption for these variables had been met. Hence, equal variance t-tests were undertaken to determine whether the gender of instructors impacted each of the four classroom assessment literacy constructs.

Table 7.4 Classroom Assessment Literacy Variables as a Function of Gender

	Male (N= 80)		Female (N= 28)		Significance of t-test
	Mean	SD	Mean	SD	
IM	0.01	1.26	-0.06	1.18	n.s
QP	0.10	1.37	-0.08	1.49	n.s
GB	0.03	1.98	0.01	1.40	n.s
CAK	0.02	0.83	-0.05	1.17	n.s

Note: n.s means not significant at the 95% confidence interval level

As can be seen from Table 7.4, there were no significant differences between males and females in relation to each of the four constructs of classroom assessment literacy. This result indicates there is no evidence that gender is related to the four constructs (i.e., CAK, GB, IM and QP) of classroom assessment literacy.

To further investigate each classroom assessment literacy variable by gender, a number of cross-tabulations of band levels by gender were undertaken. Similar to the previous background characteristics groups, the cross-tabulation analyses and their results are depicted in each bar chart presented in Figure 7.5.

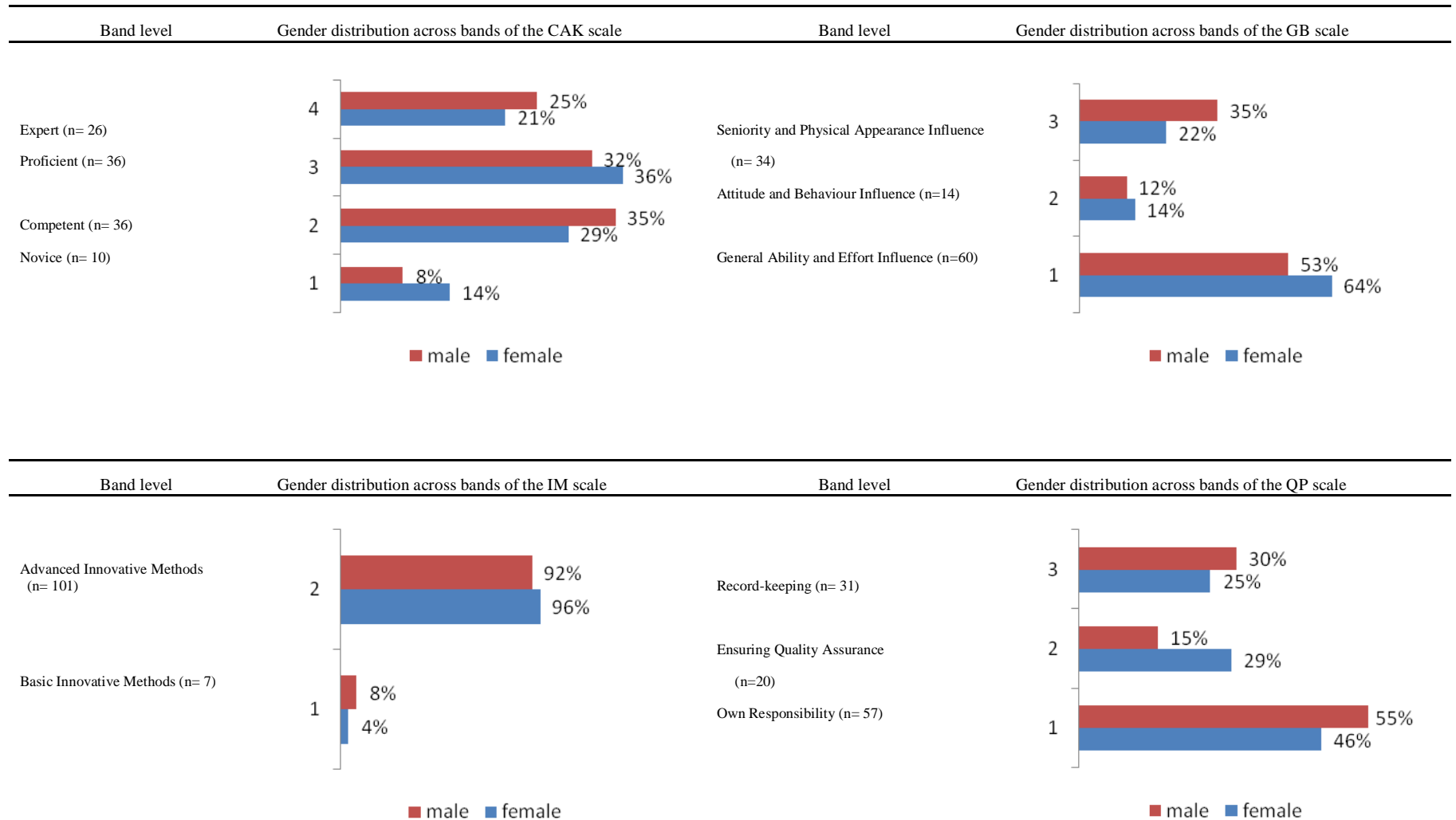


Figure 7.5 Recoded instructor gender variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

In relation to the Classroom Assessment Knowledge scale, Figure 7.5 showed that of the 80 male instructors who completed the Classroom Assessment Knowledge test, the majority were operating at Band Levels 2 (Competent), 3 (Proficient), and 4 (Expert), as were the female instructors. An examination of the band levels positioning of the instructors on the Innovative Methods, Grading Bias and Quality Procedure scales also indicated there was no marked difference between male and female instructors.

7.2.7 Classroom Assessment Literacy Variables as a Function of Departmental Status

To examine whether the departmental status (i.e., English-major versus English non-major departments) influences each of the four constructs of classroom assessment literacy, a series of independent t-tests were performed after the homogeneity assumption for equal variance was checked. Furthermore, in instances where a statistical difference has been found by the independent t-test, the effect size (Cohen, 1969) has also been calculated to help determine the size of the statistical difference (Izard, 2004b).

Table 7.5 Classroom Assessment Literacy Variables as a Function of Departmental Status

	English-major Instructors (N= 59)		English non-major Instructors (N= 49)		Significance of t-test	Effect Size (ES)
	Mean	SD	Mean	SD		
IM	0.17	1.18	-0.22	1.29	n.s	
QP	0.37	1.31	-0.34	1.41	P< .01	.52
GB	-0.48	1.76	0.64	1.76	P< .01	-.61
CAK	0.38	0.74	-0.45	0.93	P< .01	.99

Table 7.5 showed several significant findings. First, the independent two-tailed t-test revealed a significant difference for the quality procedure variable (QP ($t(106)=2.697$, $p< .01$), with English-major instructors having significantly higher average logit scores than English non-major instructors. Using Cohen's (1969) guidelines, such a difference was regarded to be a medium magnitude of effect size ($ES= .52$). This indicated that English-major instructors believed more in the importance of quality

assessment procedures than their English non-major counterparts. Secondly, the English-major instructors had significantly lower mean logit scores for the grading bias variable (i.e., GB) than English non-major instructors ($t(106) = -3.304$, $p < .01$). Moreover, the effect size of such a difference for both groups was found to be a medium and negative magnitude ($ES = -.61$). This suggested that English-major instructors were less influenced by students' personal characteristics (i.e., effort and/or attitude) than their English non-major counterparts. Third, the results shown in Table 7.5 revealed that the two groups of instructors had significantly different average logit scores for the classroom assessment knowledge variable (CAK ($t(106) = 5.181$, $p < .01$). Such a difference for both groups was further found to be a large magnitude of effect size ($ES = .99$). This indicated that the English-major instructors demonstrated a higher classroom assessment knowledge level than their English non-major counterparts.

To further investigate whether departmental status (i.e., English-major versus English non-major) has an impact on the band level positioning of each classroom assessment literacy variable, a number of cross-tabulations were undertaken. The cross tabulation analyses and their results displayed in each bar chart in Figure 7.6 followed a similar pattern as to that previously reported for gender groups.

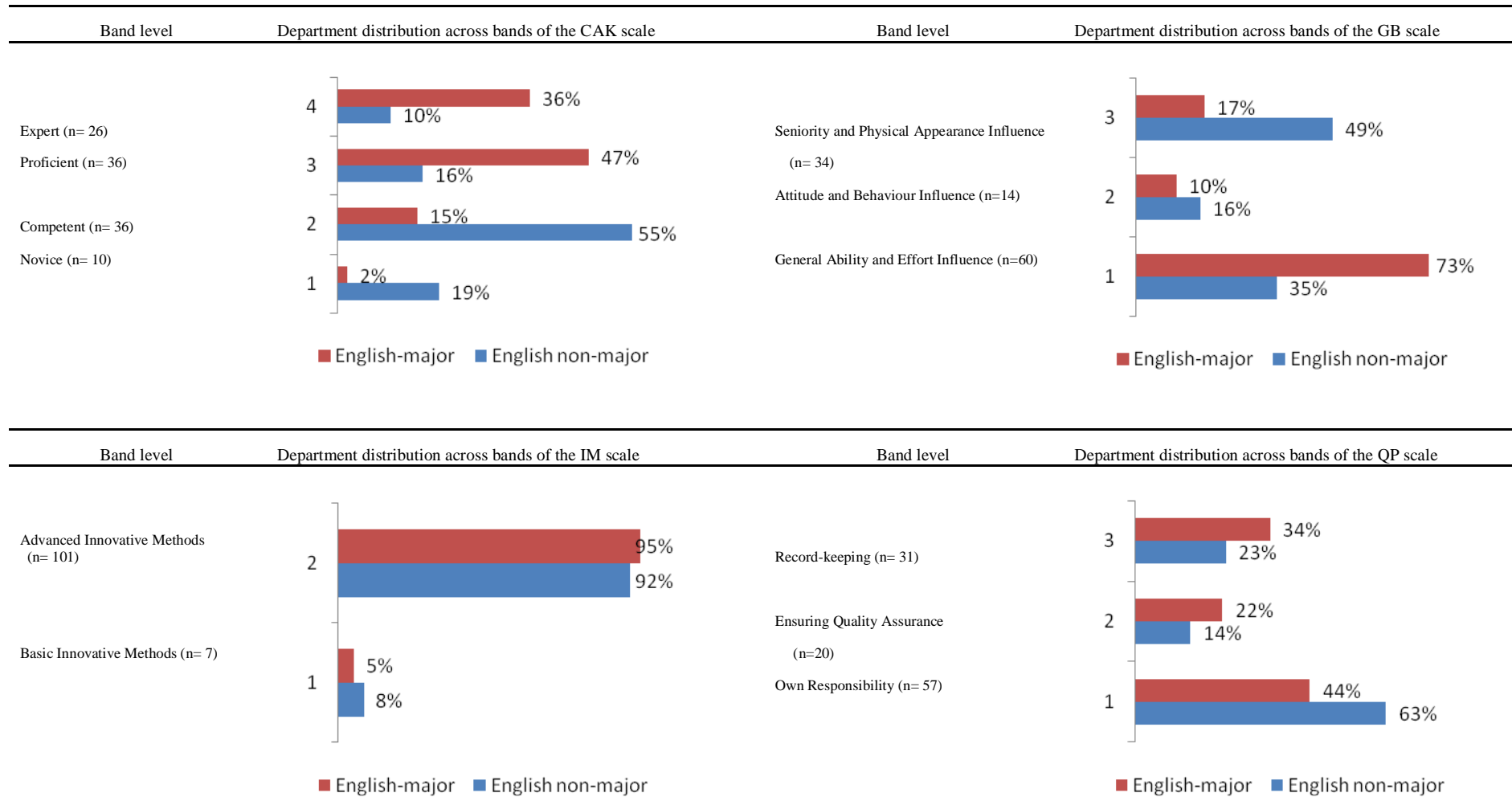


Figure 7.6 Recoded instructor department variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

Figure 7.6 illustrated that, of the 49 instructors who were classified as English non-major instructors on the Classroom Assessment Knowledge (CAK) scale, the majority (74%) were operating at Band Levels 1 (Novice) and 2 (Competent). In contrast, of those classified as English-major instructors, the majority (83%) were operating at Band Levels 3 (Proficient) and 4 (Expert). The percentage distribution patterns for the Grading Bias (GB) scale showed that more than half of English-major instructors were located at Band Level 1 (i.e., where they were typically influenced by students' general ability and effort when marking their work). However, about half of the English non-major instructors were located at Band Level 3 (i.e., where they were typically influenced by students' age, gender and appearance when marking their work). With respect to the Quality Procedure (QP) scale, more than half of the English non-major instructors (63%) tended to be performing at Band Level 1 (i.e., where they considered it their own responsibility to implement quality assessment procedures), while less than half of the English-major instructors were at Band Level 1. In relation to the Innovative Methods (IM) scale, both English-major and English non-major instructors tended to be at Band Level 2, where they valued the usefulness of employing reflective journal, portfolio, peer assessment and self-assessment in assessing students' learning.

Hence, a closer inspection of the band level attainment of instructors indicated that departmental status (i.e., English-major versus English non-major) was associated with their classroom assessment literacy level, which supported the findings from the independent two-tailed t-tests results that indicated there were different mean logit scores between English-major and English non-major instructors on three of the four measures (i.e., CAK, GB and QP) of classroom assessment literacy. The majority of English-major instructors demonstrated a higher level attainment for classroom assessment knowledge and a higher level of endorsement for quality assessment procedures than the majority of their English non-major counterparts. That is, most English-major instructors tended to have more favourable attitudes toward the importance of record-keeping and quality assurance than their English non-major counterparts. Furthermore, most English-major instructors appeared to be more influenced by students' general ability and effort, and less influenced by students' age, gender and appearance when marking students' work than their English non-major counterparts.

7.2.8 Classroom Assessment Literacy Variables as a Function of Academic Qualifications

To examine whether the level of academic qualifications of instructors influences each of the four classroom assessment literacy constructs, a series of independent t-tests were performed after the homogeneity assumption for equal variance was checked.

Table 7.6 Classroom Assessment Literacy Variables as a Function of Academic Qualifications

	Master (N= 64)		Bachelor (N= 44)		Significance of t-test
	Mean	SD	Mean	SD	
IM	0.08	1.23	-0.14	1.25	n.s
QP	0.14	1.38	-0.08	1.43	n.s
GB	-0.07	1.82	0.17	1.87	n.s
CAK	0.07	0.99	-0.09	0.83	n.s

As Table 7.6 shows, there were no statistically significant differences between the instructors who obtained a master's degree and those who had a bachelor's degree in relation to each of the four constructs of classroom assessment literacy. This result indicates that there is no evidence that academic qualifications have an association with the four constructs (i.e., CAK, GB, IM and QP) of classroom assessment literacy.

To further investigate whether the level of instructors' academic qualifications impacts the band level positioning of each classroom assessment literacy variable, a set of cross-tabulations were conducted. Prior to undertaking cross-tabulations, the instructors' academic qualifications were recoded into two groups as presented in Table 5.1 (see section 5.2.3.1 in Chapter 5). As with the previous background characteristics groups, the cross-tabulation analyses and their results illustrated in each bar chart are presented in Figure 7.7.

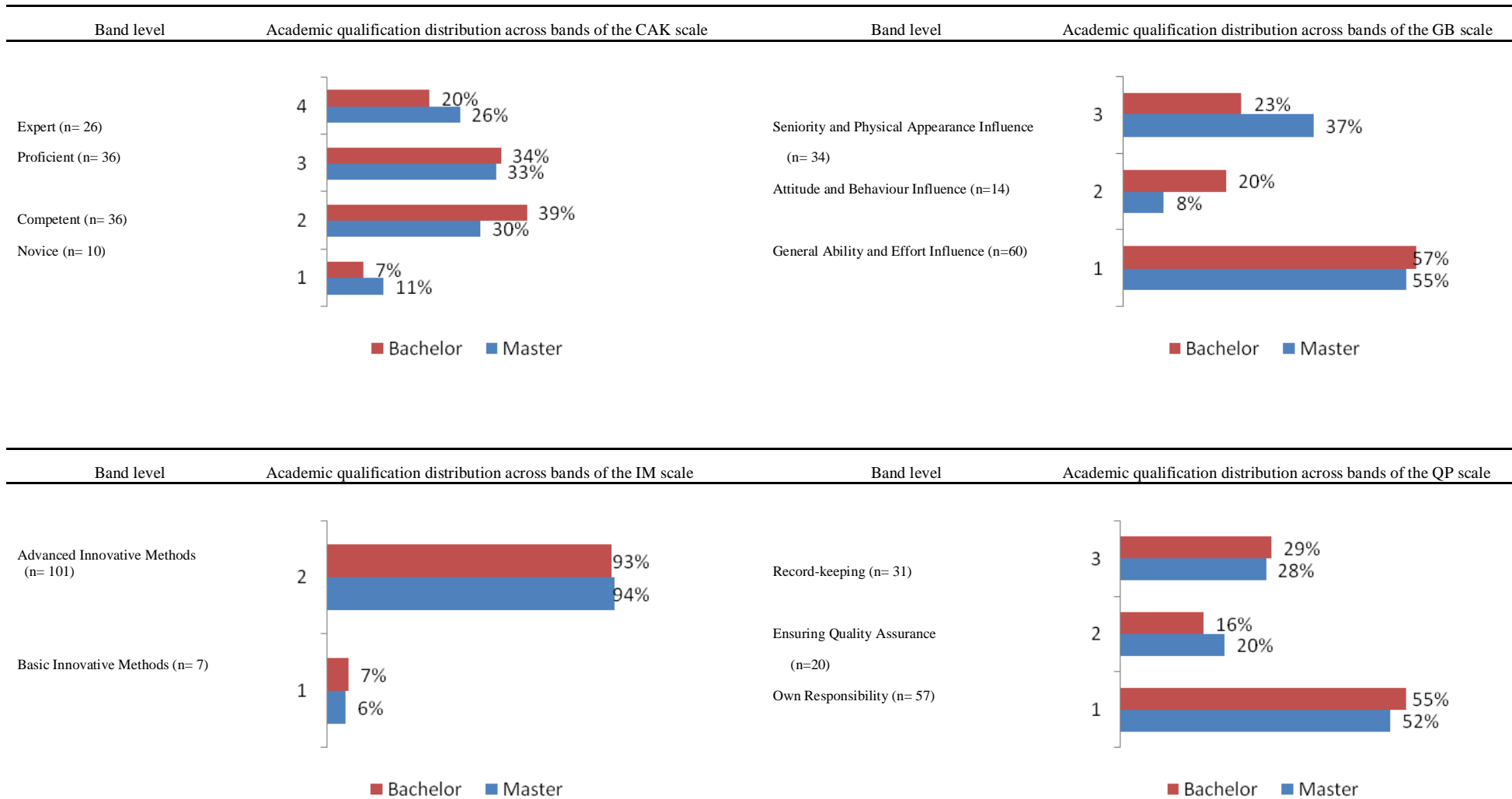


Figure 7.7 Recoded instructor academic qualification variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

With regard to the influence of academic qualifications on the level of instructor progression along the Classroom Assessment Knowledge (CAK) scale, Figure 7.7 showed that most instructors were operating at Band Levels 2 (Competent), 3 (Proficient) and 4 (Expert) irrespective of the level of the academic qualification they had. This indicated that these groups of instructors tended to predominantly operate at the middle and upper end of this scale. In relation to the Grading Bias (GB) scale, Figure 7.7 illustrated that there was approximately equal representation of instructors with a master's degree and bachelor's degree at Band Level 1 (55% and 57% respectively), where they were typically influenced by students' general ability and effort when marking their work. This suggested that their qualifications did not influence the amount of bias present in the marking process of instructors. Similarly, the percentage distribution patterns for the Innovative Methods (IM) scale showed that the majority of the instructors tended to operate at Band Level 2 (Advanced Innovative Methods) irrespective of the level of the academic qualifications they obtained, suggesting that these groups of instructors tended to mainly locate at the upper end of this scale. Similar trends were found for the Quality Procedure (QP) scale, whereby instructors with a master's degree tended to be performing at Band Levels 1 (Own Responsibility) and 3 (Record-keeping), as were instructors with bachelor's degrees, indicating that these groups of instructors tended to be predominantly represented at the lower and upper ends of this scale. Hence, a closer examination of the band level positioning of instructors indicated there was no marked difference between instructors with bachelor's degrees and master's degrees on each of the classroom assessment literacy measures.

7.2.9 Classroom Assessment Literacy Variables as a Function of Pre-service Assessment Training

To examine the influence of pre-service assessment training of instructors on each of the four classroom assessment literacy constructs, a series of independent t-tests were computed after the homogeneity assumption for equal variance was checked.

Table 7.7 Classroom Assessment Literacy Variables as a Function of Pre-service Assessment Training

	With Pre-service Assessment Training (N= 64)		Without Pre-service Assessment Training (N= 44)		Significance of t-test	Effect Size (ES)
	Mean	SD	Mean	SD		
IM	0.22	1.10	-0.34	1.31	$p < .05$.47
QP	0.28	1.32	-0.29	1.44	$p < .05$.41
GB	-0.21	1.75	0.37	1.93	n.s	
CAK	0.28	0.92	-0.39	0.79	$p < .01$.76

Table 7.7 showed several significant findings. First, the t-test revealed a significant difference for the innovative assessment methods variable (IM ($t(106) = 2.349$, $p < .05$). Furthermore, such a difference was found to be a medium magnitude of effect size ($ES = .47$). This indicated that the instructors who had pre-service assessment training during their undergraduate studies demonstrating that they believed in the usefulness of innovative assessment methods more than the instructors who had no pre-service assessment training. Second, the instructors with pre-service assessment training had significantly higher mean logit scores for the quality assessment procedures (i.e., QP) than instructors without pre-service assessment training ($t(106) = 2.133$, $p < .05$). Moreover, such a difference was found to be a small magnitude of effect size ($ES = .41$). This suggested that the instructors with pre-service assessment training demonstrated that they believed in the importance of quality assessment procedures (i.e., QP) more than those without pre-service assessment training. Third, the two groups of instructors had significantly different average logit scores for the classroom assessment knowledge variable (CAK ($t(106) = 3.892$, $p < .01$). Such a difference for both groups was also found to be a large magnitude of effect size ($ES = .76$). This indicated that the instructors who had pre-service assessment training demonstrated a higher classroom assessment knowledge level than those without pre-service assessment training.

To examine whether the duration of pre-service assessment training of the instructors influences each of the four classroom assessment literacy constructs, a series of independent t-tests were undertaken after the homogeneity assumption for equal variance was checked.

Table 7.8 Classroom Assessment Literacy Variables as a Function of Assessment Training Duration

	One or More than One Semester (s) Pre-service Assessment Training (N= 37)		Less than One Semester Pre-service Assessment Training (N= 27)		Significance of t-test
	Mean	SD	Mean	SD	
IM	0.30	1.14	0.11	1.16	n.s
QP	0.30	1.43	0.27	1.19	n.s
GB	-0.46	1.81	0.14	1.63	n.s
CAK	0.44	0.94	0.05	0.84	n.s

Table 7.8 showed that the independent two-tailed t-test revealed no significant differences between these groups of instructors in relation to each of the four classroom assessment literacy constructs. This result indicates there is no evidence that the duration of pre-service assessment training impacted the four constructs (i.e., CAK, GB, IM and QP) of classroom assessment literacy.

To further examine the association of instructors' perceptions about the level of their classroom assessment preparedness gained from their pre-service assessment training with each of the four classroom assessment literacy constructs, a series of one-way Anova tests were conducted.

Table 7.9 Classroom Assessment Literacy Variables as a Function of the Level of Preparedness of Assessment Training

	Very Prepared (N= 13)		Prepared (N= 46)		Unprepared (N= 5)		Significance of one-way Anova test	Effect Size (ES)
	Mean	SD	Mean	SD	Mean	SD		
IM	0.01	1.15	0.30	1.15	0.02	1.23	n.s	
QP	0.26	1.42	0.24	1.37	0.77	0.34	n.s	
GB	0.36	1.78	-0.41	1.78	0.15	1.15	n.s	
CAK	-0.29	0.62	0.46	0.96	0.03	0.49	P< .01	-.82

As can be seen from Table 7.9, there was one significant difference amongst these groups on the classroom assessment knowledge scale (CAK ($F(3,107)= 8.230$, $P< .01$). A

close look at the result of the Post Hoc test, Bonferroni, revealed that there was a significant difference only between the instructors who perceived they were “*prepared*” and those who perceived they were “*very prepared*” at the .05 level. Such a difference for both groups was further found to be a large and negative magnitude of effect size ($ES = -.82$). This suggested that the instructors who perceived they were “*prepared*” demonstrated a higher classroom assessment knowledge level than those who perceived they were “*very prepared*” (refer to section 9.2.2 in Chapter 9 for a detailed explanation).

To further investigate the influence of the instructors’ assessment training on the band level positioning of each of the classroom assessment literacy constructs, a set of cross-tabulations were undertaken. Similar to the previous background characteristics groups, cross-tabulation analyses and their results have been presented in Figure 7.8.

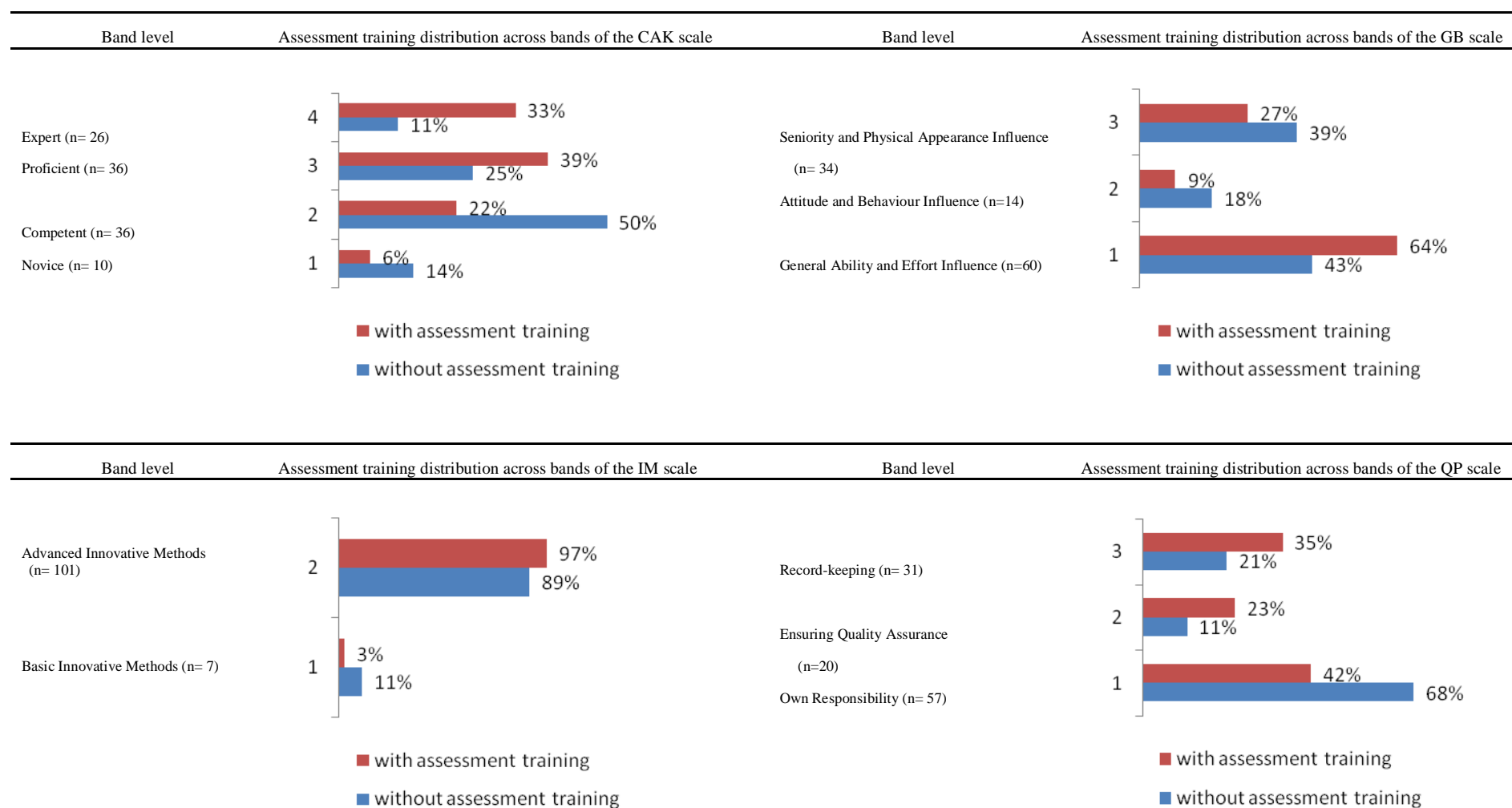


Figure 7.8 Recoded instructor assessment training variable across the band level of the CAK, GB, IM, and QP scales

Note: CAK= Classroom Assessment Knowledge, GB= Grading Bias, IM= Innovative Methods, and QP= Quality Procedure

Figure 7.8 illustrates that more than half of the instructors who had received the assessment training were performing at Band Levels 3 (Proficient) and 4 (Expert) on the Classroom Assessment Knowledge scale (72%). In contrast, a large proportion of those who had no pre-service assessment training (64%) were performing at Band Levels 1 (Novice) and 2 (Competent). In relation to the Quality Procedure scale, more instructors having assessment training seemed to progress to Band Levels 2 (Ensuring Quality Assurance, 23%) and 3 (Record-keeping, 35%), whereas most instructors without assessment training seemed to be operating at Band Level 1 (Own Responsibility, 68%). With respect to the Grading Bias scale, Figure 7.8 showed that more than half of the instructors with assessment training (64%) tended to be at Band Level 1 (i.e., where they were typically influenced by students' general ability and effort when marking their work) as opposed to those without assessment training (43%). However, more instructors without assessment training (39%) tended to be represented at Band Level 3 (i.e., where they were typically influenced by students' age, gender and appearance when marking their work) as opposed to those with assessment training (27%). The percentage distribution patterns for the Innovative Methods scale showed that, irrespective of assessment training, most instructors tended to be performing at Band Level 2 (Advanced Innovative Methods).

Thus, a closer inspection of the band levels attainment of instructors suggests that pre-service assessment training influences instructors' classroom assessment literacy, which supports the independent two-tailed t-test results that indicate there were statistically significant differences between the instructors with assessment training and those without assessment training on three of the four constructs (i.e., CAK, IM and QP) of classroom assessment literacy. The instructors with pre-service assessment training demonstrate a higher classroom assessment knowledge level, as well as a higher level of endorsement towards the use of innovative assessment methods and quality assessment procedures in their classroom assessment practices than instructors without pre-service assessment training. That is, the instructors with pre-service assessment training tended to have more favourable attitudes toward the use of self- and peer assessments, portfolios, reflective journals and individual conferences as well as the use of record-keeping and quality assurance in their assessment practices than those without assessment training. A

closer inspection of the band level attainment also suggests that instructors with pre-service assessment training appeared to be less influenced by students' attitude, behaviour, age, gender and appearance, and they were more influenced by students' general ability and effort when marking their work than those without assessment training.

7.3 Multivariate Results

This section presents the multivariate results with regard to the evaluation of the extent to which the hypothesised one-factor congeneric measurement model fitted the data through the use of confirmatory factor analysis (CFA).

7.3.1 Congeneric Measurement Model Development

Using the confirmatory factor analysis (CFA), a one-factor congeneric model was tested to determine how well this model fitted the data (refer to section 5.2.4.2 in Chapter 5). This model was defined by the interrelationships amongst the four constructs of classroom assessment literacy (i.e., IM, QP, GB and CAK). These four variables were considered as the primary constructs of classroom assessment literacy. Furthermore, the second order construct for the one-factor congeneric model was named "Classroom Assessment Literacy". The assessment of this hypothesised measurement model was determined by the statistical indices of fit and the theoretical basis of the model.

7.3.1.1 One-factor Congeneric Model: Classroom Assessment Literacy

The one-factor congeneric model was assessed with all primary variable parameters set free and the variance of the latent, second order variable (i.e., Classroom Assessment Literacy) fixed to 1. Prior to performing the confirmatory factor analysis, the assessment of the normality of each variable as well as the distribution of normality of all variables (Mardia's (1970, 1974) estimate of multivariate kurtosis) was computed. The tests indicated that each scale had a univariate normality and the distribution of all scales

had a multivariate normality (refer to Table 7.2). Thus, the assumption of multivariate normality of all scales was met.

Inspection of the absolute fit of the model revealed that the one-factor congeneric model had a non significant Chi-square (χ^2) of 0.68, with 2 degrees of freedom, and a p-value of 0.967, with a fit ratio (χ^2/df) of 0.34, a goodness of fit index (GFI) of 1.00, and an adjusted goodness of fit index (AGFI) of 1.00. The GFI indicated that the model explained 100% of the variance in the current data set and the AGFI also suggested that the model fitted the data well. Moreover, the model generated a Root Mean Square Error of Approximation (RMSEA) of 0.00, with a test of the hypothesis that $\text{RMSEA} \leq 0.05$ (also referred to as PCLOSE) of 0.97, and a χ^2/df ratio of 0.34, suggesting the model fitted the data well. An examination of comparative fit of the model showed that Normed Fit Index (NFI) of 1.00, Comparative Fit Index (CFI) of 1.00 and Tucker-Lewis Index (TLI) of 1.00, indicating a good model fit. Further inspection of the modification indices and residuals indicated no further ways to improve the model from a statistical perspective. Hence, this model was found to fit well to the current study data.

Table 7.10 displays all standardised parameter estimates (λ_x), critical ratio (CR also referred to as t-value), standard error of estimate for each parameter (SE), and the proportion of variance of the observed variable accounted for by the second order factor-Classroom Assessment Literacy (R^2). Table 7.11 illustrates the goodness of fit measures for the one-factor congeneric model.

Table 7.10 Maximum-likelihood (ML) Estimates for One-factor Congeneric Model:

Classroom Assessment Literacy					
Scale	Variable Name	λ_x	CR	SE	R^2
Innovative Methods	IM	0.55	4.88	0.14	0.31
Quality Procedure	QP	0.49	4.30	0.16	0.24
Grading Bias	GB	-0.66	-5.69	0.21	0.44
Classroom Assessment Knowledge	CAK	0.58	5.04	0.11	0.33

Table 7.11 Goodness of Fit Measures for One-factor Congeneric Model: Classroom Assessment Literacy

Goodness of Fit Index	Value
Absolute Fit	
Chi-square (χ^2)	0.68
Degree of freedom (df)	2
Fit ratio (χ^2/df)	0.34
P-value	0.967
Goodness of Fit index (GFI)	1.00
Adjusted Goodness of Fit index (AGFI)	1.00
Root Mean-Square Error of Approximation (RMSEA)	0.00
A test of the hypothesis that $\text{RMSEA} \leq 0.05$ (also known as PCLOSE)	0.97
Comparative Fit	
Normed Fit Index (NFI)	1.00
Comparative Fit Index (CFI)	1.00
Tucker-Lewis Index (TLI)	1.00

The results showed that the four variables had statistically significant regression path coefficients with the classroom assessment literacy latent construct. These variables comprised IM, QP, GB and CAK. The GB variable had the largest standardised loading of -0.66 and a squared multiple correlation coefficient of 0.44, suggesting that the classroom assessment literacy construct accounted for 44% of the variance in the GB variable. In addition, the classroom assessment literacy construct further accounted for 33% variance in the CAK variable, 31% in the IM variable, and 24% in the QP variable. Hence, the R^2 values for each observed variable suggest that the four variables serve well as a measure of the single latent Classroom Assessment Literacy construct.

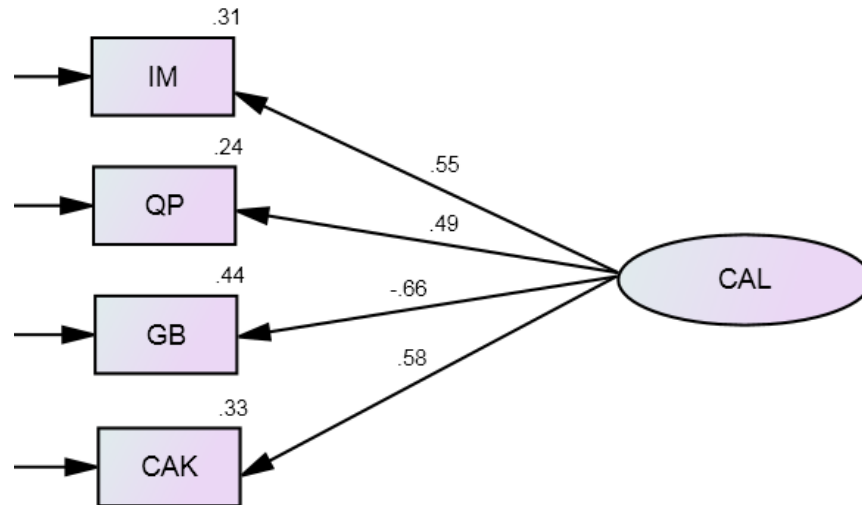


Figure 7.9 One-factor congruic model: Classroom Assessment Literacy

In the next chapter, the results from the qualitative phase of the study are reported.

Chapter 8: Qualitative Results

This chapter is organised into five sections. Section 8.1 presents the content analysis of learning goals documents of two university departments selected for the current study. Section 8.2 presents the content analysis of departmental assessment-related policies and procedures documents that governed the classroom-based assessment practices within these two departments. Section 8.3 presents the general background characteristics of the six instructors who participated in the semi-structured interviews. Section 8.4 presents the thematic analysis of the semi-structured interviews, with a particular focus on classroom assessment literacy. The final section (see 8.5) provides the summary of key findings regarding how this classroom assessment literacy relates to departmental assessment-related policies and procedures, as well as background characteristics of instructors.

8.1 Learning Goals of University Departments

Although the two university departments selected for the current study (i.e., English-major and English non-major departments) were located within the same Cambodian city-based university, they differed in terms of their learning goals and programmes delivered. The English-major department offered a four-year Bachelor of Education in Teaching English as a Foreign Language (TEFL) degree as well as a Bachelor of Arts in English for Work Skills (EWS) degree to English-major students over an eight-semester period. The department had the following three core learning goals that underpinned the delivery of its two degree programs:

- To develop highly-qualified students who are able to contribute to the labour market needs and national moral development;
- To prepare students for lifelong independent learning and to promote patriotism and community service learning; and
- To address the market needs of IT-assisted EFL education, research and management.

In contrast, the English non-major department offered a three-year English Enrichment non-degree programme to English non-major undergraduate students who studied in all university departments, with the exception of English-major students, over a six-semester period. The department's primary learning goal was to provide English language training to English non-major students in order to develop their English language proficiency, to ensure they could read and write academic work in English, in order to meet their goals in major subjects.

Hence, there were differences in the core learning goals of the two departments. The English-major department appeared to have broader learning goals than the English non-major department. That is, English-major departmental learning goals tended to mainly focus on developing high independent learning skills, knowledge and attributes (i.e., creativity, critical thinking, problem-solving, decision-making, flexibility, initiative, appreciation for diversity, communication, collaboration and responsibility) that matched the needs of the local and/or global labour market and society. In contrast, the English non-major department tended to emphasise developing students' high English language proficiency in reading and writing domains to enable them to read and write in English in their major subjects.

8.2 Departmental Assessment-related Policies

A content analysis of both departments' assessment-related policy documents was undertaken to identify similarities and/or differences within each policy feature. The comparisons have been presented in Table 8.1.

Table 8.1 Assessment Policies of the English-major and English Non-major Departments

Feature	English-major Department	English Non-major Department
Purpose of assessment	Summative function	Summative function
Preferred methods	Traditional Assessment (i.e., tests and exams) and Innovative Assessment (i.e., performance-based assessments)	Traditional Assessment (i.e., tests and exams) and Innovative Assessment (i.e., performance-based assessments)
Assessment weightings		
Ongoing Assessment	50%	60%
Final Examination	50%	40%
Timing of assessments	At the end of every lesson, month, and semester	At the end of every lesson, mid-semester, and semester
Location	Classroom	Classroom
Instructor roles/responsibilities in relation to assessments	All types of assessments developed by subject instructors	All types of assessments developed by subject instructors
Assessment Responsibility		
• Task Design & Development	Individual subject instructor	Shared between- all instructors within the subject
• Task Administration		
• Marking/Grading		
• Moderation	Administrators and subject instructors	Administrators and subject instructors
• Record-keeping & Reporting Review		
Cut-points for pass	50% or 50 marks	50% or 50 marks
Required compulsory student class attendance for passing the subject	70% or 80%	66%

Table 8.1 shows that both departments administered a combination of ongoing assessments and final examinations to determine students' learning achievements/grades. The ongoing assessment administered within the English-major department typically accounted for 50% of the final grade and comprised homework, class tests, oral presentations, assignments and class participation. Whereas, the English non-major department's ongoing assessment, worth 60%, typically included quizzes, mid-term exams, written paragraphs/essays and class participation. In both departments, final examinations were administered at the end of each semester, and were worth 50% and 40% for the English-major and English non-major departments respectively.

With regard to the conduct of classroom-based assessment, instructors within both the English-major and English non-major departments were responsible for implementing the entire assessment process in assessing students' learning achievements. Within the English-major department, the development of the ongoing assessment (comprising assessment topic selection, task specifications and marking criteria) and the final

examination was the responsibility of the team of instructors who taught the same subject. Individual instructors, however, were responsible for constructing ongoing assessment tasks including quizzes and class tests. In contrast, within the English non-major department, assessment task development was the shared responsibility of all subject instructors who taught the same subject. In all instances, a cut-off score for a subject pass for both departments had been set at 50%, which was compulsory at the university level. Within both departments, instructors were typically responsible for reporting each type of ongoing assessment grade (i.e., numerical grades) to their students in classes after completion of each assessment task. Instructors also submitted the records of their final course grades, comprising all types of ongoing assessment grades and final exam grades, in terms of numerical grades to the administrators at the end of the semester in a timely manner. Students' final course grades (numerical grades) were typically reported by the administrators via posting them publicly on the departmental noticeboard in order to communicate the learning achievements and/or grades to students.

Typically, students in the English-major department were required to take three or four subjects per semester depending on their year levels, as opposed to students in the English non-major department, with a requirement of taking one English related subject. To be promoted to the next level, students had to pass each subject and satisfy the attendance requirements. That is, they had to achieve a minimum grade level expectation of 50% for each subject in order to successfully pass. With respect to attendance policy, students' class attendance was compulsory in both the English-major and English non-major departments. The percentage of required attendance for English-major students was either 70% or 80%, depending on the year levels and status of the students (scholarship versus fee paying), as opposed to English non-major students, who were required to have at least a 66% attendance rate. If students failed the required attendance of a particular subject, their exam grades for that subject, irrespective of how high they were, were considered invalid. In other words, they would repeat that subject automatically if their attendance was less than 70% and 66% for English-major and English non-major departments respectively. Within the English-major department, if students failed three subjects in one semester after taking the supplementary exams, they had to repeat all subjects within that particular semester. If students failed three subjects

in both semesters, they would not be promoted to the next level. In contrast, within the English non-major department, if students failed the compulsory English subject, they were required to take remedial classes offered in the summer school programme for a one-month period.

Hence, there appeared more similarities than differences with regard to the two departments' assessment-related policies. In relation to commonalities, both departments had a compulsory attendance policy for all students. Failure to fulfil the required attendance could severely impact students' academic success. Such an attendance policy is interesting, given that one of the departmental learning goals was about lifelong learning, which is typically associated with recognising that learning can occur anywhere, and not just be confined to the classroom or school settings (see Guskey, 2013). Furthermore, student learning outcomes were assessed by ongoing assessments and final examinations, which were strongly based on traditional assessment methods such as quizzes, tests and examinations for summative purposes. This process indicates that the two departments' assessment policies regarded classroom assessment as an end-point judgment, serving the purposes of promotion and/or certification rather than an integral part of teaching and learning (i.e., serving both formative and summative purposes). The absence of formative assessment in the departments' assessment policies is surprising, given that considerable research has indicated that assessment for formative purposes has the potential to promote students' independent lifelong learning skills and enhance their learning development (Black & Wiliam, 1998a; Shute, 2008).

In both departments, instructors were responsible for reporting numerical grades to their students after completion of each ongoing assessment task, as well as record their final course grades (i.e., numerical grades) to the administrators at the end of the semester. Typically, the administrators posted the final course grades publicly on departmental noticeboards to communicate them to students. This type of reporting (i.e., merely numerical grades) provides limited information to all relevant assessment stakeholders, such as students, parents, administrators and instructors, as assessment results fail to convey the students' learning progress adequately (Black & William, 1998a; Brookhart, 1999). In other words, this type of reporting simply reflects that classroom assessment is mainly used to judge students' learning success on discrete

bodies of taught content in the subject. Posting the students' final course grades publicly further raises questions on the ethicalness or fairness on the part of students (i.e., consequential validity), as it violates students' privacy (i.e., course grades), which can affect their self-image. Bachman and Palmer (2010) and Miller et al. (2013) strongly recommend that assessment reports need to be kept confidential in order to protect the rights of students with respect to fundamental fairness.

At the university level, a cut-off score for pass has been set at 50% (or 50 marks) with respect to promoting students to the next level. Such a decision-making model, based on score totals, is commonly referred to as percentage grading (Sadler, 2005; Guskey & Jung, 2013; Quinn, 2013; Waugh & Gronlund, 2013). While this model gives the impression of precision and it is easy to operationalise, it has been criticised on the grounds that the cut-off scores are not related to the mastery of specific skills or learning outcomes, and it is typically left to the instructors to work out the cut-off scores for each assessment. Such a model raises concerns about how the marks are generated in terms of validity, sampling adequacy, assessment task quality, marking standards and marking reliability (Sadler, 2005; O'Connor, 2009; Guskey & Jung, 2013; Quinn, 2013; Waugh & Gronlund, 2013).

Finally, there was one marked difference with respect to the two departments' assessment policies in dealing with students who failed the subject. The English-major department provided supplementary examinations to students who failed the subject for each semester, as opposed to the English non-major department, which only offered remedial classes in the summer school programme for one month. The different ways of dealing with students who failed the subject may be due to different learning goals set by these two departments.

8.3 Background Characteristics of the Interviewees

The six instructors who participated in semi-structured interviews were of a similar age, but varied somewhat in academic qualifications and years of teaching experience (see Table 8.2). In the following sections, the coded "LM" represents the English-major instructor and the coded "LN" indicates the English non-major instructor.

Table 8.2 Background Characteristics of the Interviewees

Instructor Code	English Department		Gender		Highest Qualification in TEFL/TESOL	Teaching Experience (years)	Age (years)
	Major	Non-major	Male	Female			
LM1	<input checked="" type="checkbox"/>			<input checked="" type="checkbox"/>	Bachelor	11	34
LM2	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		Masters	8	31
LM3	<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>		Bachelor	5	30
LN4		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	Bachelor	6	31
LN5		<input checked="" type="checkbox"/>		<input checked="" type="checkbox"/>	Bachelor	5	30
LN6		<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>		Bachelor	4	31

All six instructors in Table 8.2 had at least a bachelor degree qualification in TEFL from the same city-based university within Cambodia. The three English-major instructors pursued further studies internationally, graduating with different master's degrees and areas of specialisation. Instructor LM1 had a master's degree in Leadership and Management from one Australian university, while instructor LM2 obtained a master's degree in Teaching English to Speakers of Other Languages (TESOL) from one American university. In contrast, instructor LM3 received a master's degree in Literature from one university in the Philippines. The three English non-major instructors, however, were not given the opportunity to pursue further studies after completing their undergraduate studies in Cambodia.

8.4 Classroom Assessment Literacy

A thematic analysis of semi-structured interviews identified three main themes related to instructors' classroom assessment literacy: (1) their perceived assessment competence, (2) their notions of the ideal assessment, and (3) their knowledge and understanding of the concepts of validity and reliability. In the following sections, each theme associated with instructors' classroom assessment literacy has been presented, together with a discussion of how departmental assessment policies and instructors' background-related factors could influence their classroom assessment literacy.

8.4.1 Perceived Assessment Competence

Because perceived competence can affect instructors' expended effort, persistence, motivation and confidence (Bandura, 1997) in implementing high quality assessments, it is important to better understand each instructor's perception of his/her expertise in classroom assessments. In an attempt to uncover their self-awareness of the current level of assessment competence, the instructors were asked to score themselves on their knowledge and skills in classroom assessment out of a maximum score of 10, in which a 10 would mean they had mastered all the knowledge and skills (i.e., competence) required to conduct a good quality classroom assessment. They were also asked to justify their score. Table 8.3 displays the self-reported measures of perceived classroom assessment competence for each instructor.

Table 8.3 Self-reported Measure of Instructor Classroom Assessment Competence

Instructor Code	Self-rating (1-10 score range)
LM2	10
LN4	8
LN6	7
LM3	6-7
LM1	5-6
LN5	5

Five instructors were of the views that their current assessment competence was limited, while one instructor perceived his assessment competence as sufficient for their teaching role within the university. Instructor LM2, who had eight years teaching experience and held a masters in TESOL from a USA university (see Table 8.2), perceived his classroom assessment knowledge and skills were at the highest level (i.e., a rating of 10). He argued that he was able to construct assessment tasks as well as independently undertake research on his own assessment practice to improve the quality of his assessment implementation. Specifically, he justified:

I would give myself a 10...I think to be sufficient in doing any kind of testing and assessments is to say that a person can independently research...[his] daily practice with a student and required the ability to collect information related to the use of different tests and assessments devices in the classroom and it required reflection from that also...And I am able to do it.

LM2

However, instructors LN4, LN6 and LM3, who had on average five years' teaching experience, all with bachelor degrees in TEFL, self-rated their classroom assessment competence as above moderate (i.e., a rating of 6-8). In particular, instructor LN4, who had six years' teaching experience, argued that her classroom assessment competence was above moderate (i.e., a rating of 8), because she could conduct a good quality classroom assessment. She asserted she could implement good quality assessments because she was familiar with the subject being taught and got to know her students well, indicating that she associated knowledge in her teaching domain as synonymous with her knowledge and skills in designing and implementing classroom assessment. For example, she justified:

[My assessment knowledge and skills was] 8...Because for the assessment, one thing...[I] just know clearly about the subject what...[I am] going to teach...And one more thing...[I] know the students...So, when...[I] know...[my] subject clearly and...[I] know the students, ...[I] can design the test better.

LN4

Similarly, instructor LN6, who was the least experienced teacher, also scored his classroom assessment competence as above moderate (i.e., a rating of 7). He defended his rating on the grounds that he could adapt the assessment material from the available test books to suit his students' levels and needs, although he admitted that his capacity in adapting assessment material needed further improvement. As he asserted:

Yes, [I score my current assessment knowledge and skills] 7. The reason is that you know whenever I design the quizzes or the final exams, let me tell [you] about my materials. I used World English...[as] the materials, [and] actually World English has the CD-Rom. The CD-Rom has the tests, [and] the tests include the vocabulary, grammar, reading and writing...I don't actually follow everything from the CD-Rom...[I] try to make it [test] works for the students. So I guess the way that...[I adapt the test] is not really perfect.

LN6

Similarly, instructor LM3, who had five years' teaching experience and held a masters degree in Literature, argued that he rated his classroom assessment competence

as above moderate (i.e., a rating of 6-7), because of his perceived limited capacity to implement various types of assessments methods that could provide his students with opportunities to critically self-reflect on their learning material, rather than require them to remember details and facts from the learning material. Specifically, he justified:

May be [I scored myself] 6 to 7 [at] the most...Because after the tests, most of the time I just feel sometimes students are doing well in my class or read a lot, but when I ask them to answer questions, and then some of them cannot answer because...some questions are too detailed on the factual information in the [course] books, not all the questions are about what they can reflect...there must be a way...[I can] learn to do better next time...If...[I] don't really know how to make it better and then...[I] will keep doing the same thing.

LM3

In contrast to her counterparts, instructor LM1, who was the most experienced teacher and had a master's degree in Leadership and Management, scored her classroom assessment knowledge and skill level as moderate (i.e., a rating of 5-6). She argued that she scored herself at this level because she had learnt classroom assessment a long time ago, and since then had had little or no opportunity to share/discuss assessment-related ideas and understandings with her younger colleagues owing to a perceived generational gap. As she justified:

Okay let [']s say 5 or 6...what I have learned from Teaching Methodology, it was more than 10 years ago...I have learned a lot from my former colleagues and now they [have] moved...to other [work] places. And now we have...new generations, the gap between the young generation and the senior generation...becomes, you know bigger and bigger. And team work...you know we have less time to share our experience with each other...I think I should have been trained more often.

LM1

Similarly, instructor LN5, who had five years' teaching experience and held a bachelor degree in TEFL, also self-assessed her assessment competence as at the moderate level (i.e., a rating of 5). She reported she did not have an opportunity to undertake further study regarding classroom assessment-related principles and procedures after completing her undergraduate studies within her teacher preparation programme in Cambodia. Specifically, instructor LN5 justified:

I think I scored 5 for myself...Actually, I don't have great amount of knowledge in terms of testing...after four years [undergraduate study] at the English department, I...have [not had] any chance, okay, to further [my] study about testing.

It appears as though the instructors' pre-service training and professional development opportunities in educational assessment were the underpinning factors that influenced their self-ratings of their current classroom assessment knowledge and skills levels. For example, instructor LM2, who scored his assessment competence at the highest (i.e., a rating of 10), asserted that whilst his pre-service assessment training in Cambodia was inadequate, he was fortunate to have had further opportunities to study educational assessment overseas via a scholarship fund, and to attend a number of international conferences focusing on educational assessment. As he explicitly stated:

I took a subject called Teaching Methodology in semester 2...[I] had [teaching] practicum [too]...the material I read, the material that is presented in class by teachers are very much related to different types of tests, the purposes of using different types of the tests, different test items, how different test items are used, reliability issues, validity issues, and we discussed that very broadly and then my experience go beyond just this four-year degree program at the English department. I'd been to five months' training in Applied Linguistics at the Regional Language Centre in Singapore and I took a subject called Language Testing...besides courses I had attended, I had also been to several international workshops, seminars, conferences, in which some of the papers had been focused on the issues of testing and assessments...[through this further assessment training] I think I had been very prepared in the knowledge of testing and assessments.

LM2

Unlike their counterpart, the five remaining instructors (i.e., LN6, LM3, LN5, LM1, and LN4) asserted they did not have further opportunities to study educational assessment after they completed their teacher preparation programme in Cambodia. They also reported that the insufficient pre-service assessment training impacted their current classroom assessment expertise. They commented that the assessment units they took were embedded within the Teaching Methodology subject, which also included the Teaching Practicum component for only one semester period. They further pointed out that the assessment training emphasised mainly the traditional assessment theories and provided them with little opportunity to develop the actual assessment tasks. Instructor LN5, who scored her assessment competence the lowest (i.e., a rating of 5), also stated that she learned the assessment units from an instructor who lacked expertise in educational assessment. For example, one instructor reported:

Yes, [I learned assessment units from Teaching] Methodology subject...[I] learned...[mostly] in the theories...[and I] put...[them] into practice [when I am an in-service university lecturer]...[I] just try to apply [them into practice]...they [course instructors] just teach [traditional assessment theories], but they don't ask us to design the tests... it [assessment training] would not be sufficient.

LN4

A lack of access to in-service professional development workshops on educational assessment within the two departments also appears to relate to the instructors' self-ratings. For example, three of the six instructors (i.e., LN4, LM1 and LM3) reported that the professional workshops provided by their departments mostly emphasised teaching techniques and/or other relevant teaching aspects. They therefore lacked the opportunities to learn current, innovative educational assessment-related principles and procedures. As such, their assessment knowledge and skills would remain at the level when they received their pre-service assessment training, or even decrease. Specifically, they asserted:

Just one or two workshops [provided by the department since I have worked here]...related to assessments...[other] workshops...related to our teaching, how to teach the students in this way or that way.

LN4

Sometimes...[I] have [professional] workshops...It's [professional workshop] not about assessments.

LM1

Most of the professional workshops [provided in the department] focus more on teaching basically, but not really on assessments...Most of the time we deal with techniques in teaching.

LM3

To validate the six instructors' self-reported levels of assessment competence, their self-ratings were compared with their classroom assessment knowledge test results (i.e., their band levels attainment obtained through the Rasch analysis administered in phase one of the study) (refer to Table 6.2 in Chapter 6 that reported the band levels attainment).

Table 8.4 Validation of the Self-reported Measure of Instructor Classroom Assessment Competence

Instructor Code	Phase 2: Self-reported Measure	Phase 1: Assessment Knowledge Test	
	Self-rating (1-10 score range)	Assessment Knowledge Test Result: Logit Score	Band Level
LM2	10	1.34	Level 4: Expert
LN6	7	0.43	Level 3: Proficient
LM3	6-7	0.03	Level 3: Proficient
LN5	5	0.03	Level 3: Proficient
LM1	5-6	-0.77	Level 2: Competent
LN4	8	-1.20	Level 1: Novice

As can be seen in Table 8.4, all six instructors were placed in various band levels ranging from level 1 to level 4 with regard to their performance on the Classroom Assessment Knowledge (CAK) scale administered in phase one of the study. For example, instructor LM2, who had an international master's degree in TESOL, was located in Band Level 4: Expert, which was the highest level on the CAK scale.

Five of the six instructors interviewed had self-ratings that were consistent with their performance on the CAK scale. Whereas, one instructor (i.e., LN4) scored her classroom assessment competence considerably higher (i.e., 8 out of 10) than her performance on the CAK scale (i.e., Level 1: Novice), indicating that she grossly overrated her classroom assessment knowledge level. While it could be possible that she perceived that her high level skills in assessment (which in this study has not been objectively measured) compensated for her limited knowledge (and thus she rated herself highly), it is more likely that such a discrepancy was due to the social desirability response bias (i.e., the overreporting of socially desirable behaviours and underreporting of socially undesirable behaviours) that typically occurs with the use of self-reported measures (see section 9.2.2 in Chapter 9 for a detailed explanation).

In summary, it can be seen that five out of six instructors perceived their current level of classroom assessment competence was limited for their teaching role within the university. Such perceptions were in accordance with their actual performance on the objectively administered and scored Classroom Assessment Knowledge test.

8.4.2 Notion of the Ideal Assessment

Because instructors are key agents in the classroom assessment process (Klenowski, 2013a) and their personal beliefs have the potential to impact on their actual assessment implementation (Rogers et al., 2007), it is essential to have a better understanding of their assessment related beliefs. To examine their beliefs of high quality classroom assessments, the instructors were asked to describe what their ideal assessments would look like, if they had the time and resources to design and administer good quality assessments.

In relation to instructor LM2, who can be classified as a “very confident expert” assessor (see Table 8.4), he asserted that his ideal assessment would entail the use of portfolio assessment, as it could generate accurate learning achievements of the students in the course, despite the fact that it demanded more effort and time to assess and provide feedback. This indicates that his assessment competence influenced him in terms of placing greater value and endorsement towards the use of innovative assessment to assess his students’ learning. As he stated:

I believe in what they call a portfolio type of assessment which is very rigorous, which takes a lot of time, and which takes a lot of effort from the teachers also in giving feedback...Yes, it’s not summative, it’s ongoing I mean you don’t assess students once and then make generalisations about students’ ability by just using one time assessment or test...and it’s more reliable because you have a lot of time to cross-check students’ progress throughout the semester...you get to know students better and you can help students better also.

LM2

It also appears that for this “very confident expert” instructor, class size was the underpinning factor that impacted his ability to implement his ideal assessments in his actual assessment practices. He pointed out that despite strongly endorsing the use of portfolios to accurately assess student learning, he decided to exclude it from his courses, given he had a large class size. He stated that it was impractical for him to employ portfolios to assess students’ learning in five different classes comprising 150 students for each semester. As he asserted:

[I] have an average of 30 students in [each of] the classes, then what it means is... [I have] 30 portfolios for a class. If...[I am] teaching five classes, then it multiplies by five, then...[I] have how many [students], it's impossible to do [portfolio assessment].

LM2

With respect to instructor LM3, who was a “proficient, moderately confident assessor,” he assessed his students’ oral abilities via the use of performance-based assessment tasks, such as debating and meeting discussions, as he considered that they reflected the real-life needs of the students. As he stated:

To tell you the truth, I don't think teachers assess students' speaking, well, in CE, Core English classes. There's no session for students, for teachers to test, and that would be the case. So...[I] focus more on the area that...[I] ignore so far...the way that they debate, the way that they appear in the meeting, so if...[I] can create [assessments]...like that, I think it will be more purposive and more meaningful than just writing the answers to [test/exam] questions all the time...it's the only way to show how much the students can do.

LM3

In relation to instructor LN4, who was a novice, yet very confident assessor (see Table 8.4), she argued that her ideal assessments would include a combination of paper-based and oral assessments comprising tests and exams, as well as reflective journals to assess students’ learning. This indicates that her current assessment competence played no role in her endorsement towards valuing both traditional and innovative assessments, given she had very limited assessment knowledge, as demonstrated by the objectively administered and scored Classroom Assessment Knowledge test. Hence, something else must be influencing her to value both types of assessment. As she asserted:

The [ideal] assessment [was] through the written and the oral [tests/exams]...and one more thing...[I] can ask students for their reflections, too...[Hence, I] can know or learn more clearly about the students' [abilities].

LN4

With regard to instructors LM1, LN5 and LN6, who were competent and proficient assessors respectively (see Table 8.4), they asserted that their ideal assessment would cover traditional assessment like tests and exams to assess their students’ learning. Two of these instructors (i.e., LM1 and LN5) also added that their ideal assessments mainly focused on what they had taught their students in class, indicating that they may be narrowing the curriculum to only assessing what was taught, as opposed to what was

specified in the curriculum with regard to their departmental learning goals. As they explicitly stated:

Assessment [test/exam] has to reflect what...[I] have taught...not too much on memory or may be not too much on...details or facts...questions have to be about critical thinking skills...Cambodia lacks...people who are creative, so thinking skills, okay, critical thinking skill is very important for students in Cambodia.

LM1

The [ideal] assessment will reflect, okay, what I have taught to the students. I will think about their ability whether...their ability fits with the tests or not, [their ability matches] the content of the tests, okay, [and] the language use, okay...I need to pilot [the test] for myself whether I can finish this test within the time limit or not...because when...[I] think about these factors, ...[I] can [design the tests that] reflect the real ability of the students.

LN5

[For my ideal assessment], I will adapt the resources to fit the students' levels and the students' backgrounds...[For example] when...[I] test reading, ...[I] test the skills not the knowledge. So if...[I] choose the knowledge that is far beyond the students' backgrounds, they cannot get it. And that can be a mistake for assessments.

LN6

The underpinning factors that appear to induce the six instructors to have either negative or positive endorsements toward the use of traditional assessment in assessing their students' learning tended to be more directly related to their assessment experience as students during their pre-service teacher preparation programme. Two instructors (i.e., LM2 and LM3 who were classified as expert and proficient assessors respectively) devalued traditional assessment (e.g., tests/exams) as they perceived the association between memorisation or rote-learning and the use of tests/exams in assessing their learning. They, therefore, endorsed the use of innovative assessment (e.g., portfolios and performance-based assessment tasks, such as debating and meeting discussions) in assessing their students' learning. As instructor LM2 reported:

[With regard to the] subjects like Literature, Core English, [and] Global Studies...there could be certain contents that...[I] need to remember...I mean there are some materials that...[I] memorise with very little understanding...so the only way to do well in the tests is to remember the answers to particular questions...whether...[I] understand it or not, it may not matter a lot as long as...[I] can give the answers back to the teachers, then...[I] get the scores.

The four remaining instructors (i.e., LN5, LN6, LM1 and LN4) who were identified as proficient, competent and novice assessors respectively, tended to endorse traditional assessment. They reported that their learning was mainly assessed through the use of tests and exams when they were students during their pre-service teacher preparation programmes. This indicated that such assessment experience influenced their beliefs toward endorsing and placing greater value on traditional assessment, such as tests and exams, although two of them (i.e., LN5 and LN6) demonstrated an appropriate level of classroom assessment knowledge (i.e., Level 3: Proficient) that enabled them to implement innovative assessment. As one instructor explicitly stated:

If I take the Core English [subject tests/exams], what I need to memorise is the vocabulary. But if I take the Culture or Literature [subject tests/exams]...I need to memorise the lessons in order to answer the questions because the tests usually not test only vocabulary but [also assess] comprehension that I need to memorise...what I did is [to] make sure that I can remember [the lessons] during the tests, and after the tests, I don't care...so [the lessons were] not staying in [my] mind for long, I guess after one week or two weeks I still remember, but not much.

LN6

It also appears as though the assessment policies of both departments influenced the instructors' perceptions of an ideal assessment. For instance, given departmental assessment policies placed greater value on traditional assessment, in terms of their weighted percentages contributing to the final course grade for summative purposes (see Table 8.1), the types of assessment employed by the instructors, therefore, tended to be confined to traditional assessment, including tests and exams, more than innovative assessment, such as portfolios and performance-based assessments. This indicates that the assessment policies of the two departments may exert an influence on these instructors' beliefs toward endorsing traditional assessment and thus placing weighted percentages on exams and tests in determining final course grades. As one instructor reported:

[I] have classroom [ongoing] assessment [50%] and another 50% is for final exam...[For] ongoing assessment...[I] have two kinds of progress tests, quizzes,...[one] assignment, ...presentations, ...class participation, and homework.

LM3

It further appears as though the high teaching load of instructors negatively impacted their capacity to implement their ideal assessments in their actual assessment

practices. On average, all six instructors reported that they taught over 20 hours per week in order to earn reasonable incomes to support their living expenses. Thus, these instructors appeared to have lacked time to construct the new assessment tasks and/or check the quality of their assessments. Three instructors (i.e., LM2, LM3 and LM1) also commented that they endorsed teaching as many hours as they could, because they got paid based on the hourly rate. Four instructors (i.e., LM3, LM1, LN4 and LN5) further said that they reused their previous year tests with their current students without making any changes, because they had no time to construct new ones. For example, two instructors reported:

The current hours that...[I am] teaching, if you understand that,...[I'm] paid by the hours...[I] teach, so the more...[I] teach, the more...[I] earn...[If the department head] give me more money, I don't have to teach a lot of classes and I just need to probably put a lot more effort in looking at the quality of teaching [or assessments employed].

LM2

It would be 33 hours per week...because...[I] teach many hours, many sessions in the week, and then...[I] will find...[myself] that...[I'm] always only in the middle of teaching and not having enough time to design the tests...and [I] don't have time to design [the new tests], so...[I am] forced to use the same tests for different students.

LM3

Similarly, three instructors (i.e., LN6, LN4 and LN5) stated that their endorsements toward teaching many hours per week was due to the fact that they received an extremely low monthly salary from teaching the scholarship students and received quite low hourly rates from teaching fee-paying students. Instructor LN5 also revealed that the reuse of her previous tests with her current students resulted in the leaking of test information to some of the students before the tests had been administered. For instance, one instructor reported:

[I] teach so many hours a week, so how can...[I] have the time to prepare the appropriate tests for the students' level...[I] copy and paste from the other materials in order to have a test for the students to do...this year, ...[I] use the same [test], for some [students] they know the answers already, so [they] just come and then write down the answers...[I am] the lecturer, ...[I] don't want to teach many hours, but...[I am] forced to do so...If the salary is good, ...[I am] willing, okay, to design the good tests to help students, but the pay rate [of the teaching hour] is very low...[I get paid one hour for] five US dollars, [teaching in the] private program [class]...[and received one month of] 100 US dollars [for

teaching per week of] 12 hours in the [two scholarship program] classes...[the payment for teaching the scholarship students is] extremely low.

LN5

In summary, two instructors preferred their ideal assessments to be relevant to innovative assessment such as performance-based assessments and/or portfolio assessments. In contrast, three instructors favoured their ideal assessments to be associated with traditional assessment like tests/exams. One instructor, however, preferred her ideal assessment to be related to both traditional assessment (i.e., tests/exams) and innovative assessment (i.e., reflective journals). It also appears that two instructors may have been narrowing the curriculum to only assessing what was taught, as opposed to what was stated in their departments' learning goals. Furthermore, the instructors' endorsement towards traditional assessment tends to be associated with the assessment policies of their departments, as well as their prior assessment experience as students. Moreover, four instructors reported that they reused the previous year's test without making any changes with their current students, which can result in the leak of the test information to some of the students prior to test administration. This has implications for validity. Maintaining test security is, therefore, a major challenge for such instructors, as the assessments could be unethical or unfair to those students who do not have access to such information. In addition, it appears that the instructors' background, such as large class sizes and their high teaching load, tend to impact their capacity to implement their ideal assessments in their actual assessment practices. The high teaching load of these instructors seems to be related to the low hourly rate and monthly salary.

8.4.3 Knowledge and Understanding of the Concepts of Validity and Reliability

Given validity and reliability characteristics play a crucial role in providing accuracy and appropriateness of the interpretations and uses of assessment results (Miller et al., 2013; McMillan, 2014; Popham, 2014), having in-depth knowledge and understanding of these two concepts is important for instructors to implement high

quality assessments. To explore the instructors' knowledge and understanding of the concepts of validity and reliability and to examine the quality of their assessment implementation, the instructors were asked: to define the concepts of validity and reliability; to explain the methods used to enhance the validity and reliability of their assessments; how they typically graded their students' work; and how their decision making was made in relation to final course grades plus explaining the ways in which they handled borderline students.

The interviews revealed a limited understanding of the technical implications of the concepts of validity and reliability and these impacted instructors' assessment implementation. With regard to the concept of validity, one instructor (i.e., LM2 who was an expert assessor) had a broader understanding of the concept of "content validity". However, two instructors (i.e., LN6 and LM3 who were both proficient assessors) demonstrated a narrow understanding of the concept of "content validity". Other instructors (i.e., LN5, LM1 and LN4 who were classified as proficient, competent and novice assessors respectively) seemed to show a lack of understanding of this concept. Similarly, three instructors (i.e., LM2, LN5 and LN6) demonstrated a narrow understanding of the concept of reliability, whereas the other three instructors (i.e., LM1, LM3 and LN4) showed a lack of understanding of this concept. Instructor LM2 was able to broadly define the meanings of the concept of "content validity" (i.e., the extent to which the assessment tasks provide a relevant and representative sample of the learning domains to be measured) (refer to section 2.2.1 in Chapter 2 for a detailed discussion). Instructor LM2 however narrowly associated the concept of reliability with mainly the test-retest method (i.e., the same assessment tasks are administered to the same group of students twice with a sufficient interval time between these two periods of administration). This suggests that his further overseas assessment training and his attendance at a number of international conferences on educational assessment focused more on content validity than reliability characteristics, which may have contributed to his broader understanding of the concept of content validity than reliability. As he asserted:

The test itself can be used to test what we claim to test...Yes, if you use it to measure something else, then it does not measure what it claims to measure. And therefore it's not valid...if the test is reliable, if you test a person once, and you test that same person at different time, and that person has not make any progress or change, then the test should produce the same results. Then the test is reliable.

LM2

Although instructors LN6, LM3 and LN5 were placed at the proficient level from the assessment knowledge test (see Table 8.4), they had varying levels of knowledge and understanding associated with the concepts of validity and reliability. For example, instructors LN6 and LM3 associated the concept of validity mainly with a low level of content validity, by means of confining the assessment tasks to the contents of the learning material covered in classes, rather than aligning their assessments with the skills, knowledge and attributes specified in the curriculum. Instructor LN5, however, could not provide any meaningful definition of this concept. Similarly, instructor LN6 associated the concept of reliability solely with the intra-rater method (i.e., the extent to which the consistency in marking the students' responses by the same instructor across different times), whereas instructor LN5 associated the concept of reliability with mainly the coefficient alpha method (i.e., the assessment tasks are administered a single time to a group of students and the coefficient alpha obtained from these assessment tasks provides evidence of internal consistency). In contrast, instructor LM3 could not provide any meaningful definition of the concept of reliability. Specifically, they defined the concepts of validity and reliability as follows:

I compare to what I taught with the test materials whether it really matches or... it's far beyond what I taught them. This is a validity issue...for the writing test, I don't [think] it's reliable...[I] actually have the criteria of grading, but sometimes it's not that much fair for each student...Actually...[I have] clear criteria but sometimes it depends on the idea of the teacher [myself].

LN6

If...[my] test is valid, it tests what it's supposed to test...If my test can distinguish between the one who mastered the materials well and the one who did [not]...In general, which means if you teach grammar, you're supposed to test grammar, you're teaching something, you're supposed to test that thing...And reliability is more on how you design the test, whether there are enough number of testers to evaluate the test, whether the criteria you use is okay or not, or whether it's reliable in terms of condition that students have done in the class, whether students are familiar with the test format or design.

LM3

Reliability, I think when I give this test to the students, okay, this class and then I give another class, and...[I] get the same results from two groups of students, who are in the same level. And for validity, this test I can use this year, I get this result, so I expect next year when I give the same test to the students, and I get the similar results, too.

LN5

In contrast, the other two instructors (i.e., LN4 and LM1), who were novice and competent assessors respectively (see Table 8.4), were not able to provide any meaningful definitions of validity and reliability associated with educational assessment. Instructor LM1 admitted that she was unable to define both of these concepts, given she just memorised them during her pre-service teacher education training and had since completely forgotten their meanings. This indicated that she learnt through rote-learning when she was a student teacher, which may have attributed to her current lack of understanding of assessment-related principles and procedures. For instance, they defined the concepts of validity and reliability as follows:

Reliability means that it is reliable with the scores, for example, with the correction [of the test papers]...And validity whether it is correct or not. For example, when...[I] correct the students, sometimes...[I] make mistakes...when the students give reasons.

LN4

It [reliability] reflects the results, the outcomes that the students have learned...Validity, it's valid. I don't remember the terms. I mean I don't remember these technical words.

LM1

It appears that the pre-service assessment training and lack of access to in-service professional educational workshops contributed to the narrow understanding of the concepts of validity and reliability of these instructors. As instructor LM 3 stated:

When I was training in Teaching Methodology...[I also had a teaching practicum] for one month and [a] half...[I] must learn how to teach and also how to assess. That's on the second semester syllabus...I don't think it prepared me a lot. At that [time]...[I was] discussing on theories and not really practical I mean. What I mean is that...[I] don't really have time to see the tests and then design the actual tests for students...the actual way of designing the tests was not implemented in my class...theory learning seems to be insufficient...Most of the professional workshops [provided in the department] focus more on teaching basically, but not really on assessments...Most of the time we deal with techniques in teaching.

In relation to enhancing validity, owing to their limited knowledge and understanding of the concept of validity, the more tangible notion of content validity was given the most attention in the six instructors' classroom assessment implementation. To enhance a high level of content validity, there is widespread agreement in the literature that assessment tasks must be aligned with the learning goals specified in the curriculum of the two departments and the use of innovative assessment, such as self-assessments, peer assessments, performance-based assessments and portfolio assessments (refer to Chapter 2 for a detailed discussion). However, when assessments are associated with mainly traditional assessment, such as tests and exams, and these assessments are aligned merely with how well the students have mastered the taught content, rather than based on knowledge, skills and attributes stated in the curriculum, such assessments tend to focus on the superficial level of content validity. Of the six instructors, one appeared to enhance a higher content validity than the five remaining instructors. For example, instructor LM2, who was an expert assessor with eight years' teaching experience, reported that he focused on enhancing a higher level of content validity. That is, he ensured that his assessment tasks were similar to what was taught in class and assessed the skills intended. This indicated that he had a broader understanding of the concept of content validity. As he explicitly stated:

So...[I] look for the materials that...[I] think is testing students' ability that...[I] want to test...[I] look for certain reading that is representative of what they have [been] covered in class.

LM2

Four other instructors (i.e., LN6, LM3, LN4 and LM1), however, concentrated mainly on emphasizing the superficial level of content validity by means of having their assessments covering merely what they had taught the students in class, indicating they had a narrower understanding of the concept of content validity. It also suggests that the course books used by these instructors could mediate between their assessment tasks and the content taught. That is, the course books were found to not only exert an influence on the content taught by the instructors in class, but also their assessment tasks used in assessing students' learning. For example, they reported:

I compare to what I taught with the test materials whether it really matches or... it's far beyond what I taught them...I will look at the results...How many percents that the students can achieve after I assess them. So, I will look at what is the mistake, is it the validity [issue?]

LN6

I try to cover most of the important points not any detail of it, but the important one I think it's useful to remember...If my test can distinguish between the one who mastered the materials well...[and] the one who did not master materials, I think that's [a] good test.

LM3

In order to make it [test] more valid, ...[I] just design the test [that matches] ...what...[I] have taught...and...[matches] the students' level.

LN4

Assessment [test/exam]...has to reflect what...[I] have taught, [and] what the lessons have been designed.

LM1

Unlike her counterparts, instructor LN5, who was a proficient assessor with five years' teaching experience, reported that she neglected to enhance validity of her assessments. This indicated that owing to her narrower understanding of the concept of validity, she had not focused on enhancing validity in her assessment practice. Specifically, she asserted:

I don't do anything [to enhance the validity of my tests/exams]...[I] never care about validity.

LN5

It should be noted that owing to the instructors' limited knowledge and understanding of the concept of validity, as well as the tendency to narrow the curriculum to only assessing what was taught (as opposed to what was stated in the curriculum regarding the departments' learning goals), instances of negative and unintended washback effect on teaching and student learning emerged. That is, the teaching content appeared to predominantly influence these instructors' assessment tasks and such teaching content was likely to be based on the course books they were using. As such, the departments' learning goals tended to be neglected by these instructors. Furthermore, while such assessment implementation (i.e., focusing mainly on the superficial level of content validity through using objective assessment methods) tended to have the benefits for fast and easy marking on the part of instructors, at the same time it had negative and

unintended impact on teaching and students' learning by narrowing the curriculum content. This has been referred to as a negative washback effect (see Cheng, 2008; Wall, 2012), as the assessments are less likely to be able to identify where students are in their learning, and to identify appropriate teaching intervention strategies to help the students improve and develop their language skills, knowledge and attributes specified in both departments' learning goals.

Regarding the concept of reliability, these instructors' understanding was also limited. Three out of six instructors (i.e., LM2, LN5 and LN6) associated the concept of reliability with the test-retest method, coefficient alpha method and intra-rater method respectively, while the remaining three instructors (i.e., LM1, LM3 and LN4) were not able to provide any meaningful definition of the concept of reliability. Such limited knowledge and understanding of the concept of reliability by these instructors could also have a negative impact on their assessment implementation.

With respect to enhancing reliability, the instructors defended their assessments on the grounds that they perceived the assessments to be reliable, without necessarily collecting any evidence to support such claims. For instance, instructor LM2, who was a confident expert assessor with eight years' teaching experience, believed that his assessment tasks were reliable, although he admitted he never undertook any statistical analysis to confirm the reliability of his assessment tasks. His narrow knowledge and understanding of the concept of reliability could have influenced his assessment practice. As he asserted:

[I] have never done statistical analysis formally. So...[I] assume mostly...[my] test is reliable even without using the test again and without doing the test [items] analysis.

LM2

Similarly, instructors LN6, LM1 and LN4, who were proficient, competent and novice assessors respectively, checked the reliability of their assessments by merely taking a cursory glance at the scores awarded and/or students' average scores, without undertaking any systematic review. That is, they checked whether the students' results conformed to their expectations of students' prior performance (referred to as the spill

over effect). This suggests that instructors' tacit knowledge (i.e., their own beliefs and values) influenced their marking of the students' work. For example, they reported:

I have never officially okay done that [statistical analysis of my tests]. But I have browsed through, for example, when I mark all the papers, I have browsed through it and I see that okay whether it's reliable or not reliable, ...[I] can see the scores, ...[I] can reflect [on it] okay. So, basically I think I would say that it's acceptable [reliable].

LM1

I will look at the results...How many percents [sic] that the students can achieve after I assess them. So, I will look at what is the mistake, is it reliability [issue?]

LN6

[I] make sure that it is so reliable when...[I mark student work by means of paying my attention to]...the fair correction...[for example, I] think that he [student] should get 85 or 90 [marks]...but he just gets only 65 or 70 [marks]...[I] just check...what is the...reasons [behind].

LN4

In line with her counterparts, instructor LN5 further reported that she neglected to enhance the reliability of her assessment tasks, suggesting that she devalued it. She also had the same attitude toward validity. Specifically, she stated:

I don't do anything [to enhance the reliability of my tests/exams]...[I] never care about...reliability.

LN5

Unlike his counterparts, instructor LM3, who was a proficient assessor, reported that he enhanced the reliability of his assessment tasks, yet his explanation did not provide any evidence to support such claims. As he asserted:

[To enhance] the reliability [of my test]...I try to make my [test] instructions clear [and] I try to make...[the test] items clear enough.

LM3

The instructors' limited knowledge and understanding of the concepts of validity and reliability may have negatively impacted their capacity to implement their self-reported grading practices. With regard to marking students' work (i.e., paragraphs/essays), three instructors (i.e., LN6, LM3 and LM1) appeared to have marked with a lack of accuracy and consistency, as they reported being influenced by their personal values and beliefs in judging students' work rather than using solely the marking criteria. Although these instructors appeared to have employed a criterion-referenced

framework through using analytic scoring methods (i.e., marking criteria including content, vocabulary, grammar and organisation) to assess each student's work, it appeared that they relied heavily on their overall impression of the content over other aspects (i.e., referred to as a holistic judgment). While holistic judgment allows instructors to mark their students' work quickly, as it is based on a cursory glance of the pieces of writing, it can raise questions about the validity (i.e., construct underrepresentation) and reliability (i.e., inconsistencies in marking) in relation to students' work (see Miller et al., 2013). For example, the instructors reported:

For the writing test, I don't [think] it's reliable...[I] actually have the [marking] criteria of grading [student work], but sometimes it's not that much fair for each student...[despite I have] clear [marking] criteria...sometimes it depends on the idea of the teacher [my beliefs concerning the quality of the content of student work].

LN6

When...[I] mark the tests [comprising an essay]...[I] just make sure...[I] make [my] marking fast [glancing through the content of the student work]...[so I] don't look [or read all aspects] closely...[regarding] what the students have written.

LM3

[When I] teach more, ...[I] become...very tired...[so I] don't have time for marking, so...[I] always try to find way...to do the marking easier [by means of glancing through the content of the student responses].

LM1

The findings suggest that due to limited knowledge and understanding of the concepts of validity and reliability, all six instructors were being influenced by their personal values and beliefs in terms of placing greater values and endorsements on the processes of learning, such as student class attendance and student efforts, as the main justifications of their grading practices. These types of grading practices can raise questions about validity (i.e., construct-irrelevant variance) and reliability of the assessment results in showing the true learning achievements of students in the courses. For example, instructor LM1, who was found to be a competent assessor (see Table 8.4), highly endorsed student class participation in her grading practice. She believed that class participation was one of the vital aspects that needed to be included in her course grades. As she asserted:

I strongly agree that we have to count it [class participation] as one of the necessary assessments, yes not only the results of the tests and the results of the exams.

LM1

Four other instructors (i.e., LM2, LN4, LN5 and LN6) further showed their endorsements toward student class attendance in their grading practices. As they stated:

I think there could be [a] strong correlation between class attendance and students' performance. So, I believe...[when] students are coming to class more often, it's more likely to help them learn better also.

LM2

The attendance, the class participation and homework...can be mixed...because the teachers can know the students better...when they have the low attendance, it means that they are...[often] absent, and not active in the class...So...[I] just mix this one [the combination of class participation, attendance and homework] into 10%.

LN4

[Class] attendance covers class participation, whether the students come or not, that's class participation of attendance. [For] some students, they just come and take attendance without doing anything.

LN5

I believe that 10% is not too much...I give them [students] credit, I encourage them to come to the class.

LN6

To make decisions as to whether to pass or fail borderline students, all instructors stated they took into consideration students' non-academic achievement factors. This has strong implications for validity, and in particular construct validity as to what they are really assessing. Furthermore, when awarding extra scores, all instructors reported that they largely confined scores to borderline students, and they defended their action on the grounds that this particular group of students needed extra scores to pass their courses. Such grading practices can raise questions about the validity (i.e., construct-irrelevant variance regarding including students' non-achievement factors in course grades and consequential validity concerning ethicalness/fairness for students who already got the passing grades) (refer to section 2.2.1 in Chapter 2 for a detailed discussion) and reliability of the assessment results in reflecting students' actual learning outcomes. It highlights that Cambodian culture plays a partial role in these instructors' grading practices, as it tends to value students' efforts and participation in addition to their actual

abilities. For example, one instructor (i.e., LN5) reported that she gave additional tasks to her borderline students to do in order to help them obtain extra scores to pass the course, although she admitted her action was unfair on other students in the class. As she asserted:

If I notice...[the] students just fail [by] 2 or 3 points [or marks], I find the other way to help them like asking them to do extra work in order to get the supplement scores...It's not fair [to add extra marks to only the borderline students], but...[I] cannot find...[a] better way to help those students...they already passed.

LN5

Three other instructors (i.e., LM2, LM3 and LM1) stated that they considered borderline students' active involvements in class activities and expend efforts in their work when adding scores to pass them. Specifically, they stated:

If any students scored below 50...based on the rule, it's a fail...[but] I reward students, I reward hard work in addition to real students' ability [in] doing things... out of a 100, ...[I] give a 5% [and]...this 5% would be added depending on how much effort...[I] feel the students are putting into their studies throughout the semester.

LM2

Well, I consider the other kinds of performance whether the student is working hard...involving in class [activities such as] whether they are doing their homework, [and] whether they participate activities at the classroom discussions...So I'll consider those factors and then come up with the issue: fail them or pass them...I just add to the one who needs to pass...there is no policy or rule that say if...[I] add scores [to a borderline student], and then...[I] must add to everybody [in the class].

LM3

If...[any students] got 48 % which [is]...based on the rule of the school, it means they fail. But to me I look at students, okay. If...[those] students...[are] very active, okay, and...they're capable enough as well...Unfortunately they don't do well in the exam, so, they lose the marks so I push...[these students]...I give extra scores [to them]...I think there is nothing in the world that is fair all the time.

LM1

The two remaining instructors (i.e., LN6 and LN4), however, commented that they considered students' subject majors and students' attitudes and behaviours respectively, when making their decisions as to whether to pass or fail borderline students. One instructor (i.e., LN6) believed that students who majored in Social Work or Psychology would have had more opportunities to learn English from their own majored departments in subsequent years. Unlike her counterpart, instructor LN4, who was a

novice assessor, stated that she took into consideration good attitudes and behaviours in class when deciding whether to pass or fail borderline students. For instance, they reported:

If my students [are] from the Social Work, [or] Psychology [major]...I will let them pass because I believe that they must have a lot English training...[later with] their [majored] subjects...[in addition to their studies with] English department, so I guess they can catch [up with] the others for the next [coming] year...it is not fair [to add the marks to only the borderline students], but...[I] have no choice because...[I] just would like them to pass.

LN6

[If] they [students] have tried their best already, and they have good attitude, good manner, in the class, just 1 mark or 2 marks, ...[I] just add [marks to their course grade to pass them]...I think it would be okay [it's fair] because they have passed already. Why don't they get so jealous with only 1 mark because...[I have] just learned that this student is good and [1 mark] should be added [to his course grade].

LN4

In relation to grading decision making, all six instructors stated they employed score totals (also referred to as percentage grading) for summing up students' learning outcomes at the end of the semester with reference to the cut-off score for passing (i.e., 50%) for summative purposes. That is, all six instructors stated they used average scores through aggregating the raw scores obtained from all types of assessment tasks employed from the beginning to the end of the course. In other words, if the students achieved 50 marks or above (i.e., the average scores obtained from all types of assessment tasks), they would be promoted to the next level accordingly. Such a grading decision model has the potential to provide the impression of precision and it is easy to operationalise. The grade cut-off score in this decision-making model, however, was not generally related to the mastery of specific skills or learning outcomes, and it was typically left to the instructors to work out the cut-off score for each assessment task. As such, this grading decision model can raise concerns associated with the validity and reliability of students' overall learning achievements. It raises questions about how the marks were generated in the first place in terms of sampling adequacy, the difficulty of assessment tasks, quality of the assessment tasks, marking standards and consistency, as well as the incorporation of students' non-academic achievement factors (i.e., effort and class attendance) into the overall learning achievements in the course. Furthermore, all six instructors viewed that

their final course grade was primarily used to pass or fail their students from one level to the next level. This indicates that they conducted their classroom assessment primarily for summative purposes, as opposed to formative functions. As one instructor stated:

The 100 scores consists of...ongoing assessment...[which is worth of] 50% [and] exam is 50%. Generally, in ongoing assessment, ...[I] have test 1 and...test 2...presentations...class participation...[and] homework...[course grade is used] to determine whether the students can pass...from one semester to another semester from year 1 to year 2 or not.

LM1

The instructors' endorsement towards such grading practices also appeared to be associated with departmental assessment policies. That is, these policies (see Table 8.1) appeared to exert an influence on the instructors' personal values and beliefs in relation to their summative grading practices, given the fact that these assessment policies imposed the cut-off score for passes of 50% and placed greater value on student class attendance (i.e., compulsory class attendance) and student class participation (i.e., one component of the ongoing assessment). For example, one instructor reported:

I have [50% for ongoing] assessment and another 50% is for final exam...I must follow what the school [department's assessment policies stated]...regarding criteria for assessments [and/or my grading practice].

LM3

In summary, one out of six instructors demonstrated a broader knowledge and understanding of the concept of content validity, although he showed a narrow understanding of the concept of reliability. This instructor was found to be an expert assessor with a high level of confidence. He reported that he enhanced a higher level of content validity on the basis of his broader understanding of this concept, although he reported his neglect in examining the reliability of his assessments. The other five instructors demonstrated a narrower knowledge and understanding of the concepts of content validity and reliability. These five instructors' backgrounds, such as pre-service assessment training and in-service educational assessment training, could have contributed to their narrow understanding of these concepts. Because of their insufficient comprehension of such concepts, they stated that they had neglected to examine the reliability of their assessments and limited their checks of content validity by simply

taking what was taught in class to set the items in the tests (which tended to be drawn from the course books used in class). These types of assessment implementation (i.e., focusing largely on the superficial level of content validity) tend to have an association with the use of traditional assessment methods, such as true/false and multiple-choice questions, and therefore have the benefits for fast and easy marking on the part of instructors. Such assessment implementation, however, is likely to have a negative impact on teaching and students' learning as they narrowed the curriculum content to what was taught in class, rather than based on the skills, knowledge and attributes specified in the curriculum. Furthermore, instructors' limited knowledge and understanding of the concepts of validity and reliability, their personal values and beliefs and departmental assessment policies appeared to be related to their poor grading practices.

8.5 Summary

The three main themes associated with the six instructors' classroom assessment literacy comprised their perceived assessment competence, their notions of ideal assessments and their knowledge and understanding of the concepts of validity and reliability. As Figure 8.1 below illustrates, both departmental assessment policies and background-related factors appear to influence these instructors' classroom assessment literacy in implementing their assessments. The departmental assessment policies include the purpose of assessment (i.e., merely summative functions), preferred assessment methods (i.e., mainly emphasis traditional assessment like tests/exams), assessment weightings (i.e., heavy percentage weightings given to tests and exams), student class attendance (i.e., compulsory), student class participation (i.e., one component of the ongoing assessment) and the cut-off score for a pass (i.e., 50%). The instructors' backgrounds comprised: their pre-service assessment training; overseas assessment training and attendance at a number of international conferences focusing on educational assessment; access to in-service professional workshops on educational assessment; prior assessment experience as students; hourly payment rate and monthly salary; number of teaching hours per week; and class size. The next chapter discusses the results from both quantitative and qualitative phases of the study.

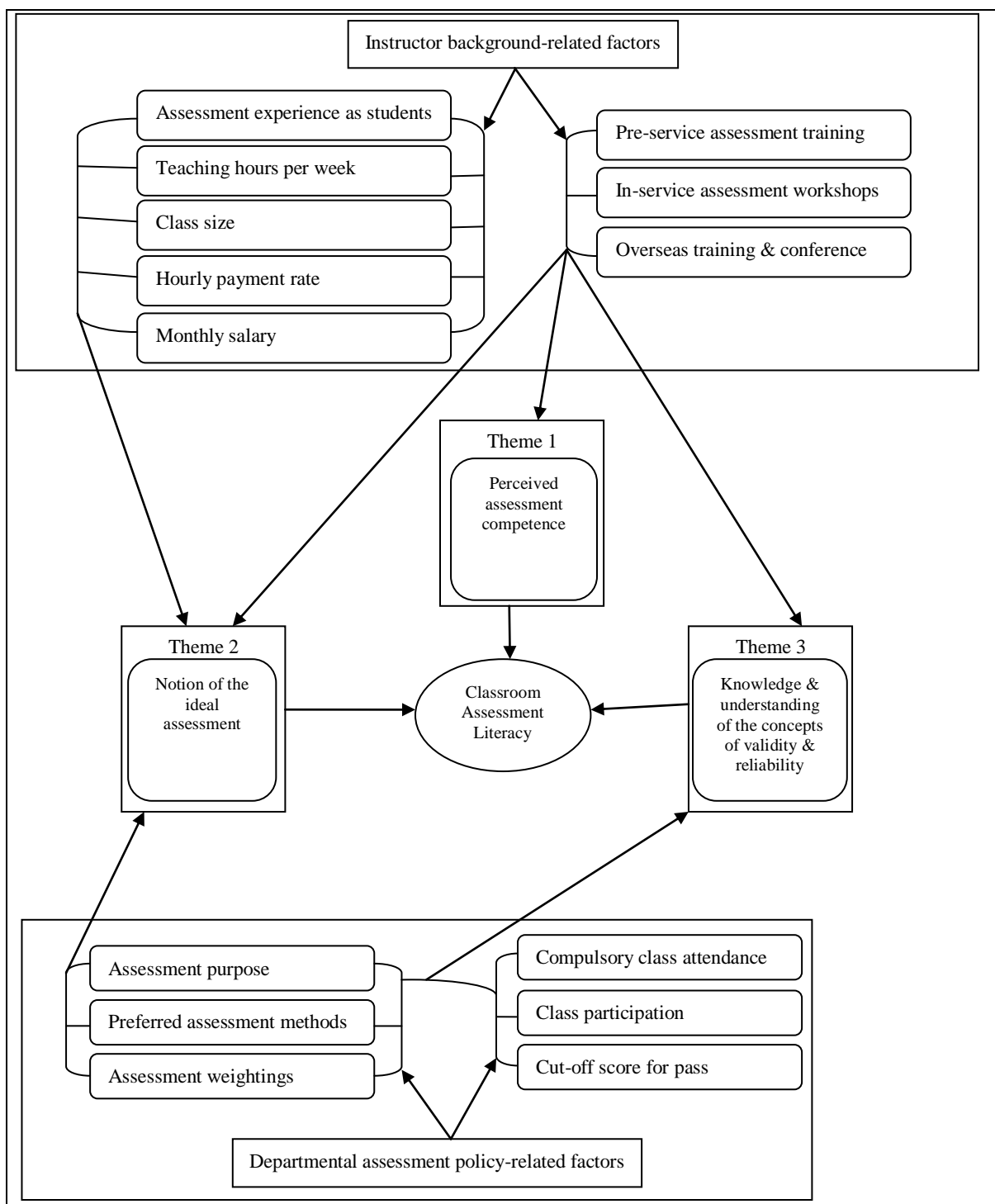


Figure 8.1 The relationship between instructors' classroom assessment literacy, their backgrounds and departmental assessment policies

Chapter 9: Discussion and Conclusion

This chapter comprises three sections. Section 9.1 presents a review of the rationale for the study and the methodology. Section 9.2 discusses the quantitative results and addresses the main research question pertaining to measurement related issues. This section also integrates the results from both quantitative and qualitative phases to address the three subsidiary research questions, which are concerned with practice and development related issues. Section 9.3 discusses the implications of such findings for theory, policy and practice, and the design of pre-service teacher education programmes. Finally, the study's limitations and future directions for research in the area of classroom assessment literacy are discussed (see sections 9.3.2 and 9.3.3).

9.1 Overview of the Study

9.1.1 Review of Rationale of the Study

The main purpose of this study was to develop and validate a set of scales to measure the classroom assessment literacy development of EFL instructors. There were three main rationales for examining classroom assessment literacy of the instructors within EFL programmes in a Cambodian higher education setting. The first rationale was associated with the need to examine the factors that underpinned classroom assessment literacy. The second rationale related to the need to better understand the level of instructors' classroom assessment literacy and its associated impact on their actual assessment implementation. The third rationale was linked to the need to have an improved understanding of the influence of instructors' background characteristics on their classroom assessment literacy development. Each will be considered in turn next.

At the time of this study, the literature had widely documented that assessment knowledge was crucial for instructors to possess in order for them to implement high quality assessments, to enhance their instruction and foster students' learning. Equally, the literature had repeatedly highlighted that instructors' personal beliefs about assessment played an important role in the ways in which they implemented their

assessments. As such, it could be concluded that both instructors' classroom assessment knowledge and personal beliefs about assessment underpinned classroom assessment literacy. There is, however, a lack of empirical research that has examined the theoretical foundations of classroom assessment literacy from this perspective.

The second rationale for this study was linked to the need to have an improved understanding of the instructors' classroom assessment literacy level and the impact it may have had on their assessment practices. Within EFL programmes in Cambodian tertiary education contexts, assessments have typically been undertaken by instructors. The nature of this classroom-based assessment primarily serves summative purposes (i.e., to pass/fail students and/or to certify the degree to students) (Tao, 2012). Given the high stakes, assessment outcomes obtained must be valid, reliable, fair and transparent to ensure comparability of standards in conducting assessments across and between classes and universities. The design and implementation of high quality assessments require high levels of instructors' classroom assessment literacy. However, there is a current lack of empirical studies that have examined classroom assessment literacy levels of Cambodian EFL university instructors.

Thirdly, given that instructors were the main agents in implementing assessments in their universities, it was vital to have a better understanding of the influence of their background characteristics on their classroom assessment literacy development.

Unfortunately, there has also been a lack of empirical research that has examined the factors influencing classroom assessment literacy development of tertiary instructors within EFL programmes, despite the increasing recognition of the crucial role of such programmes in Cambodian society. To date, acquiring a sufficient level of English language proficiency has been seen as vital for Cambodians, to enable them to fully participate and actively engage within both everyday activities, such as work and study in their society, as well as within the ASEAN community. The lack of research in this area is a concern, because the body of studies has provided adequate evidence on the direct relationships between the quality of classroom assessments used and the quality of instruction and students' learning (Black & Wiliam, 1998a; Shute, 2008; Stiggins, 2008; Wiliam, 2011).

The present study examined classroom assessment knowledge and personal beliefs about assessment of the instructors. Classroom assessment knowledge comprised nine standards, related to key stages of the assessment process (refer to Chapter 2). Personal beliefs about assessment of the instructors included their perceptions of the usefulness of innovative assessment methods (e.g., performance-based assessments), factors influencing the way in which they marked students' work (e.g., attitude and behaviour) and the importance of quality procedures employed (e.g., ensuring assessments are valid and reliable). This study also investigated the instructors' classroom assessment literacy levels and its associated impact on their actual assessment practices. Furthermore, it examined the influence of instructors' background characteristics (e.g., class size) on their classroom assessment literacy and implementation.

This study also tested the hypothesis that both classroom assessment knowledge and personal assessment beliefs underpinned classroom assessment literacy. Previous research had not accounted for the interplay of instructors' classroom assessment knowledge and their assessment beliefs. As such, the magnitude of such interaction was unknown. There were no other studies that had examined the interaction amongst classroom assessment knowledge and assessment beliefs of instructors. Therefore it was essential to first develop and calibrate a set of scales to measure the classroom assessment literacy development of instructors, and second to test the hypothesis that classroom assessment knowledge and assessment beliefs together tap into the construct of classroom assessment literacy.

The main research question to be explored in this study was: *“To what extent did assessment related knowledge and beliefs underpin classroom assessment literacy and to what extent could each of these constructs be measured”*? Subsidiary research questions comprised:

1. To what extent was classroom assessment literacy developmental?
2. What impact did classroom assessment literacy have on assessment practices?

3. How did the background characteristics of instructors (i.e., age, gender, academic qualification, teaching experience, teaching hours, class size, assessment training, and departmental status) influence their classroom assessment literacy?

9.1.2 Review of Methodology

9.1.2.1 Quantitative Phase

The primary aims of the quantitative phase were to develop and validate a set of scales to measure the classroom assessment literacy development of instructors. It also aimed to test a hypothesised one-factor congeneric measurement model of instructors' classroom assessment literacy. This phase further aimed to examine the influence of instructors' background characteristics (e.g., pre-service assessment training) on their classroom assessment literacy development.

Data Collection Procedures

Purposive sampling procedures were employed to select the largest university that offered EFL programmes in Cambodia. Within this university, the EFL programmes comprised English-major and English non-major departments.

Both the classroom assessment knowledge test and questionnaire were administered to 123 instructors. Seventy of them were currently teaching in an English-major department, while the remaining 53 instructors were teaching in an English non-major department. One hundred and eight instructors completed the classroom assessment knowledge test and questionnaire, yielding a return rate of 88%. The sample consisted of 59 English-major and 49 English non-major instructors comprising 80 males and 28 females with varied background characteristics. Details of the sample were described in Chapter 7.

The respondents initially completed the "Classroom Assessment Knowledge" multiple-choice test. They then completed the beliefs questionnaire by rating each of 22

items on a 4-point rating scale. They were also asked to provide their demographic information (e.g., age).

Scale Development Procedures

A set of scales were developed for the present study to measure the instructors' classroom assessment literacy including Classroom Assessment Knowledge, Innovative Methods, Grading Bias and Quality Procedure. The Classroom Assessment Knowledge scale was a multiple-choice test and designed to measure the instructors' classroom assessment knowledge. The three remaining scales were constructed based on self-reported measures to examine instructors' personal beliefs about assessment. All scales were constructed drawing on findings in classroom assessment literature and using psychometric methodology for scale development. In particular, the Classroom Assessment Knowledge scale was constructed by means of expanding the seven standards for Teacher Competence in Educational Assessment of Students (AFT, NCME, & NEA, 1990) to nine standards. These standards were expanded for this study to take into consideration criticisms associated with the narrow aspects of the original standards, particularly with respect to classroom-based assessment activities required by instructors (Schafer, 1991; Stiggins, 1995, 1999; Arter, 1999; Brookhart, 2011a). These nine standards were considered to cover all key stages of the assessment process. Details of each key stage of the assessment process have been provided in Chapter 2. The nine standards of the Classroom Assessment Knowledge scale comprised:

1. Choosing Appropriate Assessment Methods;
2. Developing Assessment Methods;
3. Administering, Scoring, and Interpreting Assessment Results;
4. Developing Valid Grading Procedures;
5. Using Assessment Results for Decision Making;
6. Recognising Unethical Assessment Practices;
7. Keeping Accurate Records of Assessment Information;
8. Ensuring Quality Management of Assessment Practices; and
9. Communicating Assessment Results.

Each scale was calibrated using the Rasch Simple Logistic Model (Rasch, 1960) for dichotomous items and/or the Partial Credit Model (Masters, 1982) for polytomous items by employing ConQuest software version 2.0 (Wu et al., 2007). Item response modelling procedures generated both item difficulty/parameter and person/respondent ability/perception on the same measurement scale. That is, the same measurement scale was used to refer to item difficulty/parameter and respondent ability/perception. Details of the scale development processes have been presented in Chapter 6.

Data Analysis Procedures

Various data analysis techniques were employed in this study. To analyse the relationships amongst the instructors' classroom assessment knowledge and their assessment related personal beliefs, Pearson product-moment correlations were undertaken. An examination of the influences of instructors' background characteristics on their classroom assessment literacy was carried out using various data analysis techniques including Pearson product-moment correlations, t-tests, one-way Anova tests and cross-tabulations. These analyses used IBM SPSS Statistics version 20. To further examine the interrelationships amongst the four constructs that underpinned classroom assessment literacy, a confirmatory factor analysis was conducted using IBM SPSS Amos version 20. The assessment of this measurement model was determined by the statistical indices of fit and the theoretical basis of the model.

9.1.2.2 Qualitative Phase

The main aims of the qualitative phase were to gain an in-depth understanding of the instructors' classroom assessment literacy level and the impact it may have had on their actual assessment practices. It aimed to further explore the influence of instructors' background characteristics (e.g., prior assessment experience as students) and departmental assessment-related policies (e.g., assessment purpose) on their classroom assessment literacy and practice.

Data Collection Procedures

Data collection procedures consisted of three stages. First, departmental documents regarding learning goals and assessment-related policies were collected. Second, stratified random sampling procedures were employed to select three English-major instructors and three English non-major instructors from the 33 instructors who volunteered to participate in the second phase of this study. Details of the selection criteria have been presented in Chapter 5, and details of the qualitative phase sample have been provided in Chapter 8.

Third, a semi-structured one-on-one interview with each of the six selected participants was undertaken. The informal conversational styles of interviews were employed to allow in-depth interactions between the interviewer and interviewee in a relaxed atmosphere using a series of open-ended questions.

Data Analysis Procedures

Content analysis was used to analyse departmental documents (i.e., the learning goals and assessment-related policies) while a thematic analysis was employed to analyse the interview transcripts. The steps taken to analyse the documents and transcripts included: (1) preliminary reading through the documents and transcripts and written memos to better understand the data; (2) using inductive coding (i.e., the process that permits content/themes to emerge directly from the data) and deductive coding (i.e., the process in which predetermined codes are derived from the theoretical framework used to generate content/themes from the data); (3) verifying the codes by means of using inter-coder agreement check; (4) developing content/themes based on the codes; and (5) generating descriptions and content/themes for each of the documents and participants and comparing within and cross-departments content and thematic analyses.

9.2 Discussion

9.2.1 Main Research Question: To what extent did assessment related knowledge and beliefs underpin classroom assessment literacy and to what extent could each of these constructs be measured?

The Rasch analyses demonstrated that all four scales comprising Classroom Assessment Knowledge, Innovative Methods, Grading Bias and Quality Procedure had satisfactory measurement properties.

In relation to the assessment belief constructs (i.e., Innovative Methods, Grading Bias and Quality Procedure), the correlations analyses showed there was a positive small relationship between instructors' beliefs of the usefulness of implementing innovative assessment methods (e.g., performance-based assessments) and the importance of implementing quality assurance procedures (e.g., ensuring that the assessments are valid and reliable). Furthermore, the analyses revealed that the more instructors were influenced by students' personal characteristics (e.g., attitude and behaviour) when marking students' work, the less likely they were to perceive the importance of implementing quality assurance procedures and implementing innovative assessment methods. Such findings were consistent with what would be expected.

There were also small and moderate relationships amongst the four constructs thought to underpin classroom assessment literacy (i.e., Classroom Assessment Knowledge, Grading Bias, Innovative Methods and Quality Procedure) as indicated by the correlations analyses. For example, instructors who had higher classroom assessment knowledge level were found to be:

- more likely to believe that innovative assessment methods were useful;
- more likely to perceive the importance of implementing quality assurance procedures and checks in their assessment practices; and
- less likely to be influenced by the students' personal characteristics when marking their work.

Such findings suggest that both classroom assessment knowledge and personal assessment belief factors could be regarded as tapping into the same construct.

The confirmatory factor analysis also demonstrated that Classroom Assessment Knowledge, Innovative Methods, Grading Bias and Quality Procedure variables served well as a measure of the single latent Classroom Assessment Literacy construct. These findings supported the hypothesis of the current study. That is, instructors' classroom assessment knowledge base and their personal beliefs about assessment were hypothesised as important facets to reflect their classroom assessment literacy. These findings also substantiated Fishbein and Ajzen's (1975, 2010) reasoned action theory, Ajzen's (1991, 2005) planned behaviours theory and Bandura's (1989, 1997) social cognitive theory that postulate that individuals' knowledge/skills and beliefs/attitudes underpin their behaviours and/or task performances (see section 9.3.1.1).

9.2.2 Subsidiary Research Question 1: To what extent was classroom assessment literacy developmental?

The findings in both phases of this study demonstrated that instructors' classroom assessment literacy was limited. In relation to an examination of instructors' classroom assessment knowledge, it was found that the instructors had limited classroom assessment knowledge for conducting high quality assessments. These findings confirmed a wealth of research that had rapidly reported that school teachers had insufficient assessment knowledge for implementing high quality assessments for over five decades (Mayo, 1967; Plake, 1993; Quilter & Gallini, 2000; Mertler, 2005; Schaff, 2006; Chapman 2008; King, 2010; Alkharusi et al., 2012; Davidheiser, 2013; Gotch & French, 2013).

With regard to an examination of instructors' assessment beliefs, the findings in both phases of this study highlighted that the majority of instructors demonstrated signs of bias in their grading practices. The findings further showed that the instructors mainly employed traditional assessment methods, such as tests/exams, and neglected to examine the extent to which these assessment tasks were valid and reliable, despite showing their endorsements toward the use of innovative assessment methods (e.g., performance-based

assessments) and quality assessment procedures in their assessment practices (see section 9.2.3 for a more detailed discussion).

The findings from the current study also highlighted the presence of social desirability response bias (i.e., an issue of overreporting of socially desirable behaviours and underreporting of socially undesirable behaviours) occurring in the instructors' self-reported measures of their classroom assessment competence. The quantitative phase revealed that those instructors, who perceived they were "*prepared*" in conducting classroom assessments, had a significantly higher classroom assessment knowledge level than those who perceived they were "*very prepared*". This was further highlighted in the interviews, in which one instructor rated her classroom assessment competence level as much higher than her actual classroom assessment knowledge level, as indicated by the test instrument. Such findings were in line with those of Chapman (2008), Alkharusi et al. (2011) and Alkharusi et al. (2012) who reported that instructors tended to overrate their assessment expertise in comparison to their actual assessment knowledge level, as measured by an objective test. The sociologist Erving Goffman (1959) and psychologists Barry Schlenker and Michael Weigold (1989) and Barry Schlenker (2012) had explicitly provided their theoretical views regarding the occurrence of social desirability response bias in any self-reported measures. They explained that individuals tended to influence the way in which they were perceived by others in order to pursue the goals of social interaction. Being perceived favourably by others promotes the individuals' views that they may experience an increase in rewards and a reduction in punishments. Such a perception therefore may have motivated instructors to present their self-images to appear much greater than they actually were, possibly due to the potential perceived threat of disclosure impacting on career opportunities and performance appraisals. As such, low assessment-literate instructors may have overrated their competence in classroom assessment in order to make themselves look good in their professionalism and to avoid any perceived negative consequences in the future. Alternatively, ignorance may have come into play in these low assessment-literate instructors' self-rating of their current assessment expertise.

9.2.3 Subsidiary Research Question 2: What impact did classroom assessment literacy have on assessment practices?

The findings in both phases of this study showed that the instructors' limited classroom assessment literacy negatively impacted their assessment practices. Firstly, the majority of instructors demonstrated signs of bias in their grading practices, although they were aware that their course grades were solely used for summative purposes. They reported that they were not only influenced by students' personal characteristics (e.g., attitude and behaviour) when marking their work, but also incorporated students' non-academic achievement factors (e.g., class participation and attendance) into their final course grades, as well as adding extra marks to borderline students' results in order to pass them. Such findings were not surprising, given these instructors had limited classroom assessment knowledge and were therefore influenced by their implicit personal beliefs. These findings reinforced previous research (Harlen, 2005b; Read et al., 2005; Dennis, 2007) and the assertion amongst educational assessment specialists (Price, 2005; Sadler, 2005; Popham, 2014) that when marking students' work, instructors were often influenced by their implicit personal beliefs and values. These findings also supported considerable research that had repeatedly reported that instructors incorporated both students' academic achievement factors and their non-academic achievements factors into their final course grades (Brookhart, 1993; Cross & Frary, 1999; McMillan & Nash, 2000; McMillan, 2001; Greenstein, 2004; Sun & Cheng, 2013).

These findings also suggested that Cambodian culture played a role in these instructors' assessment practices. In Cambodian society, students' non-academic achievement factors, such as effort and class participation, were perceived as important factors contributing to their learning achievements. Such findings were also consistent with the proposition by Wyatt-Smith and Klenowski (2013), that social and cultural practice was inherent in the instructors' assessment practices. These findings, however, contradicted the recommendations of educational assessment experts, who strongly advised instructors to avoid bias in their grading practices, as this type of bias could inflate the students' actual academic achievements, resulting in inaccurate assessment outcomes (Wormeli, 2006; O'Connor, 2009; Brookhart, 2011b; Chappuis et al., 2012;

Miller et al., 2013; Popham, 2014; Schimmer, 2014). That is, the bias in instructors' grading practice was considered as unethical or unfair to students, especially when assessment results were associated with having consequences on students' academic lives (Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Douglas, 2010; Brown, 2012; Miller et al., 2013). In particular, when assessment results were high stakes (i.e., to pass/fail students), it was recommended that any bias in instructors' grading practice must be guarded against (Lamprianou & Athanasou, 2009; Douglas, 2010; Miller et al., 2013; Popham, 2014).

Secondly, the findings demonstrated there were contradicting results found in the comparison of the findings of the quantitative and qualitative phases, with respect to instructors' endorsements toward the use of innovative assessment methods and quality assessment procedures in their assessment practices. In relation to innovative assessment methods, although the quantitative results indicated that the majority of instructors highly endorsed such methods (e.g., performance-based assessments), all six instructors interviewed reported employing largely traditional assessment, such as tests and exams, in their actual assessment practices. Such discrepancies may be due to the instructors' limited assessment knowledge, leading to a lack of confidence and skills to design and implement such methods to assess their students' learning. These findings were consistent with previous research that reported the tendency for teachers to employ traditional assessment methods in assessing students' learning within school settings (Oescher & Kirby, 1990; Bol & Strage, 1996; Greenstein, 2004; Tsagari, 2008) and higher education institutions (Cheng et al., 2004; Rogers et al., 2007).

With regard to quality assessment procedures, despite the quantitative results showing that instructors had favourable endorsements toward enhancing validity and reliability of their assessments, the qualitative results demonstrated that all six instructors neglected to examine the extent to which their assessments were reliable, particularly when there was a heavy emphasis on traditional assessment methods (e.g., multiple-choice tests). In relation to validity, most of those interviewed (i.e. five of the six) indicated that they merely examined the superficial level of content validity in their actual assessment practices. These findings were consistent with previous research (Mertler, 2000; Harlen, 2005b; Black et al., 2010) that showed that school teachers

tended to ignore the validity and reliability characteristics of assessment in their implementation, despite widespread acceptance of their importance. The inconsistency between what instructors regarded as important, and what they actually did in this study, may have been due to the fact that the instructors had limited knowledge and understanding of the concepts of validity and reliability, and their implications for assessment design and implementation. These findings supported the perspectives of Fishbein and Ajzen's (1975, 2010) reasoned action and Ajzen's (1991, 2005) planned behaviours, that postulated that when an individual lacked knowledge and/or skills to implement the tasks, there was the likelihood that s/he had little or no intentions of carrying out these tasks, despite s/he having favourable attitudes toward such tasks.

It should be noted that by limiting checks of validity to content mapping exercises, there was a tendency for the course books used by instructors to predominantly influence content taught and assessment tasks used, resulting in a narrowing of the curriculum to what was easily taught and assessed. The heavy reliance on course books for teaching and assessment may have been due to the fact that the instructors had limited classroom assessment literacy, as well as being non-native English speakers. And therefore they may have lacked confidence to develop valid and reliable EFL teaching/assessment materials. The narrowing of the curriculum to what was easily taught and assessed, particularly amongst instructors with low levels of classroom assessment literacy, was consistent with previous research and often referred to as a negative washback effect (Popham, 1991; Nolen et al., 1992; Bailey, 1996; Shohamy et al., 1996; Cheng, 2005; Luxia, 2007; Amengual-Pizarro, 2009; Tsagari, 2009).

9.2.4 Subsidiary Research Question 3: How did the background characteristics of instructors (i.e., age, gender, academic qualification, teaching experience, teaching hours, class size, assessment training, and departmental status) influence their classroom assessment literacy?

9.2.4.1 The Influence of Pre-service Assessment Training

Consistent with previous studies undertaken within school settings, the findings in both phases of this study demonstrated a relationship between instructors' pre-service assessment training and their current classroom assessment knowledge (Wise et al., 1991; Mertler, 1999; Tsagari, 2008; King, 2010). Interestingly, the quantitative results indicated that the duration of pre-service assessment training had no influence on the levels of instructors' classroom assessment knowledge. The qualitative results provided an explanation regarding a lack of influence of the duration of pre-service assessment training on instructors' classroom assessment knowledge. In their interviews, all six instructors explained that their assessment training was not conducted as a stand-alone course in their pre-service teacher education programmes. The assessment units they took were actually embedded within a Teaching Methodology course, providing them with less than one semester, despite the fact that this course actually ran for two semesters. Thus, those instructors who reported they had pre-service assessment training for more than one semester in the questionnaire might have misunderstood that the duration of the Teaching Methodology course was the duration of their pre-service assessment training. Such a misunderstanding may have been due to assessment being embedded within this course as opposed to a stand-alone unit. These findings were consistent with Schafer and Lissitz's (1987) study. They reported that educational assessment was typically embedded within another teacher education course rather than being offered as a separate course, resulting in school teachers' limited assessment knowledge. This phenomenon is not unique to pre-service teacher training in Cambodia, although there has been a shift worldwide to recognise the importance of assessment training in such programmes (see Popham, 2011; Griffin et al., 2012).

9.2.4.2 The Influence of Class Size

Larger class sizes were identified in both phases of this study as a factor that could impact on instructors' classroom assessment literacy in implementing their assessments. The quantitative phase showed that larger class sizes negatively impacted instructors' implementation of quality assessment procedures. It was also found that instructors with larger class sizes were more likely to be influenced by students' personal characteristics (e.g., attitude) when marking students' work, than instructors with smaller class sizes. The qualitative phase further revealed that larger class sizes were the main cause that prevented one instructor from employing innovative assessment methods (e.g., portfolio assessments) in his actual assessment practice, despite him demonstrating high classroom assessment knowledge level and positive endorsement towards implementing such assessment methods. Such findings were consistent with previous research. For instance, as was found in the current study, studies reported that larger class sizes influenced instructors' assessment implementation in terms of heavier reliance on traditional assessment methods (see Gibbs & Lucas, 1997), employment of fewer assessment tasks (see Nakabugo et al., 2007), as well as the tendency to incorporate students' non-academic achievement factors into their course grades such as class attendance and effort (see Sun & Cheng, 2013).

9.2.4.3 The Influence of Teaching Hours

The findings in both phases of this study highlighted that the number of teaching hours per week impacted instructors' classroom assessment literacy and practice. The quantitative results revealed that the instructors were more likely to perceive the usefulness of implementing innovative assessment methods (e.g., performance-based assessments) in assessing students' learning when they had less teaching hours per week (i.e., 5-12 hours). This was also highlighted in the interviews, in which all six instructors reported that the high teaching load (i.e., over 20 hours per week) caused them to solely employ traditional assessment, such as tests/exams, to assess their students' learning. Such findings were consistent with those of Haing (2012). For instance, Haing (2012)

found that the number of teaching hours per week (i.e., 30 hours) negatively impacted the way in which the Cambodian EFL university instructors designed their assessment tasks and the way in which they marked students' work.

In the current study, results prompted the suggestion that the low hourly rate for teaching fee-paying students plus the low monthly salary for teaching scholarship students enticed instructors to teach more hours per week. A large teaching workload was thought to be the reason behind some instructors (four of the six interviewed) reusing the previous year's tests. This raised challenges with maintaining test security, and in some instances, the test information was reported to have been leaked to some of the students prior to test administration. Such assessment implementation had been considered as unethical or unfair to those students who did not have access to such information. This assessment implementation contradicted the recommendations of assessment experts (Lamprianou & Athanasou, 2009; Bachman & Palmer, 2010; Douglas, 2010; Miller et al., 2013) who strongly advised instructors to guard against such poor assessment implementation.

9.2.4.4 The Influence of Departmental Status

Despite the lack of previous research examining the influence of departmental status (i.e., English-major versus English non-major) on instructors' assessment practices, the findings in both phases of this study showed that it impacted instructors' classroom assessment literacy and implementation. The majority of English-major instructors had higher levels of classroom assessment knowledge and demonstrated more positive endorsement towards implementing quality assessment procedures than their English non-major counterparts. Most English-major instructors were also less influenced by students' personal characteristics (e.g., age, gender and appearance) when marking students' work than the majority of English non-major counterparts. Such discrepancies in their classroom assessment literacy levels between these two groups of instructors may have been explained by the fact that the majority of English-major instructors received pre-service assessment training, whereas only a small number of English non-major counterparts did. Furthermore, most English-major instructors, on average, had lower

teaching requirements per week and smaller class sizes than their English non-major counterparts. As such, these findings were not surprising, given the background characteristics of instructors (i.e., pre-service assessment training, number of teaching hours per week and class size) were found to have impacted three of the four measures (i.e., Classroom Assessment Knowledge, Grading Bias and Quality Procedure) of classroom assessment literacy.

9.2.4.5 The Influence of Age

Although there was a lack of previous research that examined the effect of age on instructors' assessment practices, the findings in both phases of this study highlighted that younger instructors demonstrated a higher level of classroom assessment knowledge than older instructors. This finding might be explained by the fact that younger instructors had just graduated from their pre-service teacher education programme, and they could retain more of the assessment theories to be implemented in their assessment practices.

9.2.4.6 The Influence of Teaching Experience

Years of teaching experience were identified in both phases of this study as having no influence on instructors' classroom assessment knowledge. These findings supported some previous studies (King, 2010; Alkharusi et al., 2011; Gotch & French, 2013), but contradicted others (Zhang & Burry-Stock, 1997; Chapman, 2008; Alkharusi, 2011). For instance, King (2010), Alkharusi et al. (2011) and Gotch and French (2013), employing test instruments to measure school teachers' assessment knowledge, found that years of teaching experience had no influence on their assessment knowledge levels. However, studies that employed self-reported measures to examine the levels of school teachers' assessment knowledge found differences. For example, Zhang and Burry-Stock (1997), Chapman (2008) and Alkharusi (2011) reported that school teachers with more teaching experience showed a higher level of self-perceived assessment competence than those with less teaching experience. Such inconsistencies could be explained by the fact

that there were different methodologies used in the studies. The current study employed a test instrument, a series of self-reported measures and document analyses, as opposed to studies that have used solely self-reported measures. Given research showed that self-reported measures had an association with school teachers' tendency to overrate their assessment knowledge level (see Alkharusi et al., 2012), it may have been possible that school teachers with more teaching experience in Zhang and Burry-Stock's (1997), Chapman's (2008) and Alkharusi's (2011) studies overreported their perceived assessment knowledge level, than those with less teaching experience.

9.2.4.7 The Influence of Gender

In the current study, gender was also found to play no role in the instructors' assessment knowledge level. These findings were contrary to those of Alkharusi (2011), who found that female instructors perceived themselves to be more skilful in writing test items and communicating assessment results than male instructors. Given there are contradicting results between the current study and that of Alkharusi's (2011), further research is required.

9.2.4.8 The Influence of Academic Qualification

The level of academic qualification was shown, in both phases of this study, to have no influence on the instructors' assessment knowledge level. These findings were in line with those of Chapman (2008), but contradicted those of King (2010). Chapman (2008) found that school teachers' levels of academic qualifications (i.e., bachelor versus master degree) had no impact on their assessment knowledge. In contrast, King (2010) found that school teachers with advanced degrees, comprising specialist and doctoral studies in education, had significantly higher assessment knowledge levels than those who possessed bachelor or masters degree. Such inconsistencies may have been explained by the different samples recruited within these studies. King's (2010) study focused on various academic qualifications of school teachers, such as bachelor, masters and advanced degrees including specialist and doctorate, whilst the current study

emphasised primarily university instructors with bachelors and masters degrees in a variety of disciplines.

9.2.4.9 The Influence of Professional Development Workshop and Assessment Experience as Students

Although not being explicitly examined in the quantitative phase of this study, the qualitative results highlighted the importance of in-service educational assessment training on instructors' classroom assessment literacy development. It was found that three of the six instructors interviewed attributed a lack of in-service educational assessment training to their perceived low levels of assessment competence. It further revealed that the instructor's prior assessment experience as students played an important role in influencing their assessment beliefs toward endorsing traditional assessment methods (e.g., tests/exams). These findings supported Stiggins' (2010) argument, that the lack of in-service assessment training delivery worldwide was largely responsible for instructors' insufficient assessment literacy. These findings were also consistent with Green and Stager (1986) and Xu and Liu (2009) who found that the instructors' previous assessment experience as students influenced their current assessment implementation. Such findings reinforced Bandura's (1997) social cognitive theory, which contends that this type of learning experience (i.e., assessment methods that student teachers experienced during pre-service teacher preparation) can influence instructors' future assessment practices and beliefs.

9.3 Conclusion

This concluding section presents the implications of the study findings for theory, policy and practice, and the design of pre-service teacher education programmes. Finally, it discusses the study's limitations and future directions for research in the area of classroom assessment literacy.

9.3.1 Implications of the Study Findings

9.3.1.1 Implications for Theory

The present study identified assessment knowledge and personal beliefs about assessment as important facets in measuring classroom assessment literacy amongst Cambodian university instructors. This study supported Fishbein and Ajzen's (1975, 2010) reasoned action theory, Ajzen's (1991, 2005) planned behaviours theory and Bandura's (1989, 1997) social cognitive theory, that contended both individuals' knowledge/skills and their personal beliefs/attitudes underpinned their behaviours/performances. That is, both instructors' classroom assessment knowledge and their personal assessment beliefs have the potential to enable and empower them to implement high quality assessments.

Prior to this study, assessment literacy tended to be theorised solely in terms of the knowledge base and had not been examined explicitly from a classroom-based context. The current study confirmed that classroom assessment knowledge and personal beliefs about assessment as the underpinnings of classroom assessment literacy. Such findings provided support for those experimental studies reported in the field of social psychology (see Crisp & Turner, 2012; Dasgupta, 2013) that demonstrated the effects of individuals' beliefs/attitudes on their judgments, decisions and actions/behaviours (e.g., intergroup relations or social group context), equally as important as their knowledge/skills. As such, this study supported the presence of the social psychological process (i.e., beliefs/attitudes of the instructors) involved in the conduct of classroom-based assessment. The present study provided a theoretical framework for future research into classroom assessment literacy. Hence, it is proposed that "classroom assessment literacy" be redefined as:

"the instructors' acquaintance with knowledge base about the key stages within the assessment process and their capacity to explicitly examine their implicit personal beliefs about assessment".

This revised definition would entail knowledge of all key stages of the assessment process within a classroom-based context (see Chapter 2) and address Scarino's (2013) call for the inclusion of instructors' personal beliefs about assessment into the notion of classroom assessment literacy.

Finally, this study not only identified the constructs that underpinned classroom assessment literacy, but also examined the magnitude and the directions of the relationship of the interaction amongst these constructs. The classroom assessment literacy construct accounted for 44% variance in the Grading Bias variable, 33% variance in the Classroom Assessment Knowledge variable, 31% variance in the Innovative Methods variable, and 24% variance in the Quality Procedure variable. Furthermore, this study provided a methodological approach to measure classroom assessment literacy development.

9.3.1.2 Implications for Policy and Practice

The present study has important implications for policy development in relation to reducing undesirable, extraneous factors influencing instructors' judgements of students' work. For instance, one in 13 instructors reported they were influenced by students' attitudes and non-academic related behaviours when marking their work. However, the presence of these undesirable factors could be minimised when instructors had higher levels of classroom assessment knowledge and demonstrated more positive endorsements toward using innovative assessment methods, and implementing quality procedures and checks on their assessment practices.

A number of implications for instructors and policymakers and administrators have been identified. First, a set of external factors were identified to negatively impact instructors' classroom assessment literacy and implementation including larger class sizes, more teaching hours per week, lower hourly payment rate and monthly salary, ineffective assessment-related policies and lack of educational assessment workshops, all of which can be addressed through university policies and procedures. As such, these findings have direct implications for policymakers and administrators for their immediate interventions to prevent or at least minimise the influence of each of these undesirable

factors in order to maximise the instructors' classroom assessment literacy in conducting high quality assessments. In relation to assessment policy, broadening the purpose of assessment to include both formative and summative functions, and explicitly making the imposed cut-off score for pass in line with the departmental learning goals, would help improve the quality of instructors' assessment practices. In addition, smaller class sizes, and higher pay and lower teaching loads would provide instructors with the opportunity to be more innovative in their assessments. It has also been indicated in the literature that the provision of incentives for job performances can be a powerful tool in creating a good environment (Simon & Pleschová, 2013) and therefore, the introduction of such incentives (e.g., promotion, formal recognition/rewards, assessment conference/workshop attendance) may positively support instructors' classroom assessment literacy development.

Second, instructors' assessment implementation was found to be influenced by various extraneous factors. They indicated that they were influenced by students' attitudes and behaviours when marking their work. Moreover, instructors reported they were influenced by their previous assessment experience, as students, with respect to their endorsements toward using traditional assessment methods. These findings have direct implications for instructors' self-learning through self-reflections and undertaking action research, as well as collaborative learning in terms of having dialogues with their colleagues. These self-reflections enable instructors to identify their deeply personal, espoused beliefs, and critically examine those assumptions in order to justify and/or alter their initial beliefs about assessment (see Dewey, 1933/1960; Zwozdiak-Myers, 2012; Klenowski, 2013b). Instructors' engagement in undertaking action research also plays an important role in them deeply reflecting on their own assessment practices and challenging their implicit personal beliefs about assessment (see Borg, 2013; Mills, 2014). Research shows that involving instructors in undertaking action research on their own classroom assessment practices is a very effective way of improving their assessment literacy, as it helps them learn not only recent educational assessment theories, but also empowers them to alter and evaluate their assessment practices at a relatively low cost (see Swann, Andrews, & Ecclestone, 2011).

Furthermore, it is essential for instructors to engage in collaborative learning within their community of practice (Lave & Wenger, 1991; Wenger, 1998) to share their implicit personal beliefs about assessment. Through this community of assessment practice, instructors have opportunities to participate in dialogues with their colleagues in terms of: having moderation meetings to share understandings of the marking criteria and standards and assessment tasks/tests; discuss the issues encountered in their assessment practices; and challenge each other's perspectives regarding conducting high quality assessments (Bloxham & Boyd, 2007). Such a community of practice can help instructors to broaden their understandings of the theoretical underpinnings of the assessment process. Within this community of social moderation practice, all instructors, including novice, junior and senior, have the opportunities to openly justify their assessment practices (e.g., their decisions in adding marks to borderline students' results in order to pass them). By doing so, they can learn from each other about the types of extraneous factors that can exert influence on their judgments of students' work, and they can minimise the variation of their assessment practices. Research indicates that the assessment literacy of instructors can be improved through self-reflection on their own assessment practices (see Howley, Howley, Henning, Gillam, & Weade, 2013) and their engagement in dialogue with colleagues to share issues regarding their assessment practices (Howley et al., 2013; Klenowski, 2013a; Adie, 2013).

With respect to design and provision of professional development programmes, universities need to balance both educational assessment and teaching techniques, as these two are not synonymous. Because a lack of educational assessment training has been identified as a contributing factor to the instructors' limited classroom assessment knowledge in this study, there is a need for providing instructors with ongoing training regarding key stages of the assessment process. Darling-Hammond and Lieberman (2012), Zwozdiak-Myers (2012) and Borg (2013) have proposed that ongoing professional development programmes have the potential to provide instructors with the most up-to-date knowledge and best practice to fulfil their professional responsibility needs. A wealth of research highlights the crucial role of in-service professional development workshops focusing on educational assessment in promoting the assessment

knowledge of school teachers (Borko et al., 1997; O’Leary, 2008; Sato et al., 2008; Mertler, 2009; Black et al., 2010; Towndrow et al., 2010; Koh, 2011; Griffin et al., 2013).

Finally, the four classroom assessment literacy scales could be used as diagnostic instruments, both within in-service educational assessment programmes and pre-service teacher education programmes. The Classroom Assessment Knowledge test could be utilised to diagnose the levels of classroom assessment knowledge of in-service instructors or student teachers at the beginning of the educational assessment workshops or educational assessment courses, so that appropriate remedies could be used to address issues in a timely manner. It could also be used to monitor growth and/or the effectiveness of intervention. Similarly, the three remaining scales (i.e., Innovative Methods, Grading Bias and Quality Procedure) could be employed to uncover the personal beliefs of in-service instructors or student teachers in relation to the use of innovative assessment methods and quality procedures in their assessment practices, as well as the influence of their personal beliefs and values on marking students’ work. For instance, the Grading Bias measure could be used to help instructors recognise the influence of their own personal beliefs and values on the way in which they mark students’ work.

9.3.1.3 Implications for the Design of Pre-service Teacher Education Programme

The design of pre-service teacher education programmes was also found to have a direct relationship with the limited classroom assessment literacy of the instructors. This study has highlighted various key recommendations to be implemented for the effective design of pre-service teacher education programmes to promote student teachers’ classroom assessment literacy.

Firstly, one way to facilitate student teachers’ classroom assessment literacy development would be to provide them with a stand-alone educational assessment course rather than embedded within another course. Furthermore, the educational assessment course should be treated as the core component and be taught in-depth within the teacher education programme. In so doing, student teachers should be provided with sufficient

learning time to adequately acquire in-depth knowledge, skills and understandings by means of linking theoretical underpinnings of the assessment process with actual assessment practices. Moreover, student teachers should be awarded the degree/certificate on the condition that they have successfully completed the educational assessment course. Popham (2011) has argued that “what is needed to facilitate assessment literacy in teacher education is more than a brief mention of assessment in a course” (p. 265). Research shows that providing formal courses in educational assessment to student teachers can increase their confidence in implementing their assessments (DeLuca & Klinger, 2010) and has the potential to promote their assessment knowledge (Chen, 2005; DeLuca, Chavez, & Cao, 2013). Research also indicates that providing formal courses in educational assessment to student teachers can provide them with a foundation for continued learning about the process for conducting assessments throughout their careers (see DeLuca et al., 2013). Research, however, highlights that when the assessment units are embedded within another course, student teachers who enrolled in such a programme tend to gain limited assessment knowledge and they therefore cannot use the assessment data to enhance their instruction and students’ learning (Schafer & Lissitz, 1987). Griffin et al. (2012, p. 10) have aptly summed up when they note that “formal courses in assessment or educational measurement for pre-service teachers are uncommon,” which contribute to their limited assessment knowledge when entering their teaching careers.

In addition to a stand-alone educational assessment course, having good assessment course curriculum and employing appropriate assessment textbooks have the potential to enable student teachers to acquire in-depth knowledge, skills and understandings of the theoretical underpinnings of the assessment process. As Poth (2013) suggests, there is a need to have an alignment between the assessment knowledge/skills specified in the educational assessment course curriculum and the roles and responsibilities of student teachers. In relation to the use of educational assessment textbooks, Popham (2011) asserts that using appropriate textbooks has the potential to bring a more complete knowledge and understandings of the assessment process to student teachers. Furthermore, it was also proposed that care should be exercised not to adopt educational assessment textbooks which only emphasise dated psychometric

content. Popham's (2011) caution is consistent with Masters' (2013a) recent assertion that educational assessment has been treated at a superficial level in pre-service teacher education programmes. Masters (2013a) also comments that the assessment textbooks employed in such a programme are usually based on 20th century educational assessment concepts, most of which hamper rather than develop, student teachers' explicit understandings about assessment. Popham's (2011) and Masters' (2013a) propositions have been supported by research that indicates that most topics covered in the top-selling introductory educational assessment textbooks primarily focus on traditional assessment aspects (see Campbell & Collins, 2007). Research further highlights that crucial topics, such as assigning grades, selecting and constructing response items, developing rubrics and assessing performance, and interpreting assessment results, are not consistently identified as important or missing from the textbooks (Campbell & Collins, 2007). Popham (2011) has favourably argued for employing the educational assessment textbooks comprising the balance of formative and summative assessments in educational assessment course. This argument is in line with the dual role of classroom assessment in serving both formative and summative purposes (Broadfoot, 2005; Taras, 2005; Black, 2013; Sambell et al., 2013). Understanding the dual role of classroom assessment can assist student teachers in choosing appropriate assessment methods (Andrade, 2010) and balancing formative and summative assessments. According to Sambell et al., (2013), balancing formative and summative assessments can avoid an over-emphasis on summative assessment that can have detrimental effects on students' learning.

Secondly, the next way of facilitating student teachers to develop their classroom assessment literacy is through helping them uncover their explicit personal beliefs about assessment. That is, they should be given opportunities to examine their personal espoused beliefs with respect to their own assessment experience, by means of critically reflecting and explicitly articulating such assessment beliefs in small groups at the beginning of the course. In their learning of the assessment process, student teachers should be provided with opportunities to reflect and articulate their reasons (e.g., selecting the particular assessment methods). In so doing, they are encouraged to explicitly think about the learning aspects to be assessed and for what purpose, and the way in which they will use the assessment data to enhance their instruction and students'

learning. They should also be encouraged to think deeply about the grading practices that can accurately reflect their course objectives/goals. Furthermore, they should be encouraged to explicitly consider the role of validity and reliability characteristics in relation to the selection of assessment methods, assessment purposes and other key stages of the assessment process (see Shepard et al., 2005). Such deep self-reflection can contribute to a higher level of understanding and learning about the complexity (Schunk, Meece, & Pintrich, 2014) of interconnectedness of all key stages of the assessment process. Schunk and Pajares (2009) contend that individuals can alter their beliefs and behaviours accordingly through examining their thoughts and beliefs and engaging in self-evaluations.

Thirdly, another way of facilitating student teachers to promote their classroom assessment literacy is through employing constructive pedagogy and innovative assessments (e.g., self- and peer assessments and performance-based assessments) in order to assess their learning in the pre-service teacher education programme. Through this learning environment, student teachers are provided with opportunities to engage in the assessment process in terms of self-assessing their own work, assessing their peer's work and discussing their perspectives with their peers about assessment materials. These types of learning by doing (see Bandura, 1997), are valuable in helping student teachers link assessment theories with assessment practice and be engaged in critical reflections on their own learning. Recent research highlights that employing such constructive pedagogy can improve student teachers' assessment literacy in terms of broadening their understanding of assessments, developing their metacognitive skills (i.e., the ability to reflect on their own thinking on assessments and make adjustments accordingly) and altering their pre-existing beliefs about assessment (see DeLuca, Chavez, Bellara, & Cao, 2013). Research also shows that engaging student teachers in developing assessment units can broaden their understanding of the purposes of assessments, and communicate those purposes to students and other assessment stakeholders (see Bangert & Kelting-Gibson, 2006).

Fourthly, another way of facilitating student teachers to develop their classroom assessment literacy is through offering them the opportunities to have professional dialogues with their mentors (i.e., associate teachers assigned during the teaching

practicum) and observe how their mentors implement assessments (Graham, 2005). Such complex knowledge (Schunk et al., 2014) with respect to the assessment process can be effectively learnt through vicarious learning or learning by observation (Bandura, 1997).

Finally, facilitation of student teachers to promote their classroom assessment literacy could be achieved through having course instructors, who have expertise in educational assessment, provide instruction to student teachers. Course instructors play an important role in helping student teachers develop their classroom assessment literacy, because they have the capacity to model the assessment practice by not only explicitly provide instruction, but also connecting these theories with actual assessment practices (see Darling-Hammond, Hammerness, Grossman, Rust, & Shulman, 2005).

9.3.2 Limitations of the Study

As with any research, the present study faced several limitations in relation to the sample, course programmes and methodology implemented. One clear limitation was associated with the use of purposive sampling procedures and sample size. All participants in this study were from two English departments within a single city-based university, which limited the generalisability of its results across universities in Cambodia and beyond. Furthermore, the study was limited by the small sample size which may have affected the results. The small sample of participants may also contribute to the lack of power in finding significant differences. For example, the small variation in the academic qualifications of the sample may be considered as a limitation in the current research. Future research focusing on a greater number of participants with doctoral level qualifications should be considered. The second limitation of the study was that it focused on measuring the instructors' classroom assessment literacy within the realm of English course programmes and its results therefore could not be generalised to other course programmes. The third limitation was relevant to the variation in the effective separation of participants by the component scales. The implication of this variation may be associated with the attenuated Grading Bias and Quality Procedure scales (relative to the Classroom Assessment Knowledge and Innovative Methods scales) distorted the correlations reported in Table 7.3. The fourth limitation was relevant to the

model of classroom assessment literacy. Although classroom assessment knowledge and assessment beliefs were found to underpin classroom assessment literacy (as indicated by the confirmatory factor analysis), knowledge and beliefs measures accounted for 33% of the total observed variance in the classroom assessment literacy construct. This suggested that other factors may have not been taken into account in the model. The last limitation was related to the exclusion of classroom observations of instructors' assessment skills in developing and implementing their assessments and analysing their assessment tasks due to the time constraints associated with this study. This exclusion could limit further insights in relation to instructors' classroom assessment literacy. Despite such limitations, the findings from the current study are potentially important enough to invite replication and reconfirmation of the model of classroom assessment literacy with a broader population and other course programmes beyond English in various settings. Furthermore, the integration of the quantitative and qualitative methods employed in this study to examine the classroom assessment literacy of the instructors provides a platform for future research in this area.

9.3.3 Future Research Directions

In light of the current study limitations, future research in the area of classroom assessment literacy should address some potential issues, such as methodology and sample population, in order to obtain a complete understanding of this construct. With regard to methodological issues, future research should conduct a longitudinal study to measure instructors' classroom assessment literacy development. A longitudinal study that examines instructors' classroom assessment literacy development over five years may be more insightful to identify the developmental pathway of their classroom assessment literacy. Future research should also examine the relationship between classroom assessment literacy and practice by employing classroom observations into the design of the study, by means of observing the instructors' classroom assessment skills in developing and implementing assessments in their classroom settings. Such studies should further analyse instructors' assessment tasks, such as quizzes, tests, examinations, assignments/projects and their written feedback (numerical grades and descriptive

feedback) reported to students, to gain a fuller comprehension of their classroom assessment literacy. In relation to the sample population, future research should use random sampling procedures across universities, particularly with larger sample sizes, to enable greater generalisation of the findings. Moreover, further research should incorporate other relevant assessment stakeholders, beyond instructors, including curriculum developers, policymakers and administrators into a single study, given they are also important educational decision makers in relation to assessment outcomes. It has been recently argued that not only instructors, but also other key assessment stakeholders, particularly policymakers and administrators, need to be assessment-literate in order to effectively fulfil their roles and responsibilities (Stiggins, 2010; Taylor, 2009, 2013; Popham, 2014). Such roles and responsibilities may include quality assurance, communicating and disseminating assessment results, and using assessment data for evaluative/accountability purposes. Research also highlights that administrators, who have an understanding of the use of formative assessment themselves, highly engage in supporting school teachers to employ formative assessment to assess students' learning (Moss, Brookhart, & Long, 2013). Although it is still worthwhile to continue to investigate the classroom assessment literacy of instructors, it is necessary to shift our attention to examine the classroom assessment literacy of students, who are also key agents in the assessment process. As Taras (2013) has asserted, "if we can credit our students with the name of learners, then we must also make them central in the assessment processes" (p. 39). Given an increasing recognition of student agency in the assessment process, it has been recently argued that students also need to become assessment-literate (Price et al., 2011; Sambell, 2013; Popham, 2014). Research shows that students' knowledge and understandings of the assessment process have the potential to influence their learning (Price et al., 2012) and their personal beliefs about assessment also play a critical role in the ways in which they undertake assessment tasks (Scouller, 1998; Segers & Dochy, 2001; Struyven, Dochy, & Janssens, 2005).

Much of the improvement of classroom-based assessment and the appropriate use of assessment outcomes will be dependent on all key assessment stakeholders, including instructors, students, curriculum developers, policymakers and administrators, becoming classroom assessment-literate. As Popham (2014, p. 22) has asserted, a "reasonable lump

of assessment literacy is good for almost everyone” who is involved in implementing assessments and/or using assessment outcomes for educational decision-making.

References

- Adams, R., & Khoo, S. (1996). *Quest: The interactive test analysis system*. Melbourne: Australian Council for Educational Research.
- Adie, L. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice*, 20(1), 91-106.
- Ahrens, L., & McNamara, V. (2013). Cambodia: Evolving quality issues in higher education. In L. P. Symaco & C. Brock (Eds.), *Education in South-East Asia* (pp. 47-68). London: Bloomsbury.
- Airasian, P. W. (2000). *Assessment in the classroom: A concise approach* (2nd ed.). Boston: McGraw-Hill.
- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50, 179-211.
- Ajzen, I. (2005). *Attitudes, personality, and behavior* (2nd). New York: Open University Press.
- Alderson, J.C., & Hamp-Lyons, L. (1996). TOEFL preparation courses: A study of washback. *Language Testing*, 13(3), 280-297.
- Alderson, J. C., & Wall, D. (1993). Does washback exist? *Applied Linguistics*, 14 (2), 115-129.
- Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training. *Journal of Turkish Science Education*, 8(2), 39-48.
- Alkharusi, H., Aldhafri, S., Alnabhani, H., & Alkalbani, M. (2012). Educational assessment attitudes, competence, knowledge, and practices: An exploratory study of Muscat teachers in the Sultanate of Oman. *Journal of Education and Learning*, 1(2), 217-232.
- Alkharusi, H., Kazem, A. M., & Al-Musawai, A. (2011). Knowledge, skills, and attitudes of pre-service and in-service teachers in educational measurement. *Asia-Pacific Journal of Teacher Education*, 39(2), 113-123.
- Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, 20(1), 20-34.

- Amengual-Pizarro, M. (2009). Does the English test in the Spanish university entrance examination influence the teaching of English? *English Studies* 90(5), 582-598.
- American Federation of Teachers, National Council on Measurement in Education, & National Education Association (1990). *The standards for Teacher competence in the educational assessment of students*. Retrieved from <http://files.eric.ed.gov/fulltext/ED323186.pdf>
- Anderson, L. W. (2003). *Classroom assessment: Enhancing the quality of teacher decision making*. Mahwah, New Jersey: Lawrence Erlbaum.
- Andrade, H. L. (2010). Summing up and moving forward: Key challenges and future directions for research and development in formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 344-351). New York: Routledge.
- Andrade, H. L., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education: Principles, Policy & Practice* 17(2), 199-214.
- Andrade, H. L., Du, Y., & Wang, X. (2008). Putting rubrics to the test: The effect of a model, criteria generation, and rubric-referenced self-assessment on elementary school students' writing. *Educational Measurement: Issues and Practice* 27(2), 3-13.
- Angoff, W. H. (1974). Criterion-referencing, norm-Referencing, and the SAT. *College Board Review* 92, 3-5, 21.
- Antoniou, P., & James, M. (2014). Exploring formative assessment in primary school classrooms: Developing a framework of actions and strategies. *Educational Assessment, Evaluation and Accountability*, 1-24.
- Arbuckle, J. L. (1983-2011). *AMOS 20.0 user's guide*. Chicago: SPSS Inc.
- Archbald, D. A., & Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Washington, D.C.: National Association of Secondary School Principals.
- Arter, J. (1999). Teaching about performance assessment. *Educational Measurement: Issues and Practice*, 18(2), 30-44.
- ASEAN Secretariat (2009). *ASEAN socio-cultural community blueprint*. Retrieved from <http://www.asean.org/archive/5187-19.pdf>
- Association of Southeast Asian Nations (2007). *The Asian charter*. Retrieved from <http://www.asean.org/archive/publications/ASEAN-Charter.pdf>

- Athanasou, J. A., & Lamprianou, I. (2002). *A teacher's guide to assessment*. Sydney: Social Science.
- Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford: Oxford University Press.
- Bachman, L. F. (2014). Ongoing challenges in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3), (pp. 1586-1603). Oxford: John Wiley and Sons.
- Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.
- Bachman, L., & Palmer, A. (2010). *Language assessment in practice: Developing language assessments and justifying their use in the real world*. Oxford: Oxford University Press.
- Bailey, K.M. (1996). Working for washback: A review of the washback concept in language testing. *Language Testing*, 13, 257-279.
- Bandura, A. (1986). *Social foundations of thought and action: A cognitive social theory*. New York: Prentice Hall, Englewood Cliffs.
- Bandura, A. (1989). Human agency in social cognitive theory. *American psychologist*, 44(9), 1175-1184.
- Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.
- Bangert, A., & Kelting-Gibson, L. (2006). Teaching principles of assessment literacy through teacher work sample methodology. *Teacher Education and Practice* 19(3), 351-364.
- Bennett, R. E., Gottesman, R. L., Rock, D. A., & Cerullo, F. (1993). Influence of behaviour perceptions and gender on teachers' judgments of students' academic skill. *Journal of Educational Psychology* 85(2), 347-356.
- Berger, A. (2012). Creating language-assessment literacy: A model for teacher education. In J. Hüttner, B. Mehlmauer-Larcher, S. Reichl, & B. Schiftner (Eds.), *Theory and practice in EFL teacher education: Bridging the gap* (pp. 57-82). Bristol: Multilingual Matters.
- Biggs, J. (1998). Learning from the Confucian heritage: So size doesn't matter? *International Journal of Educational Research* 29, 723-738.
- Biggs, J. (2012). Enhancing learning through constructive alignment. In J. R. Kirby & M. J. Lawson, (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 117-136). New York: Cambridge University Press.

- Biggs, J., & Tang, C. (2007). *Teaching for quality learning at university* (3rd ed.). London: Open University Press.
- Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17-66). New York: Springer.
- Black, P. J. (1993). Formative and summative assessment by teachers. *Studies in Science Education*, 21, 49-97.
- Black, P. (2013). Formative and summative aspects of assessment: Theoretical and research foundation in the context of pedagogy. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 167-178). Los Angeles: Sage.
- Black, P., Harrison, C., Hodgson, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice* 17(2), 215-232.
- Black, P., & Wiliam, D. (1998a). Assessment and classroom learning. *Assessment in Education: Principles, Policy, & Practice*, 5(1), 7-73.
- Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *The Phi Delta Kappan*, 80(2), 139-148.
- Bloom, B.S., Hastings, J. T., & Madaus, G. F. (1971). *Handbook on formative and summative evaluation of student learning*. New York: McGraw-Hill.
- Bloxham, S. (2009). Marking and moderation in the UK: False assumptions and wasted resources. *Assessment & Evaluation in Higher Education*, 34(2), 209-220.
- Bloxham, S. (2013). Building 'standards' frameworks: The role of guidance and feedback in supporting the achievement of learners. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 64-74). New York: Routledge.
- Bloxham, S., & Boyd, P. (2007). *Developing effective assessment in higher education: A practical guide*. Maidenhead: Open University Press.
- Bloxham, S., & Boyd, P. (2012). Accountability in grading student work: Securing academic standards in a twenty-first century quality assurance context. *British Educational Research Journal*, 38(4), 615-634.
- Bloxham, S., Boyd, P., & Orr, S. (2011). Mark my words: The role of assessment criteria in UK

- higher education grading practices. *Studies in Higher Education*, 36(6), 655-670.
- Blunch, N. J. (2013). *Introduction to structural equation modelling using IBM SPSS statistics and AMOS* (2nd ed.). London: Sage.
- Bol, L., Stephenson, P. L., O'Connell, A. A., & Nunnery, J. A. (1998). Influence of experience, grade level, and subject area on teachers' assessment practices. *The Journal of Educational Research*, 91(6), 323-330.
- Bol, L., & Strage, A. (1996). The contradiction between teachers' instructional goals and their assessment practices in high school biology courses. *Science Education*, 80(2), 145-163.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.
- Bollen, K. A., & Long, J. S. (1993). *Testing structural equation models*. Newbury Park, California: Sage.
- Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.
- Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch analysis in the human sciences*. New York: Springer.
- Borg, S. (2013). *Teacher research in language teaching: A critical analysis*. Cambridge: Cambridge University Press.
- Borko, H., Mayfield, V., Marion, S., Flexer, R., & Cumbo, K. (1997). Teachers' developing ideas and practices about mathematics performance assessment: Successes, stumbling blocks, and implications for professional development. *Teaching and Teacher Education*, 13(3), 259-278.
- Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education*, 15(1), 101-111.
- Boud, D. (1995). *Enhancing learning through self assessment*. London: Kogan Page.
- Boud, D. (2006). Forward. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp. xvii-xix). New York: Routledge.
- Boud, D., & Molloy, E. (2013). What is the problem with feedback? In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 1-10). New York: Routledge.
- Bounchan, S. (2012). The relationship between metacognitive beliefs and academic performance of Cambodian university students. *PowerPoint presentation at the 8th Annual CamTESOL Conference on English Language Teaching*, February 25-26, Phnom Penh, Cambodia.

- Bounchan, S. (2013). *The value of literature in foreign language learning: A case study of tertiary-level literature studies in Cambodia*. Unpublished doctoral thesis. Macquarie University, Sydney, Australia.
- Bradley, E. H., Curry, L. A., & Devers, K. J. (2007). Qualitative data analysis for health services research: Developing taxonomy, themes, and theory. *Health Services Research, 42*(4), 1758-1772.
- Bradshaw, J., & Wheeler, R. (2013). Reporting error and reliability to test-takers: An international review. *Research Papers in Education, 28*(1), 106-117.
- Brindley, G. (2000). *Studies in immigrant English language assessment* (Vol.1). Sydney: National Centre for English Language Teaching and Research, Macquarie University.
- Broadfoot, P.M. (2005). Dark alleys and blind bends: Testing the language of learning. *Language Testing, 22*(2), 123-141.
- Brockmeier, J., & Olson, D. R. (2009). The literacy episteme from Innis to Derrida. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 3-21). Cambridge: Cambridge University Press.
- Brookhart, S. M. (1993). Teachers' grading practices: Meaning and values. *Journal of Educational Measurement, 30*(2), 123-142.
- Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*(4), 279-301.
- Brookhart, S. M. (1999). Teaching about communicating assessment results and grading. *Educational Measurement: Issues and Practice, 18*(1)5-13.
- Brookhart, S. M. (2004). Classroom assessment: Tensions and intersections in theory and practice. *Teachers College Record 106*(3), 429-458.
- Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria, Virginia: Association for Supervision and Curriculum Development.
- Brookhart, S. M. (2010). Mixing it up: Combining sources of classroom achievement information for formative and summative purposes. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 279-296). New York: Routledge.
- Brookhart, S. M. (2011a). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

- Brookhart, S. M. (2011b). *Grading and learning: Practices that support student achievement*. Bloomington: Solution Tree Press.
- Brookhart, S. M. (2013a). Classroom assessment in the context of motivation theory and research. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 35-54). Los Angeles: Sage.
- Brookhart, S. M. (2013b). *How to create and use rubrics for formative assessment and grading*. Alexandria, Virginia: Association for Supervision and Curriculum Development.
- Brown, C. R. (1998). An evaluation of two different methods of assessing independent investigations in an operational pre-university level examination in Biology in England. *Studies in Educational Evaluation* 24(1), 87-98.
- Brown, G. T. L. (2008). Assessment literacy training and teachers' conceptions of assessment. In C. M. Rubie-Davies & C. Rawlinson (Eds.), *Challenging thinking about teaching and learning* (pp. 285-302). New York: Nova Science.
- Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 367-393). Los Angeles: Sage.
- Brown, G. T. L., Kennedy, K. J., Fox, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice*, 16(3), 347-363.
- Brown, G. T. L., Lake, R., & Matters, G. (2011). Queensland teachers' conceptions of assessment: The impact of policy priorities on teacher attitudes. *Teaching and Teacher Education*, 27(1), 210-220.
- Brown, J. D., & Hudson, T. (1998). The alternatives in language assessment. *TESOL Quarterly*, 32(4), 653-675.
- Brown, A. (2012). Ethics in language testing and assessment. In C. Coombe, P. Davidson, B. O'Sullivan, & S. Stoyloff (Eds.), *The Cambridge guide to second language assessment* (pp. 113-121). New York: Cambridge University Press.
- Bryman, A. (2006). Integrating quantitative and qualitative research: How is it done? *Qualitative Research*, 6(1), 97-113.
- Burns, R. B. (2000). *Introduction to research methods* (4th ed.). Sydney: Pearson Education.
- Byrne, B. M. (2010). *Structural equation modelling with AMOS: Basic concepts, applications, and programming* (2nd ed.). New York: Routledge.

- Cambodia-Australia National Examinations Project (CANEP) (2002). CANEP News, *Issue 3*.
Cambridge Advanced Learner's Dictionary (3rd ed.) (2008). Cambridge: Cambridge University Press.
- Campbell, C. (2013). Research on teacher competency in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 71-84). Los Angeles: Sage.
- Campbell, C., & Collins, V. L. (2007). Identifying essential topics in general and special education introductory assessment textbooks. *Educational Measurement: Issues and Practice*, 26 (1), 9-18.
- Campbell, C., & Mertler, C. A. (2004). *Assessment literacy inventory*. Unpublished instrument.
- Care, E., & Griffin, P. (2009). Assessment is for teaching. *Independence* 34(2), 56-59.
- Carless, D. (2013a). Trust and its role in facilitating dialogic feedback. In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 91-103). New York: Routledge.
- Carless, D. (2013b). Sustainable feedback and the development of student self-evaluative capacities. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 113-122). New York: Routledge.
- Chapman, D. W. (2009). Education reforms and capacity development in higher education. In Y. Hirosato & Y. Kitamura (Eds.), *The political economy of educational reforms and capacity development in Southeast Asia: Cases of Cambodia, Laos and Vietnam* (pp. 91-109). New York: Springer.
- Chapman, M. L. (2008). *Assessment literacy and efficacy: Making valid educational decisions*. Unpublished doctoral thesis. University of Massachusetts Amherst.
- Chappuis, J., Stiggins, R., Chappuis, S., & Arter, J. (2012). *Classroom assessment for student learning: Doing it right- using it well* (2nd ed.). Upper Saddle River, New Jersey: Pearson Education.
- Chen, C., Sok, P., & Sok, K. (2007). Benchmarking potential factors leading to education quality: A study of Cambodian higher education. *Quality Assurance in Education* 15(2) 128-148.
- Chen, P. P. (2005). Teacher candidate's literacy in assessment. *Academic Exchange Quarterly* 9(3), 62-66.

- Cheng, L. (2005). *Changing language teaching through language testing: A washback study*. New York: Cambridge University Press.
- Cheng, L. (2008). Washback, impact and consequences. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7), (pp. 349-364). New York: Springer.
- Cheng, L., Rogers, T., & Hu, H. (2004). ESL/ EFL instructors' classroom assessment practices: Purposes, methods, and procedures. *Language Testing*, 21(3), 360-389.
- Cheng, L., Rogers, T., & Wang, X. (2008). Assessment purposes and procedures in ESL/EFL classrooms. *Assessment and Evaluation in Higher Education*, 33(1), 9-32.
- Cheng, L., & Wang, X. (2007). Grading, feedback, and reporting in ESL/EFL classrooms. *Language Assessment Quarterly*, 4(1), 85-107.
- Cheng, L., & Curtis, A. (2012). Test impact and washback: Implications for teaching and learning. In C. Coombe, P. Davidson, B. O' Sullivan, & S. Stoyanoff (Eds.), *The Cambridge guide to second language assessment* (pp. 89-95). New York: Cambridge University Press.
- Cizek, G. J. (2009). Reliability and validity of information about student achievement: Comparing large-scale and classroom testing contexts. *Theory into practice* 48(1), 63-71.
- Cizek, G. J., Rachor, R. E., & Fitzgerald, S. (1995). Further investigation of teachers' assessment practices. *Paper presented at the annual meeting of the American Educational Research Association*, April 18-22, San Francisco, California. (ERIC Reproduction Service No. ED 384 613).
- Clayton, S. (2008). The problem of 'choice' and the construction of the demand for English in Cambodia. *Language policy*, 7(2), 143-164.
- Clayton, T. (2006). *Language choice in a nation under transition: English language spread in Cambodia*. New York: Springer.
- Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological assessment*, 7(3), 309-319.
- Cohen, J. (1969). *Statistical power analysis for the behavioural sciences*. New York: Academic Press.
- Collins, J. (1995). Literacy and literacies. *Annual Review of Anthropology*, 24, 75-93.
- Connolly, S., Klenowski, V., & Wyatt-Smith, C. (2012). Moderation and consistency of teacher judgement: Teachers' views. *British Educational Research Journal*, 38(4), 593-614.

- Cooksey, R. W., Freebody, P., & Wyatt-Smith, C. (2007). Assessment as judgment-in-context: Analysing how teachers evaluate students' writing. *Educational Research and Evaluation* 13(5), 401-434.
- Cortina, J. M. (1993). What is coefficient alpha? An examination of theory and applications. *Journal of applied psychology*, 78(1), 98-104.
- Cresswell, M. J. (1986). Examination grades: How many should there be? *British Educational Research Journal*, 12(1), 37-54.
- Creswell, J. W. (2012). *Educational research: Planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Boston: Pearson Education.
- Creswell, J. W. (2013). *Qualitative inquiry and research design: Choosing among five approaches* (3rd ed.). Thousand Oaks: Sage.
- Creswell, J. W. (2014). *Research design: Qualitative, quantitative, and mixed methods Approaches* (4th ed.). Thousand Oaks, California: Sage.
- Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and conducting mixed methods research* (2nd ed.). Thousand Oaks, California: Sage.
- Crisp, R. J., & Turner, R. N. (2012). The imagined contact hypothesis. In J. M. Olson & M. P. Zanna (Eds.), *Advances in experimental social psychology* (pp. 125-182). Amsterdam: Elsevier.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16(3), 297-334.
- Cronbach, L. J. (1990). *Essentials of psychological testing* (5th ed.). New York: Haper Collins.
- Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, 58(4), 438-481.
- Cross, L.H., & Frary, R. B. (1999). Hodgepodge grading: Endorsed by students and teachers alike. *Applied Measurement in Education*, 12(1), 53-72.
- Cumming, J. J. (2010). Classroom assessment in policy context (Australia). In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.) (pp. 417-424). Oxford: Elsevier Science.
- Dahlgren, L. O., Fejes, A., Abrandt-Dahlgren, M., & Trowald, N. (2009). Grading systems, features of assessment and students' approaches to learning. *Teaching in Higher Education* 14(2), 185-194.

- Darling-Hammond, L. (2006). *Powerful teacher education: Lessons from exemplary programs*. San Francisco: Jossey-Bass.
- Darling-Hammond, L., & Adamson, F. (2013). *Developing assessments of deeper learning: The costs and benefits of using tests that help students learn*. Stanford, California: Stanford Center for Opportunity Policy in Education.
- Darling-Hammond, L., Hammerness, K., Grossman, P., Rust, F. & Shulman, L. (2005). The design of teacher education programs. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 390-441). New York: Jossey-Bass.
- Darling-Hammond, L., & Lieberman, A. (2012). Teacher education around the world: What can we learn from international practice? In L. Darling-Hammond & A. Lieberman (Eds.), *Teacher education around the world: Changing policies and practices* (pp. 151-169). New York: Routledge.
- Dasgupta, N. (2013). Implicit attitudes and beliefs adapt to situations: A decade of research on the malleability of implicit prejudice, stereotypes, and the self-concept. In P. Devine & A. Plant (Eds.), *Advances in experimental social psychology* (pp. 233-279). New York: Elsevier.
- Daugherty, R. (2010). Summative assessment by teachers. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.) (pp. 384-391). Oxford: Elsevier Science.
- Davidheiser, S. A. (2013). *Identifying areas for high school teacher development: A study of assessment literacy in the Central Bucks School District*. Unpublished doctoral thesis. The Drexel University.
- Davidson, D. (1963) Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), 685-700.
- Davidson, D. (2001). *Essays on actions and events* (2 ed.). New York: Oxford University Press.
- Davies, A. (1997a). Introduction: The limits of ethics in language testing. *Language testing*, 14(3), 235-241.
- Davies, A. (1997b). Demands of being professional in language testing. *Language testing*, 14(3), 328-339.
- Davies, A. (2008). Textbook trends in teaching language testing. *Language Testing*, 25(3), 327-347.
- Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher

- assessment practices in Australian and Hong Kong secondary schools. *Language Testing*, 21(3), 305-334.
- Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly*, 43(3), 393-415.
- 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.
- DeLuca, C., Chavez, T., Bellara, A., & Cao, C. (2013). Pedagogies for pre-service assessment education: Supporting teacher candidates' assessment literacy development. *The Teacher Educator* 48(2), 128-142.
- DeLuca, C., Chavez, T., & Cao, C. (2013). Establishing a foundation for valid teacher judgement on student learning: The role of pre-service assessment education. *Assessment in Education: Principles, Policy & Practice*, 20(1), 107-126.
- DeLuca, C., & Klinger, D. A. (2010). Assessment literacy development: Identifying gaps in teacher candidates' learning. *Assessment in Education: Principles, Policy & Practice* 17(4), 419-438.
- Dennis, I. (2007). Halo effects in grading student projects. *Journal of Applied Psychology* 92(4), 1169-1176.
- Dewey, J. (1933/1960). *How we think: A restatement of the relation of reflective thinking to the educative process*. Lexington, Massachusetts: D C. Heath.
- Diamantopoulos, A., & Sigua, J. A. (2000). *Introducing LISREL: A guide for the uninitiated*. Sage.
- Dick, B. (1990). *Convergent interviewing* (3rd ed.). Brisbane: Interchange.
- Douglas, D. (2010). *Understanding language testing*. London: Taylor and Francis.
- Duggan, S. J. (1996). Education, teacher training and prospects for economic recovery in Cambodia. *Comparative Education*, 32(3), 361-375.
- Duggan, S. J. (1997). The role of international organizations in the financing of higher education in Cambodia. *Higher Education*, 34, 1-22.
- Dunbar, S. B., Koretz, D. M., & Hoover, H. D. (1991). Quality control in the development and use of performance assessments. *Applied Measurement in Education*, 4(4), 289-303.

- Dweck, C. (1999). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Taylor & Francis.
- Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (2nd ed.). Thousand Oaks, California: Corwin.
- Eastman, J. K., Iyer, R., & Reisenwitz, T. H. (2011). The impact of unethical reasoning on different types of academic dishonesty: An exploratory study. *Journal of College Teaching & Learning (TLC)*, 5(12), 7-15.
- Ebel, R. (1969). The relation of scale fineness to grade accuracy. *Journal of Educational Measurement*, 6(4), 217-221.
- Edmonds, W. A., & Kennedy, T. D. (2013). *An applied reference guide to research designs: Quantitative, qualitative, and mixed Methods*. Los Angeles: Sage.
- Entwistle, N. J. (2012). The quality of learning at university: Integrative understanding and distinctive ways of thinking. In J. R. Kirby & M. J. Lawson, (Eds.), *Enhancing the quality of learning: Dispositions, instruction, and learning processes* (pp. 15-31). New York: Cambridge University Press.
- Entwistle, A., & Entwistle, N. J. (1992). Experiences of understanding in revising for degree examinations. *Learning and Instruction*, 2, 1-22.
- Falchikov, N. (2004). Involving students in assessment. *Psychology Learning and Teaching*, 3(2), 102-108.
- Falchikov, N., & Boud, D. (2008). The role of assessment in preparing for lifelong learning: Problems and challenges. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 87-99). New York: Routledge.
- Fereday, J., & Muir-Cochrane, E. (2006). Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development. *International Journal of Qualitative Methods*, 5(1), 1-13.
- Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Massachusetts: Addison-Wesley.
- Fishbein, M., & Ajzen, I. (2010). *Predicting and changing behavior: The reasoned action Approach*. New York: Psychology Press (Taylor & Francis).
- Fives, H., & Buehl, M. M. (2012). Spring cleaning for the “messy” construct of teachers’ beliefs:

- What are they? Which have been examined? What can they tell us? In K. R. Harris, S. Graham, & T. Urdan (Eds.), *APA educational psychology handbook* (Vol. 2, pp. 471-499). Washington, D.C.: American Psychological Association.
- Fleming, M., & Chamber, B. (1983). Teacher-made tests: Windows on the classroom. In W. E. Hathaway (Ed.), *Testing in the schools: New directions for testing and measurement* (pp. 29-39). San Francisco: Jossey-Bass.
- Forster, M., & Masters, G. (2010). Progression and assessment: Developmental assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.) (pp. 369-377). Oxford: Elsevier Science.
- Fox, J. (2008). Alternative assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language and education* (2nd ed., Vol. 7), (pp. 97-108). New York: Springer.
- Frary, R. B. (2002). More multiple-choice item writing do's and don'ts. In L. Rudner & W. Schafer (Eds.), *What teachers need to know about assessment* (pp. 61-64). Washington, D.C.: National Educational Association.
- Frary, R. B., Cross, L. H., & Weber, L. J. (1993). Testing and grading practices and opinions of secondary teachers of academic subjects: Implications for instruction in measurement. *Educational Measurement: Issues and Practice*, 12(3), 23-30.
- Fulcher, G. (2012). Assessment literacy for the language classroom. *Language Assessment Quarterly* 9(2), 113-132.
- Gay, G. H. (1990). Standardized tests: Irregularities in administering of tests affect test results. *Journal of Instructional Psychology*, 17(2), 93-103.
- Gibbs, G. (2006). How assessment frames student learning. In C. Bryan & K. Clegg (Eds.), *Innovative assessment in higher education* (pp.23-36). Hoboken: Routledge.
- Gibbs, G. (2007). *Analyzing qualitative data*. London: Sage.
- Gibbs, G., & Lucas, L. (1997). Coursework assessment, class size and student performance: 1984-94. *Journal of further and higher education* 21(2), 183-192.
- Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education* 1(1), 3-31.
- Gillham, B. (2000). *Case study research methods*. London: Continuum.
- Gillis, S. A. (2003). *The Domains of vocational assessment decision-making*. Unpublished doctoral thesis. The University of Melbourne, Australia.

- Gillis, S., Bateman, A., & Clayton, B. (2009). *A code of professional practice for validation and moderation*. Melbourne: National Quality Council.
- Gillis, S., & Griffin, P. (2008). Competency assessment. In J. Athanasou (Ed.), *Adult education and training* (pp. 233-256). Sydney: David Barlow.
- Gipps, C. (1994a). *Beyond testing: Towards a theory of educational assessment*. London: The Palmer Press.
- Gipps, C. (1994b). Quality in teacher assessment. In W. Harlen (Ed.), *Enhancing quality in assessment* (pp. 71-86). London: Paul Chapman.
- Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist* 18(8), 519-521.
- Glesne, C. (2011). *Becoming qualitative researchers: An introduction* (4th ed.). New York: Pearson Education.
- Goffman, E. (1959). *The presentation of self in everyday life*. New York: Doubleday.
- Gotch, C. M., & French, B. F. (2013). Elementary teachers' knowledge and self-Efficacy for measurement concepts. *The Teacher Educator* 48(1), 46-57.
- Graham, P. (2005). Classroom-based assessment: Changing knowledge and practice through pre-service teacher education. *Teaching and Teacher Education* 21(6), 607-621.
- Greaney, V., & Kellaghan, T. (1995). *Equity issues in public examinations in developing countries*. Washington, D.C.: World Bank.
- Green, A. (2014). *Exploring language assessment and testing: Language in action*. New York: Routledge.
- Green, K. E., & Stager, S. F. (1986). Measuring attitudes of teachers toward testing. *Measurement and Evaluation in Counseling and development*, 19, 141-150.
- Green, T. F. (1971). *The activities of teaching*. New York: McGraw-Hill.
- Greene, J. C. (2007). *Mixed methods in social inquiry*. San Francisco: Jossey-Bass.
- Greenstein, L. (2004). *Finding balance in classroom assessment: High school teachers' knowledge and practice*. Unpublished doctoral thesis. Johnson & Wales University.
- Griffin, P. (2007). The comfort of competence and the uncertainty of assessment. *Studies in Educational Evaluation* 33(1), 87-99.
- Griffin, P. (2009). Teachers' use of assessment data. In C. Wyatt-Smith & J. J. Cumming

- (Eds.), *Educational assessment in the 21st century* (pp. 183-208). Dordrecht [Netherlands]: Springer.
- Griffin, P., Care, E., & McGaw, B. (2012). The changing role of education and schools. In P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 1-15). New York: Springer.
- Griffin, P., Care, E., Robertson, 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 mediated human computer interaction* (pp. 39-54). Hershey, PA: IGI Global.
- Griffin, P., & Nix, P. (1991). *Educational assessment and reporting: A new approach*. Sydney: Harcourt Brace Jovanovich.
- Guerin, E. M. C. (2010). Initial findings from a pilot Italian study of foreign language teachers' stated language assessment knowledge-base and needs. *Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching, Vol.4*.
- Gullickson, A. R. (1984). Teacher perspectives of their instructional use of tests. *The Journal of Educational Research*, 77(4), 244-248.
- Gullickson, A. R. (1986). Teacher education and teacher-perceived needs in educational measurement and evaluation. *Journal of Educational Measurement*, 23(4), 347-354.
- Gullickson, A. R., & Ellwein, M. C. (1985). Post hoc analysis of teacher-made tests: The goodness-of-fit between prescription and practice. *Educational Measurement: Issues and Practice*, 4(1), 15-18.
- Guskey, T. R. (2013). Defining student achievement. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 3-6). New York: Routledge.
- Guskey, T. R., & Jung, L. A. (2013). *Answers to essential questions about standards, assessments, grading, and reporting*. Thousand Oaks, California: Corwin.
- Haertel, E. H. (1985). Construct validity and criterion-referenced testing. *Review of educational Research*, 55(1), 23-46.
- Haertel, E. H. (2006). Reliability. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 65-110). Westport, Conn Praeger. National Council on Measurement in Education and American Council on Education.
- Haing, S. (2012). *Perceptions of lecturers and students on the quality of teaching in Cambodian*

- higher education: A case study in English department of one university in Phnom Penh.*
Unpublished Master's thesis. Royal University of Phnom Penh.
- Hambleton, R. K., & Novick, M. R. (1973). Toward an integration of theory and method for criterion-referenced tests 1, 2. *Journal of Educational Measurement*, 10(3), 159-170.
- Harlen, W. (1994). Issues and approaches to quality assurance and quality control in assessment. In W. Harlen (Ed.), *Enhancing quality in assessment* (pp.12-25). London: Paul Chapman.
- Harlen, W. (2005a). Teachers' summative practices and assessment for learning – tensions and synergies. *Curriculum Journal* 16(2), 207-223.
- Harlen, W. (2005b). Trusting teachers' judgement: Research evidence of the reliability and validity of teachers' assessment used for summative purposes. *Research Papers in Education* 20(3), 245-270.
- Harlen, W. (2007). *Assessment of learning*. London: Sage.
- Harlen, W., & Crick, R. D. (2003). Testing and motivation for learning. *Assessment in Education: Principles, Policy & Practice*, 10(2), 169-207.
- Harlen, W., & James, M. (1997). Assessment and learning: Differences and relationships between formative and summative assessment. *Assessment in Education*, 4(3), 365-379.
- Hasselgreen, A., Carlsen, C., & Helness, H. (2004). *European survey of language testing and assessment needs. Part 1: General findings*. Gothenburg, Sweden: European Association for Language Testing and Assessment. Retrieved from <http://www.ealta.eu.org/documents/resources/survey-report-pt1.pdf>
- Heng, K. (2013a). The relationships between student engagement and the academic achievement of first-year university students in Cambodia. *The Asia-Pacific Education Researcher*, 1-11.
- Heng, K. (2013b). *Factors influencing college students' academic achievement in Cambodia: A case study*. *ASEAN Journal of Teaching and Learning in Higher Education* 5 (2), 34-49.
- Heritage, M. (2010). *Formative assessment and next-generation assessment systems: Are we losing an opportunity*. Washington, D.C.: Council of Chief State School Officers. Retrieved from www.ccsso.org/Documents/2010/Formative_Assessment_Next_Generation_2010.pdf
- Heritage, M. (2013a). Gathering evidence of student understanding. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 179-195). Los Angeles: Sage.

- Heritage, M. (2013b). *Formative assessment in practice: A process of inquiry and action*. Cambridge, Massachusetts: Harvard Education Press.
- Herold, B. (2011). Confession of a cheating teacher. *The Philadelphia Public School Notebook*. Retrieved from <http://thenotebook.org/blog/113913/confession-cheating-teacher>
- Heyneman, S. P. (1987). Uses of examinations in developing countries: Selection, research, and education sector management. *International Journal of Educational Development* 7(4), 251-263.
- Heyneman, S. P., & Ransom, A. W. (1990). Using examination and testing to improve educational quality. *Educational Policy* 4(3), 177-192.
- Hill, K., & McNamara, T. (2012). Developing a comprehensive, empirically based research framework for classroom-based assessment. *Language Testing* 29(3), 395-420.
- Hirosato, Y., & Kitamura, Y. (2009). Introduction. In Y. Hirosato & Y. Kitamura (Eds.), *The political economy of educational reforms and capacity development in Southeast Asia: Cases of Cambodia, Laos and Vietnam* (pp. 1-3). New York: Springer.
- Hoge, R. D., & Butcher, R. (1984). Analysis of teacher judgments of pupil achievement levels. *Journal of Educational Psychology* 76(5), 777-781.
- Holbrook, A. L., & Krosnick, J. A. (2010). Social desirability bias in voter turnout reports tests using the item count technique. *Public Opinion Quarterly* 74(1), 37-67.
- 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.
- Hoover, N. R., & Abrams, L. M. (2013). Teachers' instructional use of summative student assessment data. *Applied Measurement in Education* 26(3), 219-231.
- Howes, D., & Ford, D. (2011). Negotiating globalization: The Royal University of Phnom Penh, Cambodia. In S. Marginson, S. Kaur, & E. Sawir. (Eds.), *Higher education in the Asia-Pacific: Strategic responses to globalization* (pp. 161-177). New York: Springer.
- Howley, M. D., Howley, A., Henning, J. E., Gillam, M. B., & Weade, G. (2013). Intersecting domains of assessment knowledge: School typologies based on interviews with secondary teachers. *Educational Assessment* 18, 26-48.
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*:

A Multidisciplinary Journal, 6(1), 1-55.

Huerta-Macias, A. (2002). Alternative assessment: Responses to commonly asked questions.

In J. C. Richards & W. A. Renandya (Eds.), *Methodology in language teaching: An anthology of current practice* (pp. 338-343). New York: Cambridge University Press.

Hughes, A. (1989). *Testing for language teachers* (2nd ed.). New York: Cambridge University Press.

Huhta, A., Hirvala, T., & Banerjee, J. (2005). *European survey of language testing and assessment needs. Part 2: Regional findings*. Gothenburg, Sweden: European Association for Language Testing and Assessment. Retrieved from http://users.jyu.fi/~huhta/ENLTA2/First_page.htm

IBM Corp. (2011). *IBM SPSS Amos, version 20*. Armonk, NY: IBM Corp.

IBM Corp. (2011). *IBM SPSS statistics for windows software, version 20*. Armonk, NY: IBM Corp.

Impara, J. C., Plake, B. S., & Fager, J. J. (1993). Teachers' assessment background and attitudes toward testing. *Theory into practice*, 32, 113- 117.

Inbar-Lourie, O. (2008a). Constructing a language assessment knowledge base: A focus on language assessment courses. *Language Testing*, 25(3), 385-402.

Inbar-Lourie, O. (2008b). Language assessment culture. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopaedia of Language and Education* (2nd ed., Vol. 7). (pp. 285-299). New York: Springer.

Inbar-Lourie, O., & Donitsa-Schmidt, S. (2009). Exploring classroom assessment practices: The case of teachers of English as a foreign language. *Assessment in Education: Principles, Policy & Practice*, 16(2), 185-204.

Izard, J. F. (1998). Quality assurance in educational testing. In National Education Examinations Authority (Eds.), *The effects of large-scale testing and related problems: Proceedings of the 22nd Annual Conference of the International Association for Educational Assessment* (pp.17-23). Beijing, China: Foreign Language Teaching and Research Press.

Izard, J. F. (2004a). Gathering evidence for learning. *Paper presented at the AARE Conference*, November-December, Melbourne. Retrieved from <http://www.aare.edu.au> [search code IZA04877]. Melbourne, Vic.: Australian Association for Research in Education.

Izard, J. F. (2004b). Impediments to sound use of formative assessment (and actions we should

- take to improve assessment for learning). *Paper presented at the AARE Conference*, November-December, Melbourne.
- Izard, J. F. (2006). Quality assurance: Asking the right questions. *Paper presented at the 32nd Annual Conference of the International Association for Educational Assessment*, May 21-26, Singapore.
- Izard, J. F., & Jeffery, P. (2004). Implications for sound practice in assessment for learning. *Paper presented at the AARE Conference*, November-December, Melbourne. Retrieved from <http://www.aare.edu.au> [search code IZA04855]. Melbourne, Vic.: Australian Association for Research in Education.
- Johnson, R. B., & Onwuegbuzie, A. J. (2004). Mixed methods research: A research paradigm whose time has come. *Educational researcher* 33(7), 14-26.
- Jolly, B., & Boud, D. (2013). Written feedback: What is it good for and how can we do it well? In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 104-124). New York: Routledge.
- Jöreskog, K. G., & Sörbom, D. (1989). *LISREL 7: User's reference guide*. Chicago: Scientific Software International.
- Joriskög, K., & Sörbom, D. (1993). *LISREL 8: Structural equation modelling with the SIMPLIS command language*. Chicago: Scientific Software International.
- Joughin, G. (2009). Assessment, learning and judgment in higher education: A critical review. In G. Joughin (Ed.), *Assessment, learning and judgment in higher education* (pp. 1-15). New York: Springer.
- Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed.) (pp. 17-64). Westport, Conn Praeger. National Council on Measurement in Education and American Council on Education.
- Kao, S., & Som, S. (1996). CAMSET: Training the teachers. *Paper presented at the Seminar on English Language Teaching in Cambodia*, May 27-29, Phnom Penh, Cambodia.
- Kehoe, J. (2002). Writing multiple-choice test items. In L. Rudner & W. Schafer (Eds.), *What teachers need to know about assessment* (pp. 57-60). Washington, D.C.: National Educational Association.
- Kellaghan, T., & Greaney, V. (1992). *Using examinations to improve education: A study of fourteen African countries*. Washington, D.C.: The World Bank.

- Kelloway, E. K. (1998). *Using LISREL for structural equation modelling: A researcher's guide*. Thousand Oaks, California: Sage.
- King, J. D. (2010). *Criterion-referenced assessment literacy of educators*. Unpublished doctoral thesis. The University of Southern Mississippi.
- Klenowski, V. (2010). Portfolio assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.) (pp. 236-242). Oxford: Elsevier Science.
- Klenowski, V. (2013a). Towards improving public understanding of judgement practice in standards-referenced assessment: An Australian perspective. *Oxford Review of Education*, 1-16.
- Klenowski, V. (2013b). Editorial: Investigating the complexity of judgement practice. *Assessment in Education: Principles, Policy & Practice* 20(1), 1-4.
- Klenowski, V., & Adie, L. E. (2009). Moderation as judgement practice: Reconciling system level accountability and local level practice. *Curriculum Perspectives*, 29(1), 10-28.
- Klenowski, V., & Wyatt-Smith, C. (2010). Standards, teacher judgement and moderation in contexts of national curriculum and assessment reform. *Assessment Matters*, 2, 107-131.
- Kline, R. B. (2011). *Principles and practice of structural equation modelling* (3rd ed.). New York: The Guilford Press.
- Kögler, H. H. (2012). Agency and the other: On the intersubjective roots of self-identity. *New Ideas in Psychology*, 30(1), 47-64.
- Koh, K. H. (2011). Improving teachers' assessment literacy through professional development. *Teaching Education*, 22(3), 255-276.
- Kramsch, C. (1995). The cultural component of language teaching. *Language, Culture and Curriculum*, 8(2), 83-92.
- Kuch, N. (2013, November 13). Primary schools to add English from grade 4. *The Cambodia Daily*, P. 17.
- Kunnan, A. J. (1999). Recent development in language testing. *Annual Review of Applied Linguistics*, 19, 235-253.
- Kunnan, A. J. (2014). Fairness and justice in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3), (pp. 1098-1114). Oxford: John Wiley and Sons.
- Kvale, S. (1996). *InterViews: An introduction to qualitative research interviewing*. Thousand

- Oaks, California: Sage.
- Lai, E. R., & Waltman, K. (2008). Test preparation: Examining teacher perceptions and practices. *Educational Measurement: Issues and Practice*, 27(2), 28-45.
- Lamont, P. (2013). *Extraordinary beliefs: A historical approach to a psychological problem*. Cambridge: Cambridge University Press.
- Lamprianou, I., & Athanasou, J. A. (2009). *A teacher's guide to educational assessment (Revised edition)*. Rotterdam: Sense.
- Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. New York: Cambridge University Press.
- Lawson, K., & Yorke, J. (2009). The development of moderation across the institution: A comparison of two approaches. *Paper presented at the ATN Assessment Conference*, Melbourne, November 19-20, Australia. Retrieved from http://www.postgradolinguistica.ucv.cl/dev/documentos/90,906,Perceptions_technologies_gruba_2009.pdf#page=236
- Leighton, J. P. (2011). Editorial. *Educational Measurement: Issues and Practice* 30(1), 1-2.
- Leighton, J. P., Gokierto, R. J., Cor, M. K., & Heffernan, C. (2010). Teacher beliefs about the cognitive diagnostic information of classroom-versus large-scale tests: Implications for assessment literacy. *Assessment in Education: Principles, Policy & Practice*, 17(1), 7-21.
- Leung, C. (2014). Classroom-based assessment issues for language teacher education. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3), (pp. 1510-1519). Oxford: John Wiley and Sons.
- Leung, F. K. S. (2002). In search of an East Asian identity in mathematics education. *Educational Studies in Mathematics* 47(1), 35-51.
- Lin, C. H. S., & Wen, L. Y. M. (2007). Academic dishonesty in higher education-A nationwide study in Taiwan. *Higher Education*, 54(1), 85-97.
- Linn, R. L., & Gronlund, N. E. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, New Jersey: Prentice Hall.
- Linn, R. L., & Miller, M.D. (2005). *Measurement and assessment in teaching* (9th ed.). Upper Saddle River, New Jersey: Pearson Education.
- López Mendoza, A. A., & Bernal Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in language assessment. *Profile* 11 (2), 55-70.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. London: Addison-Wesley.
- Luxia, Q. (2007). Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China. *Assessment in Education* 14(1), 51-74.
- Malouff, J. (2008). Bias in grading. *College Teaching* 56(3), 191-192.
- Malone, M. E. (2008). Training in language assessment. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopaedia of Language and Education* (2nd ed., Vol. 7), (pp. 225-239). New York: Springer.
- Malone, M. E. (2013). The essentials of assessment literacy: Contrasts between testers and users. *Language Testing* 30(3), 329-344.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 57(3), 519-530.
- Mardia, K. V. (1974). Applications of some measures of multivariate skewness and kurtosis in testing normality and robustness studies. *Sankhyā: The Indian Journal of Statistics*, 36, 115-128.
- Marsden, P. V., & Wright, J. D. (2010). *Handbook of survey research* (2nd ed.). Bingley: Emerald.
- Marso, R. N., & Pigge, F. L. (1987). Teacher-made tests and testing: Classroom resources, guidelines and practices. *Paper presented at the annual meeting of the Midwestern Educational Research Association*, October 15-17, Chicago, Illinois. (ERIC Reproduction Service No. ED 291 781).
- Marso, R. N., & Pigge, F. L. (1993). Teachers' testing knowledge, skills, and practice. In S. L. Wise (Ed.), *Teacher training in measurement and assessment skills*. Buros Institute of Mental Measurements, University of Nebraska- Lincoln.
- Marton, F., & Säljö, R. (1976a). On qualitative differences in learning. I. Outcome and process. *British Journal of Educational Psychology*, 46, 4-11.
- Marton, F., & Säljö, R. (1976b). On qualitative differences in learning. II. Outcome as a function of the learner's conception of the task. *British Journal of Educational Psychology*, 46, 115-127.
- Marton, F., & Säljö, R. (1997). Approaches to learning. In F. Marton, D. Hounsell, &

- N. Entwistle (Eds.), *The experience of learning: Implications for teaching and studying in higher education* (pp. 39-58). Edinburgh: Scottish Academic Press.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, 47(2), 149-174.
- Masters, G. N. (2013a). *Reforming educational assessment: Imperatives, principles and challenges*. Australian Education Review 57. Melbourne: Australian Council for Educational Research.
- Masters, G. N. (2013b). *Towards a growth mindset in assessment*. Occasional Essays. Melbourne: Australian Council for Educational Research.
- Mathew, R., & Poehner, M. E. (2014). Monitoring progress in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 2), (pp. 631-645). Oxford: John Wiley and Sons.
- Maxwell, G. (2006). Quality management of school-based assessments: Moderation of teacher judgments. *Paper presented at the 32nd International Association for Educational Assessment Conference*, May 21-26, Singapore.
- Maxwell, G. (2007). Implications for moderation of proposed changes to senior secondary school syllabuses. *Paper commissioned by the Queensland Studies Authority*. Brisbane: Queensland Studies Authority.
- Maxwell, G. (2010). Moderation of student work by teachers. In P. Peterson, E. Baker & B. McGaw (Eds.), *International encyclopaedia of education* (3rd ed.) (pp. 457-463). Oxford, Elsevier.
- Mayo, S. T. (1967). *Pre-service preparation of teachers in educational measurement. Final report*. Chicago, IL. Loyola University (ERIC Document Reproduction Service No. ED021784).
- Mayr, E. (2011). *Understanding human agency*. New York: Oxford University Press.
- McArthur, J., & Huxham, M. (2013). Feedback unbound: From master to usher. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 92-102). New York: Routledge.
- McMillan, J. H. (2001). Secondary teachers' classroom assessment and grading practices. *Educational Measurement: Issues and Practice*, 20(1), 20-32.
- McMillan, J. H. (2014). *Classroom assessment: Principles and practice for effective standards-based instruction* (6th ed.). Boston: Pearson Education.

- McMillan, J. H., & Nash, S. (2000). Teacher classroom assessment and grading practices decision making. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, April 25-27, New Orleans, LA. (ERIC Reproduction Service No. ED 447 195).
- McNamara, T. (2000). *Language testing*. Oxford: Oxford University Press.
- Mertler, C. A. (1999). Assessing student performance: A descriptive study of the classroom assessment practices of Ohio teachers. *Education*, 120, 285-296.
- Mertler, C. A. (2000). Teacher-centered fallacies of classroom assessment reliability and validity. *Mid-Western Educational Researcher*, 13(4), 29-35.
- Mertler, C. A. (2003). Pre-service versus in-service teachers' assessment literacy: Does classroom experience make a difference? *Paper presented at the annual meeting of the Mid-Western Educational Research Association*, October 15-18, Columbus, Ohio.
- Mertler, C. A. (2005). Secondary teachers' assessment literacy: Does classroom experience make a difference? *American Secondary Education*, 33(2), 76-92.
- Mertler, C. A. (2009). Teachers' assessment knowledge and their perceptions of the impact of classroom assessment professional development. *Improving Schools*, 12(2), 101-113.
- Mertler, C. A., & Campbell, C. (2005). Measuring teachers' knowledge and application of classroom assessment concepts: Development of the assessment literacy inventory. *Paper presented at the annual meeting of the American Educational Research Association*, April 11-15, Montréal, Quebec, Canada.
- Messick, S. (1989). Validity. In R. L. Linn (Ed), *Educational measurement* (3rd ed.), (pp. 13-103). New York: MacMillan.
- Messick, S. (1996). Validity and washback in language testing. *Language Testing*, 13(3), 241-256.
- Miles, M., & Huberman, A. (1994). *Qualitative data analysis: An expanded sourcebook* (2nd ed.). Thousand Oaks, California: Sage.
- Miller, M. D., Linn, R. L., & Gronlund, N. (2013). *Measurement and assessment in teaching* (11th ed.). New York: Pearson Education.
- Mills, G. E. (2014). *Action research: A guide for the teacher researcher* (5th ed.). Boston: Pearson Education.
- Mitchell, M. L., & Jolley, J. M. (2013). *Research design explained* (8th ed.). Belmont,

- California: Wadsworth Cengage Learning.
- Ministry of Education, Youth and Sport (MoEYS) (2010). *Educational strategic plan 2009-2013*. Phnom Penh: Publisher.
- Ministry of Education, Youth and Sport (MoEYS) (2012). *Higher education vision 2030*. Phnom Penh: Publisher.
- Molloy, E., & Boud, D. (2013). Changing conceptions of feedback.
In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 11-33). New York: Routledge.
- Molloy, E., & Boud, D. (2014). Feedback models for learning, teaching and performance.
In J.M. Spector, M.D. Merrill, J. Elen, & M.J. Bishop, (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 413-424). New York: Springer.
- Moore, S. H., & Bounchan, S. (2010). English in Cambodia: Changes and challenges.
World Englishes 29(1), 114-126.
- Moss, C. M., & Brookhart, S. M. (2012). *Learning targets: Helping students aim for understanding in today's lesson*. Alexandria, Virginia: Association for Supervision and Curriculum Development .
- Moss, C. M., Brookhart, S. M., & Long, B. A. (2013). Administrators' roles in helping teachers use formative assessment information. *Applied Measurement in Education* 26(3), 205-218.
- Mueller, R. O. (1996). *Basic principles of structural equation modelling: An introduction to LISREL and EQS*. New York: Springer Verlag.
- Munns, G., & Woodward, H. (2006). Student engagement and student self-assessment: The REAL framework. *Assessment in education* 13(2), 193-213.
- Myers, K. K., & Oetzel, J. G. (2003). Exploring the dimensions of organizational assimilation: Creating and validating a measure. *Communication Quarterly*, 51(4), 438-457.
- Nakabugo, M. G., Opolot-Okurut, C., Ssebbunga, C. M., Ngobi, D. H., Maani, J. S., Gumisiriza, E. L., Mbaga, R., Alupo, C., Byamugisha, A., Tukesiga, J., Bisikwa, R., Ndawula, R., & Bbosa, D. (2007). *Instructional strategies for large classes: Baseline literature and empirical study of primary school teachers in Uganda*. Retrieved from http://home.hiroshima-u.ac.jp/cice/publications/aa/Kampala_Uganda.pdf
- Neau, V. (2003). The teaching of foreign languages in Cambodia: A historical perspective. *Language, Culture and Curriculum*, 16(3), 253-268.

- Newmann, F. M., & Archbald, D. A. (1992). The nature of authentic academic achievement. In H. Berlak, F. M. Newmann, E. Adams, D. A. Archbald, T. Burgess, J. Raven & T. A. Romberg (Eds.), *Toward a new science of educational testing and assessment* (pp. 71-83). Albany, New York: State University of New York Press.
- Nicol, D. (2013). Resituating feedback from the reactive to the proactive. In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 34-49). New York: Routledge.
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199-218.
- Nolen, S. B., & Haladyna, T. (1990). Personal and environmental influences on students' beliefs about effective study strategies. *Contemporary Educational Psychology*, 15(2), 116-130.
- Nolen, S. B., Haladyna, T. M., & Haas, N. S. (1992). Uses and abuses of achievement test scores. *Educational Measurement: Issues and Practice*, 11(2), 9-15.
- Nostrand, H. L. (1989). Authentic texts and cultural authenticity: An editorial. *The Modern Language Journal*, 73(1), 49-52.
- Nunan, D. (1992). *Research methods in language learning*. New York: Cambridge University Press.
- Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
- Nuttall, D. L., & Skurnik, L. S. (1969). *Examination and item analysis manual*. London: National Foundation for Educational Research.
- O'Connor, K. (2009). *How to grade for learning, K-12* (3rd ed.). Thousand Oaks, California: Corwin.
- Oescher, J., & Kirby, P. C. (1990). Assessing teacher-made tests in secondary math and science classrooms. *Paper presented at the annual conference of the National Council on Measurement in Education*, April 17-19, Boston. (ERIC Document Reproduction Service No. ED 322 169).
- O'Leary, M. (2008). Towards an agenda for professional development in assessment. *Journal of In-service Education*, 34(1), 109-114.
- Olson, D. R. (2009). Literacy, literacy policy, and the school. In D. R. Olson & N. Torrance

- (Eds.), *The Cambridge handbook of literacy* (pp. 566-576). Cambridge: Cambridge University Press.
- Orr, S. (2008). Real or imagined?: The shift from norm referencing to criterion referencing in higher education. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 133-143). New York: Routledge.
- Orrell, J. (2008). Assessment beyond belief: The cognitive process of grading. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 251-263). New York: Routledge.
- Pajares, M. F. (1992). Teachers' beliefs and educational research: Cleaning up a messy construct. *Review of Educational Research*, 62(3), 307-332.
- Patton, M. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks: California: Sage.
- Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, D.C.: The National Academies Press.
- Pennycuik, D. (1991). Moderation of continuous assessment systems in developing countries. *Compare*, 21(2), 145-152.
- Plake, B. S. (1993). Teacher assessment literacy: Teachers' competencies in the educational assessment of students. *Mid-Western Educational Researcher*, 6(1), 21-27.
- Plake, B. S., & Impara, J. C. (1997). Teacher assessment literacy: What do teachers know about assessment? In G. D. Phye (Ed.), *Handbook of classroom assessment* (pp. 53-68). New York: Academic Press.
- Plake, B. S., Impara, J. C., & Fager, J. J. (1993). Assessment competencies of teachers: A national survey. *Educational Measurement: Issues and Practice*, 12(4), 10-12.
- Pond, K., Ul-Haq, R., & Wade, W. (1995). Peer review: A precursor to peer assessment. *Innovations in Education and Training International*, 32(4), 314-323.
- Popham, W. J. (1991). Appropriateness of teachers' test-preparation practices. *Educational Measurement: Issues and Practice*, 10(4), 12-15.
- Popham, J. W. (2006). Needed: A dose of assessment literacy. *Educational Leadership*, 63(6), 84-85.
- Popham, W. J. (2009). Assessment literacy for teachers: Faddish or fundamental? *Theory Into Practice*, 48, 4-11.

- Popham, W. J. (2011). Assessment literacy overlooked: A teacher educator's confession. *The Teacher Educator* 46(4), 265-273.
- Popham, W. J. (2014). *Classroom assessment: What teachers need to know* (7th ed.). Boston: Pearson Education.
- Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement 1, 2. *Journal of Educational Measurement*, 6(1), 1-9.
- Poth, C. (2013). What assessment knowledge and skills do initial teacher education programs address? A Western Canadian perspective. *Alberta Journal of Educational Research* 58(4), 634-656.
- Price, M. (2005). Assessment standards: The role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education* 30(3), 215-230.
- Price, M., Carroll, J., O'Donovan, B., & Rust, C. (2011). If I was going there I wouldn't start from here: A critical commentary on current assessment practice. *Assessment & Evaluation in Higher Education* 36(4), 479-492.
- Price, M., Rust, C., O'Donovan, B., Handley, K., & Bryant, R. (2012). *Assessment literacy: The foundation for improving student learning*. Oxford: The Oxford Centre for Staff and Learning Development.
- Quitter, S. M. (1999). Assessment literacy for teachers: Making a case for the study of test validity. *The Teacher Educator*, 34(4), 235-243.
- Quilter, S. M., & Gallini, J. K. (2000). Teachers' assessment literacy and attitudes. *The Teacher Educator*, 36(2), 115 - 131.
- Quinn, T. (2013). *On grades and grading: Supporting student learning through a more transparent and purposeful use of grades*. Lanham, Rowman & Littlefield Education.
- Raes, A., Vanderhoven, E., & Schellens, T. (2013). Increasing anonymity in peer assessment by using classroom response technology within face-to-face higher education. *Studies in Higher Education*, 1-16.
- Ramdass, D., & Zimmerman, B. J. (2008). Effects of self-correction strategy training on middle school students' self-efficacy, self-evaluation, and mathematics division learning. *Journal of Advanced Academics* 20(1), 18-41.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: The Danish Institute for Educational Research.

- Rath, C. (2010). *Designing a quality management system for a Cambodian University*. Unpublished doctoral thesis. The University of Technology, Sydney, Australia.
- Read, B., Francis, B., & Robson, J. (2005). Gender, 'bias', assessment and feedback: Analyzing the written assessment of undergraduate history essays. *Assessment & Evaluation in Higher Education* 30(3), 241-260.
- Rea-Dickins, P. (2000). Assessment in early years language learning contexts. *Language Testing*, 17(2), 115-122.
- Rea-Dickins, P. (2004). Understanding teachers as agents of assessment. *Language Testing*, 21(3), 249-258.
- Rea-Dickins, P. (2007). Classroom-based assessment: Possibilities and pitfalls. In J. Cummins, & C. Davison (Eds.), *International Handbook of English Language Teaching* (pp. 505-520). New York: Springer.
- Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitative study. *Teaching and Teacher Education*, 27(2), 472-482.
- Reynolds, C. R., Livingston, R. B., & Willson, V. (2009). *Measurement and assessment in education* (2nd ed.). Upper Saddle River, New Jersey: Person Education.
- Rieg, S. A. (2007). Classroom assessment strategies: What do students at-risk and teachers perceive as effective and useful? *Journal of Instructional Psychology* 34(4), 214-225.
- Rogers, T., Cheng, L., & Hu, H. (2007). ESL/ EFL instructors' beliefs about assessment and evaluation. *Canadian and International Education*, 36(1), 39-61.
- Rokeach, M. (1972). *Beliefs, attitudes and values: A theory of organization and change*. San Francisco: Jossey-Bass.
- Rom, M. C. (2011). Grading more accurately. *Journal of Political Science Education*, 7(2), 208-223.
- Ross, K. N. (1993). *Sample design for international studies of educational achievement*. International Institute for Educational Planning Annual Training Programme Module on Monitoring and Evaluating Educational Outcomes. Paris, France: UNESCO.
- Ross, K. N. (2005). *Sample design for educational survey research: Module 3*. Paris: International Institute for Educational Planning (UNESCO). Retrieved from <http://www.unesco.org/iiep>.
- Rudner, L. M. (1992). Reducing errors due to the use of judges. *Practical Assessment*,

- Research & Evaluation*, 3(3). Retrieved from <http://pareonline.net/getvn.asp?v=3&n=3>
- Ruiz-Primo, M. A., & Li, M. (2013). Analyzing teachers' feedback practices in response to students' work in science classrooms. *Applied Measurement in Education* 26(3), 163-175.
- Russell, M. K., & Airasian, P. W. (2012). *Classroom assessment: Concepts and applications* (7th ed.). New York: McGraw-Hill.
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119-144.
- Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education* 30(2), 175-194.
- Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education: Principles, Policy & Practice* 14(3), 387-392.
- Sadler, D. R. (2009a). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education*, 34(2), 159-179.
- Sadler, D. R. (2009b). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, 34(7), 807-826.
- Sadler, D. R. (2009c). Transforming holistic assessment and grading into a vehicle for complex learning. In G. Joughin (Ed.), *Assessment, learning and judgement in higher education* (pp. 45-63). New York: Springer.
- Sadler, D. R. (2010). Educational assessment- assessment in domains. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.) (pp. 249-255). Oxford: Elsevier Science.
- Sadler, D. R. (2013a). Assuring academic achievement standards: From moderation to calibration. *Assessment in Education: Principles, Policy & Practice*, 20(1), 5-19.
- Sadler, D. R. (2013b). Opening up feedback: Teaching learners to see. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 54-63). New York: Routledge.
- Sadler, P. M., & Good, E. (2006). The impact of self-and peer-grading on student learning. *Educational Assessment* 11(1), 1-31.
- Sambell, K. (2013). Involving students in the scholarship of assessment: Student voices on the feedback agenda for change. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students*

- (pp. 80-91). New York: Routledge.
- Sambell, K., McDowell, L., & Montgomery, C. (2013). *Assessment for learning in higher education*. New York: Routledge.
- Saravanamuthu, K. (2008). Reflecting on the Biggs–Watkins theory of the Chinese learner. *Critical Perspectives on Accounting* 19(2), 138-180.
- Sato, M., Wei, R. C., & Darling-Hammond, L. (2008). Improving teachers' assessment practices through professional development: The case of National Board Certification. *American Educational Research Journal*, 45(3), 669-700.
- Scarino, A. (2013). Language assessment literacy as self-awareness: Understanding the role of interpretation in assessment and in teacher learning. *Language Testing* 30(3), 309-327.
- Schafer, W. D. (1991). Essential assessment skills in professional education of teachers. *Educational Measurement: Issues and Practice*, 10, 3-6.
- Schafer, W., & Lissitz, R. (1987). Measurement training for school personnel: Recommendations and reality. *Journal of Teacher Education*, 38(3), 57-63.
- Schaff, C. S. (2006). *A Rasch model analysis of the standards for teacher competence in educational assessment of students using the Classroom Assessment Practices Inventory*. Unpublished doctoral thesis. Northern Illinois University.
- Schimmer, T. (2014). *Ten things that matter from assessment to grading*. Boston: Pearson Education.
- Schlenker, B. R. (2012). Self-presentation. In M. R. Leary & J. P. Tangney (Eds.), *Handbook of self and identity* (2nd) (pp. 542-570). New York: Guilford.
- Schlenker, B. R., & Weigold, M. F. (1989). Goals and the self-identification process: Constructing desired identities. In L. A. Pervin (Ed), *Goal concepts in personality and social psychology*, (pp. 243-290). Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Schmidt, K. M., & Embretson, S. E. (2013). Item response theory and measuring abilities. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology, volume 2* (2nd) (pp. 451-473). Hoboken, New Jersey: Wiley.
- Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological assessment*, 8(4), 350-353.
- Schneider, M. C., & Gowan, P. (2013). Investigating teachers' skills in interpreting evidence of student learning. *Applied Measurement in Education* 26(3), 191-204.

- Schumacker, R. E., & Lomax, R. G. (2010). *A beginner's guide to structural equation modelling* (3rd ed.). New York: Routledge.
- Schunk, D. H. (1996). Goal and self-evaluative influences during children's cognitive skill learning. *American Educational Research Journal* 33(2), 359-382.
- Schunk, D. H., Meece, J. L., & Pintrich, P. R. (2014). *Motivation in education: Theory, research, and applications* (4th ed.). Boston: Pearson Education.
- Schunk, D. H., & Pajares, F. (2004). Self-efficacy in education revisited. In D. M. Melnerney & S. V. Etten (Eds.), *Big theories revisited* (vol. 4, pp.115-138). Greenwich, Conn: Information Age.
- Schunk, D. H., & Pajares, F. (2009). Self-efficacy theory. In K. Wentzel & A. Wigfield (Eds.), *Handbook of motivation at school* (pp. 35-53). New York: Routledge.
- Schwager, M. T., & Carlson, J. S. (1994). Building assessment cultures: Teacher perceptions and school environments. *Education and Urban Society*, 26, 390-403.
- Scott, C. (2007). Stakeholder perceptions of test impact. *Assessment in Education*, 14(1), 27-49.
- Scouller, K. (1998). The influence of assessment methods on students' learning approaches: Multiple-choice question examination versus assignment essay. *Higher Education*, 35, 453-472.
- Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R. M. Gagné, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago: Rand McNally.
- Segers, M., & Dochy, F. (2001). New assessment forms in problem-based learning: The value-added of the students' perspective. *Studies in Higher Education*, 26(3), 327-343.
- Selke, M. J. G. (2013). *Rubric assessment goes to college: Objective, comprehensive evaluation of student work*. Lanham, Maryland: Rowman & Littlefield.
- Shepard, L.A. (2000). The role of assessment in a learning culture. *Educational researcher*, 29(7), 4-14.
- Shepard, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping, teaching and learning* (pp. 279-303). New York: Lawrence Erlbaum Associates.
- Shepard, L. A. (2013). Foreword. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. xix-xxii). Los Angeles: Sage.

- Shepard, L., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., Gutierrez, C., & Pacheco, A. (2005). Assessment. In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able to do* (pp. 275-326). New York: Jossey-Bass.
- Shohamy, E. (2001). Democratic assessment as an alternative. *Language Testing*, 18(4), 373-391.
- Shohamy, E., Donitsa-Schmidt, S., & Ferman, I. (1996). Testing impact revisited: Washback effect over time. *Language Testing*, 13(3), 298-317.
- Shute, V. J. (2008). Focus on formative feedback. *Review of educational research*, 78(1), 153-189.
- Shute, V. J., & Kim, Y. J. (2014). Formative and stealth assessment. In J.M. Spector, M.D. Merrill, J. Elen, & M.J. Bishop, (Eds.), *Handbook of Research on Educational Communications and Technology* (pp. 311-321). New York: Springer.
- Silis, G., & Izard, J. F. (2002). Monitoring student learning using assessment data. *Paper presented at the AARE Annual Conference*, December, Brisbane.
- Silverman, D. (2013). *Doing qualitative research* (4th ed.). London: Sage.
- Simon, E., & Pleschová, G. (2013). Creating successful teacher development programmes. In E. Simon & G. Pleschová (Eds.), *Teacher development in higher education: Existing programs, program impact, and future trends* (pp. 274-297). New York: Routledge.
- Sluijsmans, D., Dochy, F., & Moerkerke, G. (1999). Creating a learning environment by using self-, peer-, and co-assessment. *Learning Environment Research*, 1, 293-319.
- Smaill, E. (2013). Moderating New Zealand's National Standards: Teacher learning and assessment outcomes. *Assessment in Education: Principles, Policy & Practice* 20(3), 250-265.
- Smith, C. (2012). Why should we bother with assessment moderation? *Nurse education today*, 32, 45-48.
- Spear, M. G. (1984). Sex bias in science teachers' ratings of work and pupil characteristics. *European Journal of Science Education* 6(4), 369-377.
- Spearritt, D. (Ed.) (1982). *The improvement of measurement in education and psychology: Contributions of latent trait theories*. Melbourne, Australian Council for Educational Research (ACER).

- Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing*, 2(1), 31-40.
- Spolsky, B. (2014). The Influence of ethics in language assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (1st ed., Vol. 3), (pp. 1571-1585). Oxford: John Wiley and Sons.
- Stiggins, R. J. (1991a). Assessment literacy. *Phi Delta Kappan*, 72(7), 534-539.
- Stiggins, R. J. (1991b). Relevant classroom assessment training for teachers. *Educational Measurement: Issues and Practice*, 10(1), 7-12.
- Stiggins, R. J. (1995). Assessment literacy for the 21st century. *Phi Delta Kappan*, 77(3), 238-245.
- Stiggins, R. J. (1999). Evaluating classroom assessment training in teacher education programs. *Educational Measurement: Issues and Practice*, 18(1), 23-27.
- Stiggins, R. (2008). Correcting "errors of measurement" that sabotage student learning. In C. A. Dwyer (Ed.), *The future of assessment: Shaping, teaching and learning* (pp. 229-243). New York: Lawrence Erlbaum Associates.
- Stiggins, R. (2010). Essential formative assessment competencies for teachers and school leaders. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 233-250). New York: Routledge.
- Stiggins, R. J., Frisbie, D. A., & Griswold, P. A. (1989). Inside high school grading practices: Building a research agenda. *Educational Measurement: Issues and Practice*, 8(2), 5-14.
- Stobart, G., & Gipps, C. (2010). Alternative assessment. In P. Peterson, E. Baker, & B. McGaw (Eds.), *International encyclopedia of education* (3rd ed.) (pp. 202-208). Oxford: Elsevier Science.
- Street, B. (2009). Ethnography of writing and reading. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 329-345). Cambridge: Cambridge University Press.
- Struyven, K., Dochy, F., & Janssens, S. (2005). Students' perceptions about evaluation and Assessment in higher education: A review. *Assessment and Evaluation in Higher Education*, 30(4), 325-341.
- Sun, Y., & Cheng, L. (2013). Teachers' grading practices: Meaning and values assigned. *Assessment in Education: Principles, Policy & Practice*, 1-18.
- Suos, M. (1996). Pre- and in-service training of teachers: The bachelor of education (teaching

- English as foreign language) degree course at UPP. *Paper presented at the Seminar on English Language Teaching in Cambodia*, May 27-29, Phnom Penh, Cambodia.
- Swann, J., Andrews, I., & Ecclestone, K. (2011). Rolling out and scaling up: The effects of a problem-based approach to developing teachers' assessment practice. *Educational Action Research* 19(4), 531-547.
- Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Boston: Pearson Education.
- Tang, K. C. C. (1994). Assessment and student learning: Effects of modes of assessment on students' preparation strategies. In G. Gibbs (Ed.), *Improving student learning: Theory and practice* (pp. 151-170). Oxford: The Oxford Centre for Staff Development.
- Tao, N. (2007). Foreword. In N. Tao (Ed.), *CamTESOL conference on English language teaching: Selected papers 3* (pp. iii-iv). Phnom Penh: CamTESOL.
- Tao, N. (2012). An investigation into assessment practices of EFL university lecturers: Purposes, methods, and procedure. *PowerPoint presentation at the 8th Annual CamTESOL Conference on English Language Teaching*, February 25-26, Phnom Penh, Cambodia.
- Taras, M. (2005). Assessment- summative and formative- some theoretical reflections. *British Journal of Educational Studies*, 53(4), 466-478.
- Taras, M. (2013). Feedback on feedback: Uncrossing wires across sectors. In S. Merry, M. Price, D. Carless, & M. Taras (Eds.), *Reconceptualising feedback in higher education: Developing dialogue with students* (pp. 30-40). New York: Routledge.
- Tashakkori, A., Teddlie, C., & Sines, M. C. (2013). Utilizing mixed methods in psychological research. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology, volume 2* (2nd) (pp. 428-450). Hoboken, New Jersey: Wiley.
- Taylor, C. S., & Nolen, S. B. (2008). *Classroom assessment: Supporting teaching and learning in real classrooms* (2nd ed.). Upper Saddle River, New Jersey: Pearson Education.
- Taylor, L. (2009). Developing assessment literacy. *Annual Review of Applied Linguistics*, 29, 21-36.
- Taylor, L. (2013). Communicating the theory, practice and principles of language testing to test stakeholders: Some reflections. *Language Testing* 30(3), 403-412.
- Teddlie, C., & Tashakkori, A. (2009). *Foundations of mixed methods research*. Thousand Oaks, California: Sage.

- The Department of Cambodian Higher Education (2009). *Statistics of students in academic year 2008-2009*. Phnom Penh: Publisher.
- Thomas, P. R., & Bain, J. D. (1982). Consistency in learning strategies. *Higher Education*, 11, 249-259.
- Thomas, P. R., & Bain, J. D. (1984). Contextual dependence of learning approaches: The effects of assessments. *Human Learning*, 3, 227-240.
- Tierney, R. D. (2013). Fairness in classroom assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 125-144). Los Angeles: Sage.
- Timperley, H. (2013). Feedback. In J. Hattie & E. M. Anderman (Eds.), *International guide to student achievement* (pp. 402-404). New York: Routledge.
- Topping, K. J. (1998). Peer assessment between students in colleges and universities. *Review of educational research* 68(3), 249-276.
- Topping, K. J. (2013). Peers as a source of formative and summative assessment. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 395-412). Los Angeles: Sage.
- Tourangeau, R., & Yan, T. (2007). Sensitive questions in surveys. *Psychological bulletin* 133(5), 859-883.
- Towndrow, P. A., Tan, A. L., Yung, B. H. W., & Cohen, L. (2010). Science teachers' professional development and changes in science practical assessment practices: What are the issues? *Research in Science Education*, 40(2), 117-132.
- Tran, T. T. (2013a). Is the learning approach of students from the Confucian heritage culture problematic? *Educational Research for Policy and Practice* 12(1), 57-65.
- Tran, T. T. (2013b). Limitation on the development of skills in higher education in Vietnam. *Higher Education* 65(5), 631-644.
- Tsagari, D. (2008). Assessment literacy of EFL teachers in Greece: Current trends and future prospects. *PowerPoint presentation at the 5th Annual EALTA Conference*, May 9-11, Athens, Greece.
- Tsagari, D. (2009). *The complexity of test washback: An empirical study*. New York: Peter Lang.
- Tudge, J. R. H., Doucet, F., Otero, D., Sperb, T. M., Piccinni, C. A., & Lopes, R. S. (2006). A window into different cultural worlds: Young children's everyday activities in the United States, Brazil, and Kenya. *Child development*, 77(5), 1446-1469.

- Ullman, J. B., & Bentler, P. M. (2013). Structural equation modelling. In I. B. Weiner, J. A. Schinka, & W. F. Velicer (Eds.), *Handbook of psychology: Research methods in psychology, volume 2* (2nd) (pp. 661-690). Hoboken, New Jersey: Wiley.
- Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Cambridge: Harvard University Press.
- Wagner, D. A. (2009). New technologies for adult literacy and international development. In D. R. Olson & N. Torrance (Eds.), *The Cambridge handbook of literacy* (pp. 548-565). Cambridge: Cambridge University Press.
- Wall, D (2012). Washback. In G. Fulcher & F. Davidson (Eds.). *The Routledge handbook of language testing* (pp. 79-92). New York: Routledge.
- Wall, D., & Alderson, J. C. (1993). Examining washback: The Sri Lankan impact study. *Language Testing*, 10(1), 41-69.
- Walters, F. S. (2010). Cultivating assessment literacy: Standards evaluation through language-test specification reverse engineering. *Language Assessment Quarterly*, 7(4), 317-342.
- Waugh, C. K., & Gronlund, N. E. (2013). *Assessment of student achievement* (10th ed.). Boston: Pearson Education.
- Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge: Cambridge university press.
- Wiliam, D. (2008). Balancing dilemmas: Traditional theories and new applications. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 267-281). New York: Routledge.
- Wiliam, D. (2010). An integrative summary of the research literature and implications for a new theory of formative assessment. In H. L. Andrade & G. J. Cizek (Eds.), *Handbook of formative assessment* (pp. 18-40). New York: Routledge.
- Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, 37, 3-14.
- Wiliam, D. (2013). Feedback and instructional correctives. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 197-214). Los Angeles: Sage.
- Wiliam, D., & Black, P. (1996). Meanings and consequences: A basic for distinguishing formative and summative functions of assessment. *British Educational Research Journal*, 22(5), 537-548.

- Wise, S. L., Lukin, L., & Roos, L. L. (1991). Teacher beliefs about training in testing and measurement. *Journal of teacher Education*, 42, 37-42.
- Witte, R. H. (2012). *Classroom assessment for teachers*. New York: McGraw-Hill.
- Wolf, D., Bixby, J., Glenn, J., & Gardner, H. (1991). To use their minds well: Investigation new forms of student assessment. *Review of Research in Education*, 17, 31-74.
- Wormeli, R. (2006). *Fair isn't always equal: Assessing & grading in the differentiated classroom*. Portland, Maine: Stenhouse.
- 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., & Panchapakesan, N. (1969). A procedure for sample-free item analysis. *Educational and Psychological measurement*, 29(1), 23-48.
- Wright, B., & Stone, M. H. (1979). *Best test design*. Chicago: MESA Press.
- Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.
- Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modelling software*.
- Wyatt-Smith, C., & Klenowski, V. (2013). Explicit, latent and meta-criteria: Types of criteria at play in professional judgement practice. *Assessment in Education: Principles, Policy & Practice*, 20(1), 35-52.
- Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, 17(1), 59-75.
- Xu, Y., & Liu, Y. (2009). Teacher assessment knowledge and practice: A narrative inquiry of Chinese college EFL teacher's experience. *TESOL Quarterly*, 43(3), 493-513.
- Yin, R. K. (2009). *Case study research: Design and methods* (4th ed.). Los Angeles, California: Sage.
- Zhang, Z. (1996). Teacher assessment competency: A Rasch model analysis. *Paper presented at the Annual Meeting of the American Educational Research Association*, April 8-12, New York. (ERIC Reproduction Service No. ED 400 322).

- Zhang, Z., & Burry-Stock, J. A. (1994). *Assessment practices inventory*. Tuscaloosa, AL: The University of Alabama.
- Zhang, Z., & Burry-Stock, J. A. (1997). Assessment practices inventory: A multivariate analysis of teachers perceived assessment competency. *Paper presented at the Annual Meeting of the National Council on Measurement in Education*, March 25-27, Chicago, Illinois. (ERIC Reproduction Service No. ED 408 333).
- Zhang, Z., & Burry-Stock, J. A. (2003). Classroom assessment practices and teachers' self-perceived assessment skills. *Applied Measurement in Education*, 16(4), 323-342.
- Zimmerman, B. J. (1995). Self-efficacy and educational development. In A. Bandura (Ed.), *Self-efficacy in changing societies* (pp. 202-231). New York: Cambridge University Press.
- Zimmerman, B. J., & Cleary, T. J. (2006). Adolescents' development of personal agency: The role of self-efficacy beliefs and self-regulatory skill. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 45-69). Greenwich, Conn: Information Age.
- Zimny, S. T., Robertson, D. U., Bartoszek, T. (2008). Academic and personal dishonesty in college students. *North American Journal of Psychology*, 10(2), 291-312.
- Zwozdiak-Myers, P. (2012). *The teacher's reflective practice handbook: Becoming an extended professional through capturing evidence-informed practice*. New York: Routledge.

Appendices

Appendix A: Classroom Assessment Knowledge Test

DIRECTIONS

*The following items are examining your knowledge in the educational assessment of students. Please read each scenario followed by each item carefully and answer each of the items by **circling** the response you think is the best one. **Even if you are not sure of your choice, circle the response you believe to be the best. Do not leave any items unanswered.***

Scenario # 1

Mr. Chan Sambath, a first year English writing lecturer, is aware of the fact that his students will be taking a semester examination at the end of the course.

1. Mr. Chan Sambath wants to assess his students' critical thinking abilities at the end of the unit to determine if any reinstruction will be necessary prior to the exam. Which of the following methods would be the **most** appropriate choice?
 - A. multiple-choice items
 - B. matching items
 - C. gap-filling items
 - D. essay writing
2. In order to grade his students' writing accurately and consistently, Mr. Chan Sambath would be **well** advised to
 - A. identify criteria from the unit objectives and create a marking criteria.
 - B. develop a marking criteria after getting a feel for what students can do.
 - C. consider student performance on similar types of tests.
 - D. consult with experienced colleagues about a marking criteria that has been used in the past.
3. Mr. Chan Sambath wants to evaluate his students' understanding of specific aspects of their responses. Which of the following would **best** facilitate him scoring of these responses?
 - A. an objective answer key
 - B. an holistic scoring
 - C. a checklist
 - D. an analytic scoring

4. At the end of each class period, Mr. Chan Sambath asks his students several questions to get an impression of their understanding. In this example, the **primary** purpose for conducting formative assessment is to
 - A. determine the final grades for students.
 - B. determine content for the final examination.
 - C. identify individual learning needs to plan classroom instruction.
 - D. evaluate curriculum appropriateness.
5. Which grading practice being considered by Mr. Chan Sambath would result in grades that would **most** reflect his students' learning achievement against the learning outcomes?
 - A. grades based on the students' performances on a range of assessments
 - B. grades based on the amount of time and effort the student spent on the assessments
 - C. grades based on how the student has performed in comparison to his/her classmates
 - D. grades based upon the personal expectations of Mr. Chan Sambath
6. Mr. Chan Sambath is planning to keep assessment records as a part of his assessment and reporting process. Which of the following is the **least** important assessment information to be recorded?
 - A. statistical data including marks, student welfare and biographical information.
 - B. anecdotal data comprising critical incidents or reflections of both Mr. Chan Sambath and his students.
 - C. all copies of his students' assessment work.
 - D. a representative sample of each student work.
7. In a routine conference with his students, Mr. Chan Sambath is asked to explain the **basis** for assigning his course grade. Mr. Chan Sambath should
 - A. explain that the grading system was imposed by the school administrators.
 - B. refer to the information that he presented to his students at the beginning of the course on the assessment process.
 - C. re-explain the students the way in which the grade was determined and show them samples of their work.
 - D. indicate that the grading system is imposed by the Ministry of Education.
8. Mr. Chan Sambath was worried that his students would not perform well on the semester examination. He did all of the following to help increase his students' scores. Which was **unethical**?
 - A. He instructed his students in strategies for taking tests.
 - B. He planned his instruction so that it focused on concepts and skills to be covered on the test.
 - C. He allowed his students to bring in their coursebooks/materials to refer to during the test.
 - D. He allowed students to practice with a small number of items from the actual test.

9. To ensure the validity and reliability of his classroom assessment procedure, it is advised that Mr. Chan Sambath should gather together with his colleagues to discuss all of the following **except**
- A. marking criteria.
 - B. students' pieces of work.
 - C. teaching techniques.
 - D. assessment activities.

Scenario # 2

Ms. Chan Tevy is a year two English lecturer. She has just finished teaching a unit on the Industrial Revolution and wishes to measure her students' understanding of this particular unit using a multiple-choice test.

10. Based on her goal, which of the following assessment strategies would be the **most** appropriate choice?
- A. She should use the test items included in the teacher's manual from the textbook she uses.
 - B. She should design test items which are consistent with the content and skill specified in the course learning outcomes.
 - C. She should use available test items from internet that cover Industrial Revolution.
 - D. She should design test items which cover the factual information she taught.
11. In constructing her multiple-choice test items, Ms. Chan Tevy should follow all of the following guidelines **except**
- A. ensure that the correct response is unequivocally the best.
 - B. ensure that the responses to a given item are in different literary forms.
 - C. ensure the stem and any response, taken together, read grammatically.
 - D. make all distracters plausible and attractive to the ignorant test-taker.
12. Ms. Chan Tevy decides to score the tests using a 100% correct scale. Generally speaking, what is the **proper** interpretation of a student score of 85 on this scale?
- A. The student answered 85% of the items on the test correctly.
 - B. The student knows 85% of the content covered by this instructional unit.
 - C. The student scored higher than 85% of other students who took this test.
 - D. The student scored lower than 85% of other students who took this test.

13. Some of Ms. Chan Tevy's students do not score well on the multiple-choice test. She decides that the next time she teaches this unit, she will begin by administering a pretest to check for students' prerequisite knowledge. She will then adjust her instruction based on the pretest results. What **type** of information is Ms. Chan Tevy using?
- A. norm-referenced information (describes each student's performance relative to the other students in a group such as percentile ranks)
 - B. criterion-referenced information (describes each student's performance in terms of status in specific learning outcomes)
 - C. both norm- and criterion-referenced information
 - D. neither norm- nor criterion-referenced information
14. The Industrial Revolution test is the only student work that Ms. Chan Tevy grades for the current grading period. Therefore, grades are assigned only on the basis of the test. Which of the following is **not** a criticism of this practice?
- A. The test, and therefore the grades, reflect too narrow a curriculum focus.
 - B. These grades, since based on test alone, are probably biased against some minority students.
 - C. Tests administered under supervised conditions are more reliable than those assessments undertaken in less standardized conditions (e.g. homework)
 - D. Decisions like grades should be based on more than one piece of information.
15. Ms. Chan Tevy fully understands that her classroom assessment records serve all of the following purposes **except**
- A. provide information regarding assessment methods development.
 - B. provide diagnostic information to show the strengths and weaknesses of student performance.
 - C. show the extent of student progress.
 - D. provide information to assist administrative decision makers.
16. During an individual conference, one student in Ms. Chan Tevy's class wants to know what it means that he scored in the 80th percentile in a multiple-choice test. Which of the following provides the **best** explanation of this student's score?
- A. He got 80 % of the items on the test correct.
 - B. He is likely to earn a grade of "B" in his class.
 - C. He is demonstrating above grade level performance.
 - D. He scored the same or better than 80 % of his classmates.
17. Based on their grades from last semester, Ms. Chan Tevy believes that some of her low-scoring students are brighter than their test scores indicate. Based on this knowledge, she decides to add some points to their test scores, thus raising their grades. Which of Ms. Chan Tevy's action was **unethical**?
- A. examining her student's previous academic performance
 - B. adjusting grades in her course
 - C. using previous grades to adjust current grades
 - D. adjusting some students' grades and not others'

18. To enhance the quality of a new developed multiple-choice test, Ms. Chan Tevy should do all of the following **except**
- A. pilot the test items with a small number of her past students to see how well each item performs.
 - B. make all necessary changes to the test items based on the information received during her pilot.
 - C. have all of her current students undertake the test twice and make a comparison of their scores.
 - D. panel the test items through consultation with her colleagues who have assessment experience.

Scenario # 3

Mr. Peo Virak is a senior English lecturer in the Indrak Tevy University. Experienced in issues of classroom assessment, Mr. Peo Virak is often asked to respond to the questions concerning best practices for evaluating student learning.

19. Ms. Meas Chakriya, an English lecturer, asks what type of assessment is best to determine how well her students are able to apply what they have learned in class to a situation encountered in their everyday lives. The type of assessment that would **best** answer her question is called
- A. diagnostic assessment.
 - B. performance assessment.
 - C. formative assessment.
 - D. authentic assessment.
20. Ms. Keo Bopha is constructing essay questions for a test to measure her students' critical thinking skills. She consults with Mr. Peo Virak to see what concerns she would be aware of when constructing the questions. Which statement is **not** an appropriate recommendation when writing essay questions?
- A. consider the relevance of the questions for a particular group of her students
 - B. avoid determining the amount of freedom of writing responses that will be accepted
 - C. indicate the time limits for the writing responses
 - D. be clear about the skills require to be demonstrated

21. Chenda, a student in Mr. Peo Virak's class, scored 78 marks on a reading test which has a mean of 80 and a standard deviation of 4. She scored 60 marks on the writing test which had a mean of 50 and a standard deviation of 3. Based on the above information, in comparison to her peers, which statement provides the **most** accurate interpretation?
- A. Chenda is better in reading than in writing.
 - B. Chenda is better in writing than in reading.
 - C. Chenda is below average in both subjects.
 - D. Chenda is close to average in both subjects.
22. After teaching four units from his course book, Mr. Peo Virak gives his students a test to measure their learning achievement. In this example, the **primary** purpose for conducting summative assessment is to
- A. identify individual learning needs to plan classroom instruction.
 - B. motivate students to learn.
 - C. evaluate curriculum appropriateness.
 - D. determine the final grades for students.
23. Throughout instruction, Mr. Keo Ratana assesses how well his students are grasping the material. These assessments range from giving short quizzes, mid-term tests, written assignments to administering a semester examination. In order to improve the **validity** of this grading procedure, what advice should Mr. Peo Virak give to Mr. Keo Ratana?
- A. consider students' class participation and their attendance before assigning a final grade.
 - B. consider students' performance in other subjects before assigning a final grade.
 - C. weight assessments according to their relative importance.
 - D. take into consideration each student's effort when calculating grades.
24. Ms. Meas Chakriya consults with Mr. Peo Virak for advice to effectively use her observations in recording her students' activities in the classroom. Which statement is **not** an appropriate recommendation when observing her students' behaviors?
- A. make a record of the incident as soon after the observation as possible
 - B. maintain separate records of the factual description of the incident and her interpretation of the event
 - C. observe as many incidents in one long observation as possible
 - D. record both positive and negative behavioral incidents

25. Bora is a student in Mr. Keo Ratana's class. He receives a raw score of 12 items answered correctly out of a possible 15 on the vocabulary section of a test. This raw score equates to a percentile rank of 45. He is confused about how he could answer so many items correctly, but receive such a low percentile rank. He approaches Mr. Keo Ratana for a possible explanation. Which of the following is the **appropriate** explanation to offer to Bora?
- A. "I don't know...there must be something wrong with the way the test is scored."
 - B. "Although he answered 12 correctly, numerous students answered more than 12 correctly."
 - C. "Raw scores are purely criterion-referenced and percentile ranks are merely one form of norm-referenced scoring."
 - D. "Raw scores are purely norm-referenced and percentile ranks are merely one form of criterion-referenced scoring."
26. Prior to the semester examination, Mr. Keo Ratana reveals some information to his students. Which of Mr. Keo Ratana's action was **unethical**?
- A. inform his students the exam contents to be covered.
 - B. inform his students the exam methods to be used.
 - C. show the actual exam paper to a small group of his low-achieving students.
 - D. tell his students the exam duration.
27. To achieve quality management of classroom assessments, Mr. Peo Virak advises his colleagues to be involved in all of the following **except**
- A. quality assurance (concerning with quality of assessment by emphasising the assessment process).
 - B. quality teaching (dealing with the effectiveness of teaching in helping students undertake assessments successfully).
 - C. quality control (dealing with monitoring and, where necessary making adjustment to assessor judgments before results are finalised).
 - D. quality review (focusing on the review of the assessment results and processes in order to make recommendations for future improvement).

End of Test
Thank you for your kind help.

Appendix B: Questionnaire

I. Background Information

Please complete the following information about your background.

1. Age _____ years
2. Gender ☐ Male ☐ Female
3. You are currently teaching in ☐ English-major ☐ English non-major department.

4.

The degree you have attained (Tick as many that apply)	The discipline best describes your degree (Tick as many that apply)				
	Education	Law	Business	Politics	Other
Bachelor <input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Master <input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Doctoral <input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

5. Approximately, how many years have you been teaching English at university level?
_____ years
6. How many hours have you taught per week this semester? (**If you teach at another place, please also include those teaching hours**). _____ hours
7. On average, how many students are typically in one of your classes? _____ students
8. Have you undertaken any formal studies in the field of assessment during your undergraduate teacher preparation program? (**If you tick No, please skip questions 8a & 8b**)
☐ Yes ☐ No
- 8a. If yes, how long did the assessment course run for?
☐ Less than 1 semester ☐ 1 semester ☐ 2 semesters ☐ Over 2 semesters
- 8b. How well did your undergraduate teacher preparation program prepare you for designing and conducting classroom-based assessment?
☐ Very unprepared ☐ Unprepared ☐ Prepared ☐ Very prepared

II. Personal Beliefs about Assessment

<p>DIRECTIONS <i>The following items are examining your personal beliefs/attitudes toward assessments within EFL programme. To complete this questionnaire, please read each item carefully and answer each of the items by circling the response that best relates to you. There is no right or wrong answers, so please attempt to answer each statement accurately and honestly. Do not leave any items unanswered.</i></p>	<p>Key 1= not useful at all 2= a little useful 3= useful 4= very useful</p>
<p>For example: To what extent do you think it is useful to provide your students with the correct answers to the test questions when returning the test results? (If you think it is useful, circle 3).</p>	<p>1 2 3 4</p>
<p>1. To what extent is each of the following assessment types/methods useful in assessing students' learning?</p>	
<ul style="list-style-type: none">• Oral presentation	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Reflective journal	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Individual conference (a face-to-face discussion about a particular piece of your student's work with you)	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Self-assessment (students' work being assessed by themselves)	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Peer assessment (students' work being assessed by their classmates)	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Individual assignment/ project work	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Portfolio (a collection of students' work being assessed by you at the end of a semester)	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Assessments that resemble the English language use in your students' real life situations	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Assessments that provide regular feedback indicating the ways to improve your students' future performance	<p>1 2 3 4</p>
<p>2. Which of the following characteristics of your students influence you when marking their work (i.e., essays/assignments/presentations)?</p>	<p>Key 1= never 2= sometimes 3= often 4= always</p>
<ul style="list-style-type: none">• Gender	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Age	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Appearance	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Behaviour	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Attitude	<p>1 2 3 4</p>
<ul style="list-style-type: none">• General abilities	<p>1 2 3 4</p>
<ul style="list-style-type: none">• Effort	<p>1 2 3 4</p>

3. To what extent do you agree with each of the following assessment quality procedures?	Key 1= strongly disagree 2= disagree 3= agree 4= strongly agree			
• It is my responsibility to ensure that my assessments are valid and reliable before using them.	1	2	3	4
• It is important to provide students with their assessment results in a timely and effective way.	1	2	3	4
• It is important for instructors to gather together regularly to design and check the quality of the assessment process and results.	1	2	3	4
• It is just as important to maintain detailed records of the assessment process as it is to maintain records of students' results.	1	2	3	4
• It is important to employ various methods to record students' achievement.	1	2	3	4
• It is important to construct accurate reports about students' achievement for communicating to both students and administrators.	1	2	3	4

End of Questionnaire
Thank you for your kind help.

Appendix C: Tables of Matrices

I. English-major Instructors

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LM1	LM2	LM3
Perceived assessment competence		<p>“Okay let [’s] say 5 or 6...what I have learned from Teaching Methodology, it was more than 10 years ago...I have learned a lot from my former colleagues and now they [have] moved...to other [work] places. And now we have...new generations, the gap between the young generation and the senior generation...becomes, you know bigger and bigger. And team work...you know we have less time to share our experience with each other...I think I should have been trained more often”</p>	<p>“I would give myself a 10...I think to be sufficient in doing any kind of testing and assessments is to say that a person can independently research their [his] daily practice with a student and required the ability to collect information related to the use of different tests and assessments devices in the classroom and it required reflection from that also...And I am able to do it”</p>	<p>“May be [I scored myself] 6 to 7 the most...Because after the tests, most of the time I just feel sometimes students are doing well in my class or read a lot, but when I ask them to answer questions, and then some of them cannot answer because...some questions are too detailed on the factual information in the [course] books, not all the questions are about what they can reflect...there must be a way...[I can] learn to do better next time...If we [I] don’t really know how to make it better and then you [I] will keep doing the same thing”</p>
Notion of the ideal assessment		<p>“Assessment [test/exam] has to reflect what the teachers have taught...not too much on memory or may be not too much on...details or facts...questions have to be about critical thinking skills...Cambodia lacks...people who are creative, so thinking skills, okay, critical thinking skill is very important for students in Cambodia”</p>	<p>“I believe in what they call a portfolio - type of assessment which is very rigorous, which takes a lot of time, and which takes a lot of effort from the teachers also in giving feedback...Yes, it’s not summative, it’s ongoing I mean you don’t assess students once and then make generalization about students’ ability by just using one time assessment or test...and it’s more reliable because you have a lot of time to cross-check students’ progress throughout the semester...you get to know students better and you can help students better also”</p>	<p>“To tell you the truth, I don’t think teachers assess students’ speaking, well, in CE, Core English classes. There’s no session for students, for teachers to test, and that would be the case. So, we [I] focus more on the area that we [I] ignore so far...the way that they debate, the way that they appear in the meeting, so if we [I] can create [assessments]...like that I think it will be more purposive and more meaningful than just writing the answers to [test/exam] questions all the time...it’s the only way to show how much the students can do”</p>

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LM1	LM2	LM3
Knowledge and understanding of the concepts of validity & reliability	Definition of the concepts of validity & reliability	“It [reliability] reflects the results, the outcomes that the students have learned...Validity, it’s valid. I don’t remember the terms. I mean I don’t remember these technical words”	“The test itself can be used to test what we claim to test...Yes, if you use it to measure something else, then it does not measure what it claims to measure. And therefore it’s not valid...if the test is reliable, if you test a person once, and you test that same person at different time, and that person has not make any progress or change, then the test should produce the same results. Then the test is reliable”	“If your [my] test is valid, it tests what it’s supposed to test...If my test can distinguish between the one who mastered the materials well and the one who did [not]...In general, which means if you teach grammar, you’re supposed to test grammar, you’re teaching something, you’re supposed to test that thing...And reliability is more on how you design the test, whether there are enough number of testers to evaluate the test, whether the criteria you use is okay or not, or whether it’s reliable in terms of condition that students have done in the class, whether students are familiar with the test format or design”
	Enhancing validity in the assessment implementation	“Assessment [test/exam]...has to reflect what the teachers [I] have taught, [and] what the lessons have been designed”	“So we [I] look for the materials that we [I] think is testing students’ ability that we [I] want to test...we [I] look for certain reading that is representative of what they have [been] covered in class”	“I try to cover most of the important points not any detail of it, but the important one I think it’s useful to remember...If my test can distinguish between the one who mastered the materials well...[and] the one who did not master materials, I think that’s [a] good test”
	Enhancing reliability in the assessment implementation	“I have never officially okay done that [statistical analysis of my tests]. But I have browsed through, for example, when I mark all the papers, I have browsed through it and I see that okay whether it’s reliable or not reliable, we [I] can see the scores, we [I] can reflect [on it] okay. So, basically I think I would say that it’s acceptable [reliable]”	“We [I] have never done statistical analysis formally. So we [I] assume mostly our [my] test is reliable even without using the test again and without doing the test [items] analysis”	“[To enhance] the reliability [of my test]...I try to make my [test] instructions clear [and] I try to make...[the test] items clear enough”

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LM1	LM2	LM3
Knowledge and understanding of the concepts of validity & reliability	Grading practices	<p>“[When I] teach more, you [I] become very, very tired...[so I] don't have time for marking, so teacher [I] always try to find way, okay, to do the marking easier [by means of glancing through the content of the student responses]”</p> <p>“I strongly agree that we have to count it [class participation] as one of the necessary assessments, yes not only the results of the tests and the results of the exams”</p> <p>“If someone [any students] got 48 % which [is]...based on the rule of the school, it means they fail. But to me I look at students, okay. If...[those] students...[are] very active, okay, and...they're capable enough as well...Unfortunately they don't do well in the exam, so, they lose the marks so I push...[these students]...I give extra scores [to them]...I think there is nothing in the world that is fair all the time”</p> <p>“The 100 scores consists of...ongoing assessment...[which is worth of] 50% [and] exam is 50%. Generally, in ongoing assessment, we [I] have test 1 and...test 2...presentations...class participation...[and] homework...[Course grade was used] to determine whether the students can pass...from one semester to another semester from year one to year two or not”</p>	<p>“I think there could be [a] strong correlation between class attendance and students' performance. So, I believe...[when] students are coming to class more often, it's more likely to help them learn better also”</p> <p>“If any students scored below 50...based on the rule, it's a fail...[but] I reward students, I reward hard work in addition to real students' ability doing things... out of a 100, we [I] give a 5% so this 5% would be added depending on how much effort you [I] feel the students are putting into their studies throughout the semester”</p> <p>“So...we [I] have test 1 [which accounts for]...10 to 15% of the 50 percent ongoing assessment. Then we [I] have test 2 [which is] almost equal [weight], and we [I] have assignment, we [I] might have homework, or class participation...[and 50 percent of the] semester exam”</p> <p>“Well, we [I] send...[course grades] to the [administrator]...and the [administrator]...basically checks and...[the grades are] announced. So, how are they used, they are used for against the policy for promotion and demotion of the students”</p>	<p>“When you [I] mark the tests [comprising an essay]...[I] just make sure you [I] make [my] marking fast [glancing through the content of the student work]...[so I] don't look [or read all aspects] closely...[regarding] what the students have written”</p> <p>“Well, I consider the other kinds of performance whether the student is working hard...involving in class [activities such as] whether they are doing their homework, [and] whether they participate activities at the classroom discussions...So I'll consider those factors and then come up with the issue: fail them or pass them...I just add to the one who needs to pass...there is no policy or rule that say if you [I] add scores [to a borderline student], and then you [I] must add to everybody [in the class]”</p> <p>“I have [50% for ongoing] assessment and another 50% is for final exam...I must follow what the school [department's assessment policies stated]...regarding criteria for assessments [and/or my grading practice]”</p> <p>“[I] have classroom [ongoing] assessment [50%] and another 50% is for final exam...[For] ongoing assessment...[I] have two kinds of progress tests, quizzes, we [I] have assignment, we [I] have presentations, we [I] have class participation, and homework”</p>

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LM1	LM2	LM3
Background related-factors	Pre-service assessment training	<p>“I...learned [assessment units] from Teaching Methodology [subject]...[and] I had a...[teaching practicum, too]...I learned how to design [the] test, [to judge] whether the test is reliable...[or] whether...it matched...what we [I] had taught...it...[was] more like theory base rather than practice base. We [I] didn't have chance to design the test [at all]...”</p>	<p>“I took a subject called Teaching Methodology in semester 2...yes, we [I] had [teaching] practicum [too]...the material I read, the material that is presented in class by teachers are very much related to different types of tests, the purposes of using different types of the tests, different test items, how different test items are used, reliability issues, validity issues, and we discussed that very broadly and then my experience go beyond just this four year-degree program at the English department. I'd been to five months' training in Applied Linguistics at the Regional Language Centre in Singapore and I took a subject called Language Testing...besides courses I had attended, I had also been to several international workshops, seminars, conferences, in which some of the papers had been focused on the issues of testing and assessments...[through this further assessment training] I think I had been very prepared in the knowledge of testing and assessments”</p>	<p>“When I was training in Teaching Methodology...[I also had a teaching practicum] for one month and [a] half...we [I] must learn how to teach and also how to assess. That's on the second semester syllabus...I don't think it prepared me a lot. At that we were [I was] discussing on theories and not really practical I mean. What I mean is that we [I] don't really have time to see the tests and then design the actual tests for students...the actual way of designing the tests was not implemented in my class...theory learning seems to be insufficient...”</p>
	Assessments experience as a student	<p>“At that time, we [I] had to memorise a lot of vocabulary [and] we [I] had to remember [the answer to]...the question itself...It...[was] more... about vocabulary [and] about the grammar rule... I think it lasted only a few days after the test/exam [administration]...If I...[didn't] use it again, then I...[forgot] it forever”</p>	<p>“[With regard to the] subjects like Literature, Core English, [and] Global Studies...there could be certain contents that you need to remember...I mean there are some materials that you [I] memorise with very little understanding...so the only way to do well in the tests is to remember the answers to particular questions...whether you [I] understand it or not, it may not matter a lot as long as you [I] can give the answers back to the teachers, then you [I] get the scores”</p>	<p>“Before the test, teachers give students the exam specification. And most of the exams which [is] based on the [course] content...key terms, or exact information from course books. So, wanting or not, I must memorise...we [I] think that the information that we [I] memorise does not really help us [me] a lot in our [my] general understanding”</p> <p>“Yes, it [my assessment experience as a student] influences more or less...[on] the way that I assess my students”</p>

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LM1	LM2	LM3
Background related-factors	Teaching hours & teaching payment and salary	“Basically I teach 24 hours... I have 5 classes... Yes, if I teach only...15 hours per week, I think that I would have enough time to design a better test... But if I teach like 18 hours or 24 hours, okay, I think I’m too tired to spend my time...[concentrating] on [developing new]...tests”	“The current hours that we are [I am] teaching if you understand that you’re [I’m] paid by the hours we [I] teach, so the more you [I] teach, the more you [I] earn...[If the department head] give me more money, I don’t have to teach a lot of classes and I just need to probably put a lot more effort in looking at the quality of teaching [or assessments employed]”	“It would be 33 hours per week...because you [I] teach many hours, many sessions in the week, and then you [I] will find...[myself] that you [I’m] always only in the middle of teaching and not having enough time to design the tests...and [I] don’t have time to design [the new tests], so...[I am] forced to use the same tests for different students”
	Professional development workshops	“Sometimes...[I] have [professional] workshops...It’s [professional workshop] not about assessments”		“Most of the professional workshops [provided in the department] focus more on teaching basically, but not really on assessments...Most of the time we deal with techniques in teaching”
	Class size		“We [I] have an average of 30 students in [each of] the classes, then what it means is you need [I have] 30 portfolios for a class. If you are [I am] teaching five classes, then it multiplies by five, then you [I] have how many [students], it’s impossible to do [portfolio assessment]”	

II. English Non-major Instructors

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LN4	LN5	LN6
Perceived assessment competence		<p>“[My assessment knowledge and skills was] 8...Because for the assessment, one thing we [I] just know clearly about the subject what we [I] are going to teach...And one more thing we [I] know the students...So, when we [I] know our [my] subject clearly and we [I] know the students, we [I] can design the test better”</p>	<p>“I think I scored 5 for myself...Actually, I don’t have great amount of knowledge in terms of testing...after 4 years [undergraduate study] at the English department, I...have [not had] any chance, okay, to further [my] study about testing”</p>	<p>“Yes, [I score my current assessment knowledge and skills] 7. The reason is that you know whenever I design the quizzes or the final exams, let me tell about my materials. I used World English...[as] the materials, [and] actually World English has the CD-Rom. The CD-Rom has the tests, [and] the tests include the vocabulary, grammar, reading and writing...I don’t actually follow everything from the CD-Rom...[I] try to make it [test] works for the students. So I guess the way that...[I adapt the test] is not really perfect”</p>
Notion of the ideal assessment		<p>“The [ideal] assessment [was] through the written and the oral [tests/exams]...and one more thing we [I] can ask students for their reflections, too...[Hence, I] teacher can know or learn more clearly about the student [abilities]”</p>	<p>“The [ideal] assessment will reflect, okay, what I have taught to the students. I will think about their ability whether...their ability fits with the tests or not, [their ability matches] the content of the tests, okay, [and] the language use, okay...I need to pilot [the test] for myself whether I can finish this test within the time limit or not...because when we [I] think about these factors, we [I] can [design the tests that] reflect the real ability of the students”</p>	<p>“[For my ideal assessment], I will adapt the resources to fit the students’ levels and the students’ backgrounds...[For example] when we [I] test reading, we [I] test the skills not the knowledge. So if we [I] choose the knowledge that is far beyond the students’ backgrounds, they cannot get it. And that can be a mistake for assessments”</p>

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LN4	LN5	LN6
Knowledge and understanding of the concepts of validity & reliability	Definition of the concepts of validity & reliability	“Reliability means that it is reliable with the scores, for example, with the correction [of the test papers]...And validity whether it is correct or not. For example, when we correct the students, sometimes the teacher makes mistakes...when the students give reasons”	“Reliability, I think when I give this test to the students, okay, this class and then I give another class, and we [I] get the same results from two groups of students, who are in the same level. And for validity, this test I can use this year, I get this result, so I expect next year when I give the same test to the students, and I get the similar results, too”	“I compare to what I taught with the test materials whether it really matches or... it’s far beyond what I taught them. This is a validity issue...for the writing test, I don’t [think] it’s reliable...[I] actually have the criteria of grading, but sometimes it’s not that much fair for each student...Actually we [I have] clear criteria but sometimes it depends on the idea of the teacher [myself]”
	Enhancing validity in the assessment implementation	“In order to make it [test] more valid, we [I] just design the test [that matches]...what we [I] have taught...and...[matches] the students’ level”	“I don’t do anything [to enhance the validity of my tests/exams]...[I] never care about validity”	“I compare to what I taught with the test materials whether it really matches or... it’s far beyond what I taught them...I will look at the results...How many percents that the students can achieve after I assess them. So, I will look at what is the mistake, is it the validity [issue?]”
	Enhancing reliability in the assessment implementation	“[I] make sure that it is so reliable when we [I mark student work by means of paying my attention to]...the fair correction...[for example, I] think that he [student] should get 85 or 90 [marks]...but he just gets only 65 or 70 [marks], yes, we [I] just check...what is the...reasons [behind]”	“I don’t do anything [to enhance the reliability of my tests/exams]...[I] never care about...reliability”	“I will look at the results...How many percents that the students can achieve after I assess them. So, I will look at what is the mistake, is it reliability [issue?]”

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LN4	LN5	LN6
Knowledge and understanding of the concepts of validity & reliability	Grading practices	<p>“The attendance, the class participation and homework...can be mixed, yes, because the teachers can know the students better...when they have the low attendance, it means that they are...[often] absent, and not active in the class...So, we just mix this one [the combination of class participation, attendance and homework] into 10%”</p> <p>“[If] they [students] have tried their best already, and they have good attitude, good manner, in the class, just 1 mark or 2 marks, we [I] just add [marks to their course grade to pass them]...I think it would be okay [it’s fair] because they have passed already. Why don’t they get so jealous with only 1 mark because the teacher [I have] just learned that this student is good and [1 mark] should be added [to his course grade]”</p> <p>“[My course grades comprised] ongoing assessment [consisting of]...quizzes...mid-term...presentation...assignment...class participation, homework...[ongoing assessments accounted for] 50% or 60%...[and] the final semester [exam accounted for] 50%...or 40% based on the group [of lecturers in the team]...[Course grades are used] for passing the students [to the next level]”</p>	<p>“[Class] attendance covers class participation, whether the students come or not, that’s class participation of attendance. [For] some students, they just come and take attendance without doing anything”</p> <p>“If I notice...[the] students just fail 2 or 3 points [or marks], I find the other way to help them like ask[ing] them to do extra work in order to get the supplement scores...It’s not fair [to add extra marks to only the borderline students], but we [I] cannot find...[a] better way to help those students...they already passed”</p> <p>“[My course grades comprised 50% of [the] ongoing assessment]...including attendance, quizzes, [and] mid-term...and...[another] 50% for [the] final [exam]...”</p>	<p>“For the writing test, I don’t [think] it’s reliable...[I] actually have the [marking] criteria of grading [student work], but sometimes it’s not that much fair for each student...[despite I have] clear [marking] criteria...sometimes it depends on the idea of the teacher [my beliefs concerning the quality of the content of student work]”</p> <p>“I believe that 10% is not too much...I give them [students] credit, I encourage them to come to the class”</p> <p>“If my students [are] from the Social Work, [or] Psychology [major]...I will let them pass because I believe that they must have a lot English training...[later with] their [majored] subjects...[in addition to their studies with] English department, so I guess they can catch [up with] the others for the next [coming] year...it is not fair [to add the marks to only the borderline students], but we [I] have no choice because we [I] just would like them to pass”</p> <p>“My [course grades consisted of] ongoing assessment [including]...class participation...homework, quizzes, writing and mid-term...and [the] final exam...We [group of teachers in the team] changed [the percentage of course grades]...In the past 60 [% was for ongoing assessment and], 40 [% was for the final exam], but now my band [team] changed [it] to 70 [% for the ongoing assessment and], 30 [% for the final exam]...[the course grades are used]...to pass them [students]”</p>

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LN4	LN5	LN6
Background related-factors	Pre-service assessment Training	<p>“Yes, [I learned assessment units from Teaching] Methodology subject...[I] learned...[mostly] in the theories...[and I] put...[them] into practice [when I am an in-service university lecturer]...[I] just try to apply [them into practice]...they [course instructors] just teach [traditional assessment theories], but they don't ask us to design the tests... it [assessment training] would not be sufficient”</p>	<p>“I think [the extent to which the assessment training prepared me to assess student learning was] around 30% because at that time we [I] did not have chance to design the test. We [I] were asked only to...[give a critique of] the test [given by my course lecturer”</p>	<p>“[I learned assessment units from] Teaching Methodology [subject]...I guess I don't [didn't] learn much about how to design the test...I just learned the theory...[the assessment training providing was] not enough”</p>
	Assessment experience as a student	<p>“They [course lecturers] just asked us [me] to do the test, do the exam and...assessed us [me] by giving marks...sometimes we [I] had to memorise... because there...[was] only one answer that...[was] correct...[I] usually forget [forgot] it [the answer] after a short time [or] after...[completing] the test... Yes, [I used similar assessment methods with my students]”</p>	<p>“[My course] lecturers just asked [me] about the key terms [and] most of the time before the exam day, we [I] tried to memorise every key term...I think [I forgot what I memorised] just about 1 day or 2 days after the test [administration]”</p>	<p>“If I take the Core English [subject tests/exams], what I need to memorise is the vocabulary. But if I take the Culture or Literature [subject tests/exams]... I need to memorise the lessons in order to answer the questions because the tests usually not test only vocabulary but [also assess] comprehension that I need to memorise...what I did is [to] make sure that I can remember [the lessons] during the tests, and after the tests, I don't care...so [the lessons were] not staying in [my] mind for long, I guess after one week or two weeks I still remember, but not much”</p>

THEME	SUB-THEME	INSTRUCTOR CODE & QUOTATION		
		LN4	LN5	LN6
Background related-factors	Teaching hours & teaching payment and salary	<p>“[I taught] more than 20 hours...[comprising] 4 or 5 [classes per week]...[I got paid by] teaching hours and also [by] salary...[for teaching fee-paying students, I got paid by] hours...[and for teaching scholarship students, I got paid by] the salary...[If I had good payment] and...less [teaching] hours...[I would] have more hours to design materials or design tests...it would be better”</p>	<p>“We [I] teach so many hours a week, so how can we [I] have the time to prepare the appropriate tests for the students’ level...we [I] copy and paste from the other materials in order to have a test for the students to do...this year, we [I] use the same [test], for some [students] they know the answers already, so [they] just come and then write down the answers...“[I am] the lecturer, we [I] don’t want to teach many hours, but we are [I am] forced to do so...If the salary is good, we are [I am] willing, okay, to design the good tests to help students, but the pay rate [of the teaching hour] is very low...[I get paid one hour for] five US dollars, [teaching in the] private program [class]...[and received one month of] 100 US dollars [for teaching per week of] 12 hours in the [two scholarship program] classes...[the payment for teaching the scholarship students is] extremely low”</p>	<p>“Yes, [I taught] 27 [hours per week with] 5 classes...it [my workload] definitely affected the way of...[implementing my] assessments... if I don’t [didn’t] have time, I just...[adopted the test from the available printed source or the CD-Rom to assess my students’ learning achievement]”</p>
	Professional development workshops	<p>“Just one or two workshops [provided by the department since I have worked here]...related to assessments...[other] workshops...related to our teaching, how to teach the students in this way or that way”</p>	<p>“[I want to have] training...[regarding] how to design a [good] test...when we [I] design [it], what...[I] need to [consider]”</p>	<p>“[The workshops provided by the department are more about Teaching] Methodology rather than [on] assessments...I would [like to] learn the way how to design the [good] test...[because] for [many] years, I use the same way of designing [tests]”</p>