Development and Validation of a Diagnostic Rating Scale for Formative Assessment in a Thai EFL University Writing Classroom: A Mixed Methods Study

Apichat Khamboonruang
ORCID Identifier: 0000-0002-7182-3501

Doctor of Philosophy (Arts)

August 2020

Submitted in Total Fulfilment for the Degree of Doctor of Philosophy

Department of Linguistics and Applied Linguistics
School of Languages and Linguistics
Faculty of Arts
The University of Melbourne

Abstract

Aimed at identifying learners' strengths and weaknesses on specific skills or contents, diagnostic assessment can provide fine-grained information to formatively promote teaching, learning, and language development in an ongoing language classroom (Alderson, et al., 2015, Elder, 2017; Jang, 2012; Knoch & Macqueen, 2017; Lee, 2015). While much research has developed diagnostic tools for large-scale standardised assessment, few have constructed diagnostic instruments for low-stakes formative classroom assessment. To contribute to the existing knowledge of diagnostic language assessment (e.g., Alderson et al., 2015; Jang, 2012; Knoch, 2007, 2009a, 2009b, 2011; Lee, 2015), this PhD research aimed to (1) develop a diagnostic rating scale for a formative diagnostic assessment to diagnose students' strengths and weaknesses in academic writing products and support ongoing teaching and learning in an EFL university classroom, and (2) explore the validity of the assessment claims following an argument-based approach to validation (Chapelle et al., 2008, 2010; Kane, 1992, 2006, 2011, 2012, 2013, 2016a, 2016b; Knoch & Chapelle, 2018). To this end, this research employed a multistage exploratory sequential mixed-methods design (Creswell & Plano Clark, 2018) to undertake the scale development and validation over three study stages: scale construction, scale trialling, and scale implementation.

Following the line of a multisource-driven approach to scale development (e.g., Banerjee et al., 2015; Knoch, 2007, 2009b; Montee & Malone, 2014), the scale was constructed and revised on the basis of theories of L2 writing ability, existing scales, expert intuition, and classroom curriculum. The scale was operationally implemented over the course of one semester in four writing classrooms, in which 80 English-major undergraduates used the scale to write, self-diagnose, and revise their assignment essays, and five teachers applied the scale to diagnose the students' essays and use diagnostic results to support teaching and learning. The teachers and twenty students were interviewed regarding their perceptions of the scale and assessment. The diagnostic scores were analysed using Classical Test Theory, Many-Facets Rasch, correlation, regression, and ANOVA statistics, and the perception protocols were analysed following a qualitative content analysis.

Overall, findings offered reasonable support for the overarching validity argument for the scale-driven assessment system. Yet, the different writing tasks to which the scale was applied over the course of instruction made it difficult to reliably gauge student progress, highlighting the need for stronger evidence relating to the consequence inference. This limits the usefulness of a measurement-driven assessment approach in detecting learning progression over the course. In addition, the current validation framework, driven by Kane's argument-based approach, appeared not to well capture the dynamic and varying evidentiary sources of learning and writing development in the classroom assessment. The present study provides implications for developing a diagnostic rating scale for diagnostic purposes in a formative assessment, and examining the validity of the assessment within the context of EFL language classroom.

Declaration

This is to certify that:

i. the thesis comprises only my original work towards the Doctor of Philosophy degree,
ii. due acknowledgement has been made in the text to all other material used,
iii. the thesis is fewer than 100,000 words in length, exclusive of tables, language examples, list of references, and appendices.

Signed:

Apichat Khamboonruang

Preface

I would very much like to gratefully acknowledge the University of Melbourne for generously awarding me the *Melbourne Research Scholarship*, including stipend and tuition fee offset, to undertake my doctoral degree throughout the period of my PhD study in Australia.

My grateful acknowledgement also goes to the School of Languages and Linguistics as well as the Faculty of Arts for kindly supporting my PhD research and international conference presentations at LTRC 2018, LTRC 2019, AAAL 2019, and Thai TESOL & PAC 2020 with the following grants: *Graduate Research in Arts Travel Scholarship*, *Research and Graduate Studies Scholarship*, and *Riady Scholarship*.

I would also like to acknowledge Educational Testing Service (ETS) for awarding me the *TOEFL Small Grant for Doctoral Research in Second or Foreign Language Assessment* to support this PhD research project.

Finally, I would like to thank the Association for Language Testing and Assessment of Australia and New Zealand (ALTAANZ) for awarding me the *ALTAANZ LTRC Student Travel Award* to support my travel to present this PhD research at the Language Testing Research Colloquium 2018 in Auckland.

Acknowledgement

I would like to begin by expressing my gratitude to my thesis committee, Associate Professors Ute Knoch (Principal Supervisor), Catherine Elder (Co-supervisor), and Carsten Roever (Chair), for taking care of my thesis during my PhD journey.

In particular, I am deeply indebted to both of my supervisors, Ute Knoch and Catherine Elder, for their excellent supervision as well as constructive and insightful feedback on my PhD thesis. I have learned a great deal from them. Their solid and specialist knowledge and advice have significantly shaped my research skills and knowledge in the field, and helped me overcome a number of challenges during the PhD research process and thesis writing. It would be really difficult for me to complete my PhD study without their generous and continued support.

I am also grateful to Professor Tim McNamara for broadening my intellectual horizon with his insightful lectures and thought-provoking ideas while I was taking his course in Researching Language. His knowledge and expertise are really awe-inspiring.

I am very much obliged to Dr Mike Linacre for his generous and specialist advice and assistance regarding Many-Facets Rash analysis and FACETS application.

I am also very thankful to my bosses, colleagues, and staffs in the Faculty of Humanities and Social Sciences at Mahasarakham University in Thailand for their support and assistance during my PhD study and research project. This study would not have been possible without them.

As always, I would like to express my deepest gratitude to my parents for their unflagging love, support, and encouragement throughout my personal and educational life. Without them, I could not have come this far.

It has been a truly worthwhile experience pursuing my PhD degree in the School of Languages and Linguistics at the University of Melbourne and a once-in-a-lifetime experience completing my PhD thesis amidst Melbourne's harsh lockdown and the COVID-19 outbreak.

Table of Contents

List of Tables

Page

List of Figures

Chapter 1: Introduction

This PhD research employed a three-stage exploratory sequential mixed methods research design (Creswell & Plano Clark, 2018) to develop and validate a diagnostic binary rating scale for formative assessment in order to assess students' strengths and weaknesses of an academic writing product in a Thai EFL university classroom context. This formative diagnostic assessment is intended to generate diagnostic information to support teaching and learning in an ongoing writing classroom. The scale was developed and validated following an argument-based approach to validation (Kane, 1992, 2006, 2011, 2012, 2013, 2016a, 2016b) and drawing upon a multisource or hybrid approach to scale development (e.g., Banerjee et al., 2015; Knoch, 2007, 2009b; Montee & Malone, 2014; Wagner, 2015) and a diagnostic language assessment approach (e.g., Alderson et al., 2015; Knoch, 2007, 2009b, 2011; Lee, 2015). In this chapter, I introduce the background of the current research and then address the research questions. Finally, I provide an overview of the subsequent chapters of this PhD thesis.

1.1 Research Background

It can be argued that much research has thus far been dedicated to streamlining large-scale, standardised, and high-stakes assessment of English proficiency particularly for university entry or exit purposes at the expense of small-scale continuing assessment of learners during a language course, programme, or classroom (Elder, 2017; Knoch, 2016). In light of this, growing attention has recently been drawn to small-scale, non-standardised, and low-stakes formative classroom assessment (Alderson et al., 2017; Elder, 2017; Knoch, 2016; Knoch & Macqueen, 2017; Tsagari & Banerjee, 2015; Turner & Purpura, 2016) and an ever-increasing amount of work has been devoted towards the utilisation of a continuing assessment to support language learning, teaching, and development in an ongoing classroom (e.g., Jang, 2012; Knoch & Macqueen, 2017; Lantolf & Poehner, 2011; Oscarson, 2014; Poehner, 2014; Poehner & Infante, 2016; Tsagari & Banerjee, 2015; Turner, 2012; Turner & Purpura, 2016).

In the classroom, assessment is often practiced separately from teaching and learning and more emphasis is put on summative assessment of learning achievement,

which allows teachers to appraise how much students have learned in relation to particular contents taught over a certain period of time (Cheng & Fox, 2017; Lee, 2017). It is argued, however, that summative assessment is not very helpful for teachers to gain routine detailed information about learner's learning problems, provide fine-grained and targeted feedback to learners, and fine-tune instructional and remedial activities to satisfy leaners' needs (Cheng & Fox, 2017; Knoch & Macqueen, 2017; Lee, 2017). This can be made possible through integrating assessment into an ongoing classroom and such formative assessment is argued to play a more productive role in improving progressive teaching and learning and ultimately enhancing students' language development over a language course (Jang, 2012; Lee, 2017; Turner & Purpura, 2016).

In response to the demand for learning-oriented, assessment-for-learning, or formative assessment in a language classroom, several approaches have thus far been introduced with a view to helping classroom teachers integrate assessment into ongoing teaching and learning so as to continually support teaching and learning in a language classroom. Diagnostic language assessment (DLA) is one of this line of assessment approaches having been recognised as effective for targeting learners' strengths and weaknesses in specific language knowledge, skills and abilities (Alderson, 2005; Alderson et al., 2015; Cumming, 2015; Harding et al., 2015; Jang, 2012; Jang & Wagner, 2014; Knoch, 2007, 2009a, 2009b, 2011; Kunnan & Jang, 2009; Lee, 2015; Lee & Sawaki, 2009a). In particular, DLA has been shown to be useful for supporting writing teaching and learning in a classroom (Kim, 2010; Wagner, 2015). This study, therefore, aims to take advantage of DLA to promote ongoing L2 learning, teaching, and assessment of writing in a Thai EFL university writing classroom context.

DLA has long been introduced in educational, psychological, and medical fields and has recently attracted growing research interest in the field of language assessment (Alderson, 2005, 2010; Alderson et al., 2015; Alderson, Haapakangas, et al., 2015; Cumming, 2015; Doe, 2013, 2014, 2015; Elder et al., 2009; Harding et al., 2015; Jang, 2012; Jang & Wagner, 2014; Kim, 2010, 2011; Knoch, 2007, 2009a, 2009b, 2011; Knoch & Elder, 2013; Knoch & Macqueen, 2017; Kunnan & Jang, 2009; Lee, 2015; Lee & Sawaki, 2009a, 2009b; Wagner, 2015). For diagnostic purposes, a DLA tool should be designed with emphasis on evaluating learners' strengths and weaknesses and provide fine-gained feedback that is individualised to learners' needs (Alderson, 2005; Alderson et al., 2015;

Jang, 2012; Knoch & Macqueen, 2017; Lee, 2015). During a DLA process, a DLA instrument should be repeatedly administered to continually help teachers gain formative diagnostic information on students' strengths and weaknesses and use such information to give feedback to students and adjust their upcoming instruction (Knoch & Macqueen, 2017; Kunnan & Jang, 2009; Lee, 2015). Learners, on the one hand, can benefit from DLA through their awareness and reflection of their strengths and weaknesses as informed by teacher feedback and/or their self-assessment of their own work (Alderson et al., 2015; Lee, 2015). In this way, they continually develop a sense of self-regulated, self-monitoring, and autonomous learners (Jang, 2012; Knoch & Macqueen, 2017). In particular, DLA is useful for classroom writing instruction where students need specific and individualised feedback in order to improve writing ability (Lee, 2017) which is complex and multifaceted in nature (Cumming, 2016; Hirvela et al., 2016).

Despite a recent surge of interest in a DLA approach, DLA is still in its early period of development and there remains a need for further development and experimentation in terms of theory and practice to advance the knowledge of DLA (Alderson et al., 2015; Jang, 2012; Kunnan & Jang, 2009; Lee, 2015). Accordingly, little information exists regarding principles, procedures, and guidelines for DLA tool development and validation as well as DLA implementation in different contexts (Alderson et al., 2015; Jang, 2012; Knoch & Macqueen, 2017; Lee, 2015). In addition, more classroom-based assessment research is still called for to shed more light on the processes of diagnosis (Alderson et al., 2015).

To date, much effort has been put to develop DLA tools for listening and reading diagnostic assessments (Chen & Chen, 2016; Jang, 2009; Lee & Sawaki, 2009a) and for large-scale diagnostic English tests, such as *Online Diagnostic Language Assessment System (DIALANG)* (Zhang & Thompson, 2004), *Diagnostic English Language Assessment (DELA)* (Brown & Lumley, 1991; Elder & Read, 2015, p. 25-46), *Diagnostic English Language Needs Assessment (DELNA)* (Knoch, 2009b; Read, 2015), *Diagnostic English Language Tracking Assessment (DELTA)* (Elder & Read, 2015, p. 70-92; Lockwood, 2013), *Canadian Academic English Language (CAEL)* (Doe, 2015), and recently *Diagnosing Reading and Writing in a Second or Foreign language (DIALUKI)* (Alderson, Haapakangas, et al., 2015). Few studies have developed DLA tools for diagnostic writing assessment (Kim, 2010; Knoch, 2007, 2009b; Wagner, 2015) and very few have constructed diagnostic scales for

ESL writing classrooms (Kim, 2010; Wagner, 2015), and for multiple-round assessment to support learning over a course of study (Wagner, 2015).

In the light of what previously discussed, DLA is potentially profitable to support writing teaching and learning and it should be particularly situated within formative assessment in a classroom context where teachers and students need detailed assessment information at multiple points in time to point them to specific learning strengths and weaknesses and to reflect, plan, and improve learning and teaching continually all the way through a course. The newly-developed diagnostic scale was accordingly intended to support formative assessment and generate diagnostic formation to improve teachers' teaching and students' learning in an ongoing classroom. In line with the current conceptualisation of validity and validation, Kane's argument-based approach to validation is adopted as the theoretical framework for the current scale development and validation. Kane's approach has been widely used to develop and validate assessment tools as it is acknowledged as a comprehensive and flexible framework requiring multiple sources of evidence to achieve the current validity (Chapelle, 2011a, 2011b, 2012; Chapelle et al., 2008, 2010; Knoch & Chapelle, 2018; Knoch & Elder, 2013; Sireci, 2013, 2016). As widely used as it is, how well Kane's argument-based approach can fit into a classroom assessment validation, where ongoing information about learning and language development is sought rather than simply a snapshot of students' ability at one point in time, still needs further investigation.

In contributing to this line of formative diagnostic assessment in the language classroom context, this PhD study will provide further insights regarding (1) the design and development of a diagnostic writing rating scale and a formative diagnostic assessment system in the context of Thai EFL classroom assessment as well as the applicability of Kane's argument-based approach to validation in formative classroom assessment, where learning development is of interest. To this end, this study employs a multistage exploratory sequential mixed methods research design to develop the diagnostic rating scale and validate the scale-driven assessment system by drawing upon a range of qualitative and quantitative data over three study stages: (1) scale construction, (2) scale trialling, and (3) scale implementation. The scale is designed as a generic diagnostic tool in the assessment system to diagnose English-major undergraduates' strengths and weaknesses in academic writing and to provide detailed diagnostic

information to inform teachers' and students' decisions and actions geared to continually improving teaching and learning in a Thai EFL university classroom setting.

1.2 Research Questions

This research aims to develop the diagnostic rating scale for the formative diagnostic assessment in the classroom and validate the assessment purposes or claims that the scale-driven assessment system is intended to achieve. Four research questions were addressed in order to seek empirical quantitative and qualitative evidence needed to make arguments in support of the intended claims of the assessment system, which will be addressed in Chapter 2. The research questions are as follows:

1) To what extent does the diagnostic rating scale function appropriately for the formative diagnostic assessment in the EFL university writing classroom?
2) To what extent does the diagnostic rating scale function consistently for the formative diagnostic assessment in the EFL university writing classroom?
3) To what extent does the formative diagnostic assessment system support formative decisions about teaching and learning in the EFL university writing classroom?
4) To what extent does the formative diagnostic assessment system have beneficial consequences for teaching and learning in the EFL university writing classroom?

1.3 Overview of the PhD Thesis

This PhD thesis is divided into seven chapters. This first chapter has introduced the background of the current scale development and validation research. Chapter 2 presents a comprehensive review of the relevant literature underpinning this research. Chapter 3 describes the rationales underlying this mixed methods research and the scale development and validation procedures over the three study stages with an emphasis on the scale implementation stage, the main focus of the study. Chapters 4 and 5 present the quantitative and qualitative results respectively. In Chapter 6, all the findings are first synthesised and discussed in response to the research questions and then are linked to the various claims in the validity argument. Finally, the research conclusions, implications, limitations, and suggestions for future researcher are discussed in Chapter 7.

Chapter 2: Literature Review

This chapter reviews the literature related to diagnostic language assessment, rating scale development, and current perspectives of validity and validation in order to provide the rationale for the current research. To begin with, I look at the key characteristics and components of diagnostic language assessment. I then review previous research on diagnostic writing assessment, arguing that very few studies have been situated within classroom contexts, been conducted longitudinally, and explored the interface between the key components involved in diagnostic language assessment.

After that, I explore the key stages involved in developing a diagnostic scale. In this section, I first look at a definition of L2 writing construct situated in the classroom context, which highlights the multifaceted nature of the construct and relevant sources of information underlying it. Then, I consider types of rating scale frequently used in performance-based assessment, pointing to the potential format that could optimise specificity of diagnosis and meanwhile facilitate implementation especially in the classroom context where there are practical constraints on teachers. I also consider approaches to scale development, pinpointing the value of a multisource-driven approach in providing a well-grounded basis for developing a classroom-based diagnostic scale.

In addition, I review current conceptualisations of validity and validation to emphasise the importance of validation with respect to both the design and the use of an instrument and multiple sources of evidence in keeping with current conceptualisations in the field. I then review previous research related to L2 classroom assessment validation, arguing that relatively few studies have applied argument-based validation frameworks in formative classroom assessment and no research on diagnostic assessment of L2 writing has employed such validation frameworks in classroom contexts and never before in a local EFL writing classroom.

The rationale for this research is then summarised, followed by the research questions, in turn linked to the current validation framework presented at the end of this chapter.

2.1 Diagnostic Writing Assessment in the L2 Classroom

Over the past two decades or so, significant work has been done to theorise and advance diagnostic assessment within the context of second language assessment (e.g., Alderson, 2005; Alderson et al., 2015; Jang, 2012; Jang & Wagner, 2014; Knoch, 2007, 2009b; Kunnan & Jang, 2009; Lee, 2015, Lee & Sawaki, 2009a). In spite of a growing body of research in this area, there are still relatively few studies on diagnostic assessment of L2 writing. In this section, I look at essential characteristics and procedural components of diagnostic language assessment before considering the significance of self-assessment, which is deemed as an essential element of diagnostic process. I then review current practices of prior research on L2 diagnostic writing assessment.

2.1.1 Characteristics of Diagnostic Language Assessment

According to Alderson (2005, pp. 11), diagnostic tests are designed to identify strengths and weaknesses in a learner's knowledge and use of language, are more likely to focus on weaknesses than on strengths, and are typically low-stakes or no-stakes. More recently, Lee (2015, pp. 303) defined diagnostic language assessment (DLA) as "*the processes of identifying test-takers' or learners' weaknesses as well as their strengths in a targeted domain of linguistic and communicative competence and providing specific diagnostic feedback and (guidance for) remedial learning*."

The main purpose of DLA is to identify learners' strengths and weaknesses in specific skills and processes being targeted in assessment and instruction (Alderson et al., 2015; Jang, 2012; Lee, 2015) and provide detailed information to subsequently improve learning and guide instruction (Jang, 2012; Jang & Wagner, 2014; Knoch & Macqueen, 2017; Kunnan & Jang, 2009; Lee, 2015). In particular, DLA should be focused on the weakness or area for further improvement as such information will inform future remedial action (Lee, 2015).

The goal of DLA distinguishes itself from other types of assessment in that DLA needs to target specific and discrete-point skills in order to provide specific and detailed information on learners' strengths and weaknesses (Jang, 2012; Lee, 2015). Increased specificity of diagnosis and feedback is thus a distinct characteristic of DLA (Lee, 2015). Information generated by DLA needs to be specific enough to serve the diagnostic purpose (Jang, 2012; Lee, 2015). Yet, it is challenging to arrive at the desired level of DLA

specificity or granularity specific to a given learning context (Jang, 2012; Lee, 2015) since specificity is a relative rather than absolute concept (Jang, 2012; Lee, 2015).

The level of specificity needs to serve the diagnostic and pedagogical purposes, clearly distinguishes between the strength and the weakness, direct students' learning, and help students act upon diagnostic feedback (Jang, 2012; Lee, 2015). Determining the optimal level of specificity may also need to consider such factors as learner characteristics, test purposes, assessment construct, and test design (Lee, 2015). An appropriately-defined specificity can support teachers' diagnostic judgment and feedback pertaining to students' mastery status (Jang, 2012). On the contrary, too narrow or broad granularity may generate either too specific or crude diagnostic feedback, potentially resulting in unintended negative washback and undermining assessment validity (Jang & Wagner, 2014).

In addition, DLA should be situated within formative assessment to continually provide diagnostic information to promote teaching and learning (Alderson et al., 2015; Jang, 2012; Kunnan & Jang, 2009; Lee, 2015). Accordingly, the process of DLA may also require collaborations with stakeholders, varying diagnostic tools, and teachers' diagnostic assessment expertise (Alderson et al., 2015; Kunnan & Jang, 2009; Lee, 2015). A single-round administration of DLA may not be sufficient to target the entire spectrum of language development (Jang, 2012; Kunnan & Jang, 2009; Lee, 2015).

To serve the diagnostic purpose, a DLA instrument should be designed to be user-friendly, targeted, discrete and efficient to help the teacher make diagnostic decision, should be suitable for administration in the classroom, be compatible with other existing pedagogical and diagnostic tools, and generate rich and detailed feedback for teachers and learners to adjust teaching and learning (Alderson et al., 2015; Cumming, 2015; Jang, 2012; Lee, 2015). As such, diagnostic feedback generated from a DLA instrument is deemed as an essential component of DLA (Alderson et al., 2015; Jang, 2012; Jang & Wagner, 2014; Knoch & Macqueen, 2017; Kunnan & Jang, 2009; Lee, 2015).

The type of feedback needed in the classroom is formative diagnostic feedback (Kunnan & Jang, 2009). It needs to be delivered to learners as immediately and routinely as possible so that they can still recall their reasons for responding the way they did on the task and take feedback for learning improvement (Alderson, 2005; Kunnan & Jang, 2009). When immediately and routinely provided, diagnostic feedback can reach its full

potential of integrating assessment with teaching, learning, and the curriculum (Kunnan & Jang, 2009) and becomes maximally informative and relevant (Alderson, 2005). However, it is challenging to provide immediate feedback in the classroom, which can be made possible via technological assistance (Alderson, 2005; Kunnan & Jang, 2009).

In addition, diagnostic feedback should enable students to realise their current level of performance and expected levels of performance or learning goal (Jang & Wagner, 2014) and reflect on their learning so as to take further remedial actions (Lee, 2015). It is potentially more effective when yielding information on learner's progress toward the expected standard or learning goal (Jang & Wagner, 2014). To ensure its positive impact and effectiveness, diagnostic feedback should be aligned closely with remedial learning activities (Lee, 2015). While specific feedback may facilitate learners' remedial actions to improve their deficient areas, too specific diagnostic feedback increases undesirable complexity for students (Jang & Wagner, 2014).

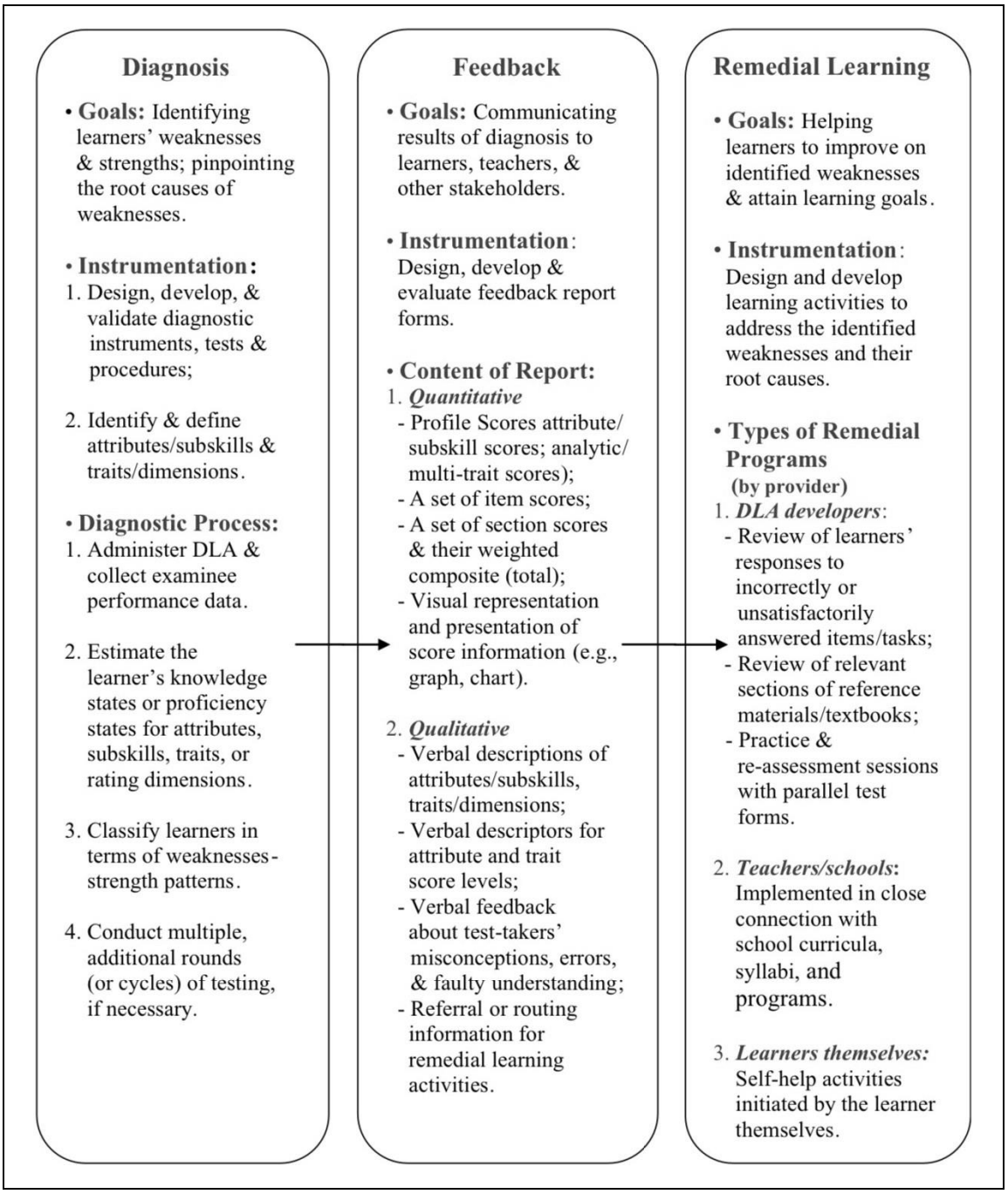
It should be noted that the effectiveness of DLA may be facilitated or impeded, depending not only on the quality of feedback generated from a diagnostic test but also on individual students' language ability and learning attitudes, teachers' diagnostic assessment competence, types and modes of feedback provision, and other relevant contextual factors (Alderson, 2005; Jang & Wagner, 2014; Kunnan & Jang, 2009).

2.1.2 Process of Diagnostic Language Assessment

Alderson et al. (2015) interviewed professionals from different fields with a view to theorising and characterising diagnostic assessment in the field of second language assessment. Based on the interview data, they proposed that a DLA system should involve four diagnostic stages: (1) listening and/or observing learners' problems, (2) an initial assessment of the problem, (3) using tools and consulting various sources of information, and (4) make decision, which required the synthesis of various knowledge strands. They argued, however, that much current diagnostic testing focuses on Stage 3, at which generic diagnostic tests are used for particular examinee populations rather than more targeted measures selected to meet the specific needs identified in Stages 1 and 2. However, Harding et al., (2015) argued that while the DLA process proposed by Alderson et al. has a firm theoretical basis, its application to the field of language assessment remains untested.

Building on the DLA literature and particularly Alderson et al. (2015), Lee (2015) proposed three key stages that should be incorporated into DLA process as shown in Figure 2.1: (1) diagnosis, (2) feedback, and (3) remedial learning.

Figure 2. 1 *Three Stages of Diagnostic Language Assessment (Lee, 2015, p. 308)*



The first stage is the diagnosis stage which is the focal stage of the process. This stage aims to identify learners' strengths and weaknesses in specific knowledge, skills, or abilities in order to determine the current state of learner's knowledge and envisage the

areas on which they need to improve. This diagnosis stage is concerned primarily with development and evaluation of diagnostic tools and interpretation of diagnostic results, which involve several activities, including (a) developing and evaluating various DLA instruments, (b) identifying and defining attributes or subskills to be assessed, (c) administering DLA instruments and scoring language performance data, (d) estimating learners' current state of the defined attributes or subskills, (e) classifying learners' patterns of strengths and weaknesses in the defined attributes or subskills, and (f), if necessary, conducting additional rounds of DLA.

The second feedback provision stage aims to report DLA results as feedback to learners, teachers, and other stakeholders. This stage involves developing and evaluating diagnostic profile reports which contain both quantitative and qualitative information describing and summarising diagnostic results in a way that is clear and interpretable for learners, teachers, and other stakeholders. It is important that diagnostic feedback be effectively conveyed to both teachers and learners so that they can take further necessary actions to improve teaching and learning.

The final remedial learning stage is aimed at taking diagnostic information to develop remedial activities or programmes in order to assist learners in improving their weaknesses, reinforcing their strengths, and ultimately fulfilling the expected learning objectives in a target learning context. This stage involves design, development, and implementation of learning activities or materials to help learners improve on their weaknesses and promote learning.

Apart from the key stages of DLA, diagnostic assessment process should incorporate stakeholder views, including learners' self-assessments and relate to some future remedial intervention (Alderson et al., 2015; Lee, 2015).

While the DLA processes suggested by Alderson et al. (2015) and Lee (2015) offer the potential to bring DLA into its optimal fruition, these processes encompass several demanding activities, including not only development and implementation of diagnostic instruments but also development and provision of diagnostic feedback and remedial intervention. To achieve this, the process of DLA may require a relatively extended period of time and necessitate close collaboration from key stakeholders, students' self-assessment, teachers' assessment and feedback literacy, and multiple-round assessment.

2.1.3 Self-Assessment in Diagnostic Classroom Assessment

As earlier pointed out, self-assessment is essential to enhance the effectiveness of DLA (Alderson et al., 2015; Kunnan & Jang, 2009; Lee, 2015) and it is thus viewed as an essential component of DLA (Alderson et al., 2015) and other forms of formative and classroom assessments (Brown et al., 2015). According to Yan and Brown (2017, pp. 1248), self-assessment of writing is "*a process during which students collect information about their own performance, evaluate and reflect on the quality of their learning process and outcomes according to selected criteria to identify their own strengths and weaknesses*."

Self-assessment is considered as valuable to be used alongside diagnostic assessment as it provides students with the opportunity to better understand the criteria used by teachers for diagnosis (Lee, 2017; Yan & Brown, 2017), and to compare teacher diagnostic feedback to their own assessment (Kunnan & Jang, 2009; Lee, 2015). This process of comparing is considered to be particularly important in helping students notice differences between their own understanding of their strengths and weaknesses and the assessment made by the teacher (Wang, 2017; Yan & Brown, 2017), and become aware of the areas they need to improve on (Andrade & Brown, 2016; Brown & Harris, 2013).

Despite the attempt to enhance students' self-assessment, a body of research has highlighted the difficulty in achieving reliable and accurate self-assessment from students. For example, Matsuno (2009) found that in an EFL Japanese university writing classroom, self-raters tended to be more severe than peer- and teacher-assessors while peer-raters were the most lenient and were more internally consistent than self-raters. Particularly, higher achieving students were not more severe and lower achieving writers were not more lenient. In an EFL Iranian university writing classroom, however, Esfandiari and Myford (2013) compared the levels of rating severity between self-, peer-, and teacher-assessors and discovered that self-assessors were the most lenient raters while teachers were the most severe raters. Baleghizadeh and Hajizadeh (2014) discovered that Iranian EFL learners' self-ratings were highly consistent with teachers' ratings. Notwithstanding the low-quality nature of self-assessment, previous studies have revealed the positive impact of self-assessment on EFL students learning in the classroom (Fung & Mei, 2015; Heidarian, 2016; Kim, 2019).

Whether students' self-assessment is reliable or not, students' beliefs about their own abilities should be considered in evaluating the impact of diagnostic feedback (Lee,

2015) and self-assessment should be used to promote students' self-regulated learning and academic achievement (Andrade, 2019; Andrade & Heritage, 2018).

2.1.4 Research on Diagnostic Assessment of L2 Writing

Despite a growing body of research on DLA, very few studies have thus far been conducted into diagnostic writing assessment. Knoch (2007, 2009b) developed and validated an analytic diagnostic rating scale to diagnose examinees' academic English expository writing ability for post-admission diagnosis in a large-scale university setting. In this study, she reviewed several theories of L2 language and writing abilities to generate a theoretical framework of writing quality features, which was then used as the basis for a discourse analysis of empirical language performances produced from the Diagnostic English Language Needs Assessment (DELNA). The framework of writing quality features was then built into the new scale and was refined based on raters' intuition. The revised scale included 8 domains of writing ability measured by 10 underlying indicators which were judged based on 4-to-5 scoring bands. The scale was operationalised with empirical writing performances produced from the DELNA. Knoch employed analysis of variance, multivariate analysis of variance many-facets Rasch model (MFRM), and thematic analysis of rater perception to evaluate the quality of the scale. Statistical results appeared to confirm acceptable psychometric properties of the theoretically-informed, empirically-derived scale. In this study, Knoch also proposed a designed diagnostic profile report for examinees. While the diagnostic analytic scale developed by Knoch was informed by the theory, language performance, and intuition, it was intended for post-admission diagnosis in a single-administration, large-scale, and standardised assessment rather than for ongoing diagnosis to support teaching and learning in the classroom.

A diagnostic writing rating scale designed particularly for classroom diagnosis was developed by Kim (2010). Kim developed and validated the Empirically-derived Descriptor-based Diagnostic (EDD) checklist to diagnose ESL adult learners' writing ability diagnosis purposes in a language programme or classroom. In this study, Kim reviewed existing rating scales, rater perception studies, written discourse analysis studies, and L2 writing theories to inform the dimensions of the construct and then employed a think-aloud protocol method to elicit experts' and teachers' perceived criteria while evaluating empirical writing performances produced by examinees of the TOEFL test in order to

develop the checklist descriptors. The checklist was revised on the basis of teachers' feedback and the pre-operational checklist consisted of five categories of writing ability and 35 descriptors, each dichotomously scored as mastered and non-mastered. The checklist was operationalised with TOEFL writing scripts. Kim adopted MFRM, correlation statistics, percent interrater agreement, dimensionality analysis, diagnostic classification model (DCM), and thematic analysis of rater perceptions to evaluate the checklist. Kim also developed a DCM-based diagnostic profile report for learners. Although the EDD checklist was designed for in-class diagnosis in the ESL writing classroom, it was informed and operationalised using TOEFL writing scripts produced within standardised testing without information linked to the curriculum which is deemed as necessary to inform classroom-based assessment. Furthermore, the validation of the EDD checklist was based on a single administration of the scale and was not operationalised in the real-world ESL classroom. It is thus questionable how well the EDD checklist functions to serve the intended diagnostic purposes in the real-world ESL classroom.

A study that investigated a broader process of DLA by implementing a diagnostic writing rating scale in the real-world classroom was carried out by Wagner (2015). In this study, Wagner developed the Diagnostic Rubric for Assessing Writing (DRAW) to diagnose ESL learners' English writing ability. However, her study focused on investigating the impact of the DRAW-generated diagnostic feedback on students' learning in the secondary classroom setting. Through an iterative process of scale development, the DRAW was developed and revised based on multiple sources of information, including standard curriculum, teachers' voices, students' voices, student writing samples, L2 writing theories, and existing rating scales. The DRAW included six categories of writing ability and 30 descriptors, each dichotomously scored as master and non-master. Wagner used the percent interrater agreement to evaluate the DRAW and rater judgement. The DRAW was repeatedly implemented in three ESL classrooms over three sequential tasks, in which students performed the writing tasks under the same testing conditions. Wagner used the scores on Task 2 to generate raw score-based diagnostic profile reports as feedback to learners and examined the impact of the feedback on student learning. In this study, teachers were asked to do peer- and student self-assessment activities. However, not all teachers were able to administer the third task and not all students completed the third task, and only a small amount of peer-and self-assessment data was obtained, thereby

limiting investigations of students' self-assessment behaviours and the formative impacts of the DRAW and its feedback on ongoing teaching and learning. Moreover, the DRAW was operationalised with standardised tasks and assessment conditions, which are not typical in formative classroom assessment.

Table 2.1 summarises the key characteristics of the three studies reviewed above. All in all, the three studies drew on multiple sources of information to develop a diagnostic writing rating scale. The features of writing quality on the three existing diagnostic writing rating scales should be useful to inform L2 diagnostic writing assessment. Therefore, these diagnostic rating scales are used to inform the dimensions of L2 writing construct in this study. It was nevertheless observed that these scales appeared to be operationalised and validated on the basis of standardised assessment situations and were not truly formatively used to support teaching and learning in the real-world classroom. Further, no studies have fully explored the interface between a diagnostic writing rating scale and other essential components of diagnostic language assessment (e.g., self-assessment, formative diagnosis) as proposed by Alderson et al. (2015) and Lee (2015). All this indicates that some of the key components in effective diagnosis relating to the provision and uptake of diagnostic feedback yielded by diagnostic assessment tools, multiple-round diagnosis, and student self-assessment in particular have thus far been underexplored and indeed never before explored in an EFL writing context which is the focus of the current study.

Table 2. 1 *Characteristics of the Studies on Diagnostic Writing Rating Scale Development*

Study features	Knoch (2007, 2009b)	Kim (2010)	Wagner (2015)
Purpose of scale use	• Post-admission diagnosis for academic university studies	• In-class diagnosis for ESL English writing classroom	• In-class diagnosis for ESL English writing course
Sources underlying scale development	• L2 writing theories • Raters' voices • DELNA examinee writing samples	• Teachers' voices • TOEFL examinee writing samples • Experts' voice • Existing rating scales • L2 writing theories	• Context curriculum • Teachers' voices • Students' voices • Student writing samples • L2 writing theories • Existing rating scales
Data elicitation methods for scale development	• Theory-driven discourse analysis of empirical writing essays	• Review of theories, previous research, and existing scales • Teacher think-aloud protocol • Expert review and discussion	• Review of curriculum, theories, and existing scales • Teacher review and discussion • Student think-aloud protocol

Study features	Knoch (2007, 2009b)	Kim (2010)	Wagner (2015)
Examinees or learners	• 100 L1/L2 English DELNA examinees	• 480 ESL TOEFL examinees	• 52 ESL secondary students
Raters or teachers	• Certified DELNA and trained raters	• Certified ETS raters and ESL teachers	• Researcher, experienced ESL teacher, and classroom teachers
Characteristics of writing task under diagnosis	• One-round standardised DELNA expository writing task (secondary data)	• One-round standardised TOEFL expository writing task (secondary data)	• Round 1: Standardised pre-test tasks • Round 2: Classroom-based task • Round 3: Standardised post-test tasks
Validation framework	• Assessment and use argument	• Argument-based approach	• No validation framework
Scale evaluation methods	• Analysis of variance statistics • Multivariate analysis of variance • Many-facets Rasch model • Rater questionnaire • Rater interview	• Many-facets Rasch model • Correlation analysis • Percent interrater agreement • Dimensionality analysis • Diagnostic classification model • Teachers questionnaire • Teacher interview	• Percent interrater agreement • Teacher review and discussion
Criteria domains	• Accuracy • Lexical complexity • Content - *Data description* - *Data interpretation* - *Data comparison (idea expansion)* • Hedging • Paragraphing • Coherence • Cohesion • Repair fluency	• Content fulfilment • Organisational effectiveness • Grammatical knowledge • Vocabulary use • Mechanic	• Idea • Organisation • Vocabulary • Sentence fluency • Mechanics • Grammar

2.2 Rating Scale Development

In the previous section, I demonstrated that previous research on diagnostic writing assessment in classroom settings is still scarce and that there is a clear need for further research on using diagnostic feedback in formative EFL settings. In this section, I review key stages involved diagnostic scale development in the classroom context. I begin by looking at a detailed definition of the construct of L2 writing assessment in the classroom context before exploring theories of L2 language and writing ability which can serve as the basis for the L2 writing construct definition. I then consider design features of a diagnostic scale, including scale formats and scale development approaches which will ensure scale practicability and adequate specificity of diagnostic assessment in an ongoing classroom.

2.2.1 Construct of L2 Writing in Classroom Assessment

The language construct is typically defined as language knowledge, skill, and ability that language learners possess (Bachman & Palmer, 2010; Schoonen, 2011) and is in itself complicated and multifaceted, encompassing several interactive language subskills (Purpura, 2008; Schoonen, 2011). Depending on assessment purposes and contexts, a meaningfully-decomposed construct needs to be informed by relevant theories and contextual variables (Bachman & Palmer, 2010; Brown, 2012; Fulcher & Davidson, 2007). The language construct can also be informed by rater decision-making behaviours in rater-mediated assessment (Cumming et al., 2001, 2002; McNamara, 1996) or informed by learning contents, language ability theories, stakeholders' needs, and learners' performances in classroom contexts (Bachman & Palmer, 2010; Brown, 2012). In addition, McNamara (1996) pointed out that existing tests or scales are useful to inform the language construct.

For diagnostic language assessment, theoretical models of language ability or development can provide a strong theoretical base for diagnostic assessment and a lack of theoretical grounds underpinning diagnostic criteria is considered as a threat to the validity for diagnostic assessment inferences (Jang, 2012). As diagnostic criteria should represent learning contents reflected through a variety of contextual sources (e.g., learning materials, syllabi, and teachers), contextual data should also inform diagnostic criteria (Jang, 2012; Knoch & Macqueen, 2017; Kunnan & Jang, 2009). Moreover, teachers can help intuitively determine diagnostic criteria they expect students to learn and achieve (Jang, 2012).

As can be seen, the language construct for diagnostic and classroom assessment can be driven by multiple sources of information. As the theory is viewed as essential to inform the construct characterisation in respective of assessment contexts, I next review theories of L2 language and writing ability in order to identify a set of writing ability features to inform the theoretical dimensions of the current L2 writing construct. It should be noted that the dimensions of the current construct are informed by the theory and existing diagnostic rating scale of writing whereas the descriptors representing the construct components are driven by the curriculum and teacher intuition in the context. The conceptual framework of the current construct will be presented in Chapter 3.

2.2.2 Theoretical Construct of L2 Writing

The construct of L2 writing is complex and multifaceted in nature (Cumming, 2016; Hirvela et al., 2016). Knoch (2011, pp. 90) asserted that no theory can serve by itself as a basis for the design of a rating scale for diagnostic writing assessment. She proposed that the theoretical models of communicative language ability, text construction and writing knowledge (Grabe & Kaplan, 1996), and rater decision-making behaviour should be considered as a basis for the design of a rating scale for diagnostic writing assessment. Weigle (2002) also argued that the communicative language ability models are useful to inform the construct of L2 writing test-takers' characteristics, assessment purpose and context. In this section, I, therefore, look primarily at the theoretical models of communicative language ability (Bachman & Palmer, 2010), text construction and writing knowledge (Grabe & Kaplan, 1996), and rater decision-making behaviour (Cumming et al., 2002) in order to define the components of the L2 writing construct in question.

2.2.2.1 Model of Communicative Language Ability

Several CLA models are proposed to describe the theoretical models of L2 ability (e.g., Bachman, 1990; Bachman & Palmer, 1996, 2010; Canale, 1983; Canale & Swain, 1980). These models describe L2 language ability as multi-componential knowledge and learners' use of linguistic competences for various communicative purposes (Purpura, 2008). As shown in Table 2.2, Bachman and Palmer's (2010) language knowledge is composed of two general interacting components: organisational and pragmatic knowledge. Organisational knowledge refers to how individual learners control language to produce grammatically correct linguistic forms whereas pragmatic knowledge deals with how individual learners communicate meaning and produce contextually appropriate utterances, sentences, and texts. Organisational knowledge includes grammatical knowledge (vocabulary, syntax, and phonology/graphology) and textual/discourse knowledge (cohesion, rhetorical organisation, and conversational organisation). Pragmatic knowledge is composed of functional and sociolinguistic knowledge. Functional knowledge has to do with knowledge of how to use organisational resources to communicate language functions, while sociolinguistic knowledge deals with how organisational resources relate to features of language use in context.

Bachman and Palmer proposed that the model is generic and can be applied to any language assessment situations. Purpura (2008) pointed out that the model not only describes various knowledge components constituting language ability but also describes a learner's ability to use this multi-componential knowledge appropriately in a specific context. For writing ability assessment, Weigle (2002) added that the model informs that L2 writing ability must essentially account for grammatical knowledge, sociolinguistic knowledge, discourse knowledge, and strategic competence. Furthermore, Connor and Mbaye (2002) argued that the model is a convenient framework for decomposing components of written discourse and therefore grammatical, discourse, sociolinguistic, and strategic competences should be reflected in scoring criteria.

Table 2. 2 *Model of Communicative Language Ability (Bachman & Palmer, 2010)*

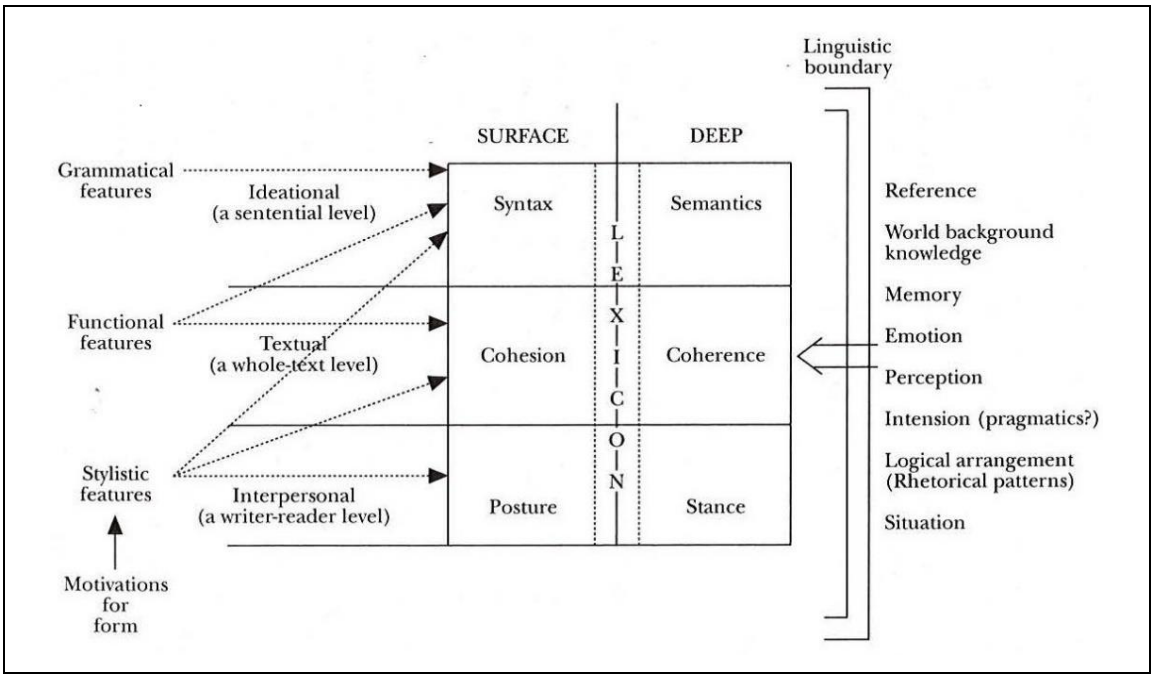
Organisational knowledge		Pragmatic Knowledge	
Grammatical	Textual	Functional	Sociolinguistic
• Vocabulary • Syntax • Phonology or graphology	• Cohesion • Rhetorical organisation	• Ideational functions • Manipulative functions • Heuristic functions • Imaginative functions	• Genres • Dialects or varieties • Registers • Natural or idiomatic Expressions • Cultural references and figures of speech

However, the CLA model does not suffice to serve the basis for writing assessment as it is developed to describe underlying competence and not performance and does not incorporate content and fluency as parts of language ability (Knoch, 2009b, 2011). The model may be difficult to apply in performance assessment since raters may attend to different components of language knowledge on rubric criteria (McNamara,1996). Further, the model does not account sufficiently for what it really means to communicate competently in an additional language (Harding, 2014). To properly describe a language construct, the model needs to be complemented by other theories/models and relevant contextual factors (Chalhoub-Deville, 2003; Chapelle, 1998) since knowledge construction is contextual, culturally embedded and socially mediated (Chalhoub-Deville, 2003; Knoch, 2011). Though the CLA model does not provide an adequate basis for the current L2 writing construct, different components of language knowledge in the model can be selected as the construct components of written product ability.

2.2.2.2 Models of Text Construction and Writing Knowledge

Not long after CLA models emerged in language assessment literature, Grabe and Kaplan (1996) draw on early CLA models (Bachman, 1990; Canale & Swain, 1980; Hymes, 1972), L1 cognitive and process-based writing models (Bereiter & Scardamalia, 1987; Flower & Hayes, 1980, 1981; Hayes, 1996; Hayes & Flower, 1980), and previous applied linguistics studies and theories on text construction to develop text construction (see Figure 2.2) and language knowledge models (see Figure 2.3).

Figure 2. 2 *Model of text construction (Grabe & Kaplan, 1996, p. 81)*



As displayed in Figure 2.2, the text construction model describes the components of how texts are constructed from at least seven inter-related basic components: (1) syntactic structure, (2) semantic senses and mappings, (3) coherent signalising, (4) genre and organisational structuring to support coherence interpretations, (5) lexical forms and relations, (6) stylistic and register dimensions of text structure, and (7) non-linguistic knowledge bases, including world knowledge. These elements are either surface or deep structures operating on sentential and textual levels and each element encapsulates numerous, interrelated sub-components which can be of use as indicators for written product assessment (Knoch, 2007, 2009a, 2009b, 2011; Weigle, 2002).

Apart from the text construction model, the language knowledge (see Table 2.3) describes language knowledge focusing on writing knowledge which is divided into (1) linguistic knowledge, (2) discourse knowledge, and (3) socio-linguistic knowledge. Linguistic knowledge is concerned with basic formal or structural components of language, discourse knowledge deals with the ways in which texts are constructed cohesively, and socio-linguistic knowledge is related to the ways in which language is used appropriately in various language use situations.

Table 2. 3 *Model of Language Knowledge (Grabe & Kaplan, 1996)*

Linguistic knowledge	Discourse knowledge	Sociolinguistic knowledge
・ Knowledge of the written code ・ Knowledge of phonology and morphology ・ Vocabulary ・ Syntactic/structural knowledge ・ Awareness of differences across languages ・ Awareness of relative proficiency in different languages and registers	・ Intra-sentential and inter-sentential making devices (cohesion, syntactic parallelism) ・ Informational structuring (topic/comment). ・ Given/new, theme, adjacency pairs ・ Semantic relations across clauses ・ Recognising main topics ・ Genre structure and genre constraints. ・ Organising schemes (topic-level discourse structure) ・ Inferring (bridging, elaborating) ・ Awareness of differences in features of discourse structuring across language and cultures. ・ Awareness of different proficiency levels of discourse skills in different languages.	・ Functional uses of written language ・ Application and interpretable violation of Grice maxims ・ Register and situational parameters ・ Awareness of sociolinguistic differences across languages and cultures ・ Self-awareness of roles of register and situational parameters.

The varying features and aspects of the text construction and writing-focused knowledge models can be useful for writing assessment (Knoch, 2011; Weigle, 2002). The models are developed based on previous models, theories, and studies in both L1 and L2 writing, incorporate language use features identified by CLA models and other contextual characteristics specific to writing (Weigle, 2002), and thus should be considered as the basis for defining the L2 construct. Since the models are product-based and provide basic aspects and variables necessary for writing research, they are useful for product-based writing assessment for defining L2 writing ability (Knoch, 2011, Weigle, 2002).

It is argued, however, that the models are developed specifically for language assessment and thus do not inform what features are more contributory to effective writing (Knoch, 2011), and how such features are structured hierarchically (di Gennaro, 2011). Moreover, the models do not account for differences in L2 writers' language

proficiency which might be related to observable features in writing performance (Grabe & Kaplan, 1996). Despite these limitations, several linguistic components and features described in the models can be of use to inform the current construct.

2.2.2.3 Model of Rater Decision-Making Behaviour

The models of CLA, text construction, and writing knowledge may not cover the writing aspects that raters attend to when evaluating language performances. As reliable and valid as a rating scale is, research reveals that raters with different backgrounds may not attend to the criteria on a rubric and may focus on different aspects of the rubric criteria (e.g., Cumming et al., 2001, 2002; Lumley, 2002, 2005; Milanovic et al.,1996; Sakyi, 2000; Smith, 2000). The models of rater decision-making behaviours based on previous research help uncover additional quality or ability aspects of language products that raters attend to when scoring performances (Knoch, 2011; McNamara, 1996). Raters' perceived criteria are beneficial to inform the development of a rating scale that can capture the intended construct (Knoch, 2011). Therefore, rater decision-making behaviour models are useful for describing the writing construct. This research focuses on Cumming et al.'s (2002) rater decision-making behaviour model since it is based on a rigorous study and widely used as the framework for rater decision-making behaviour studies (e.g., Baker, 2012; Barkaoui, 2007b, 2010; Han, 2017; Zhang, 2016).

Cumming et al. (2002) conducted a three-phase study employing a concurrent think-aloud method to develop, evaluate, and refine a framework describing experienced raters' decision-making behaviours when evaluating EFL compositions. Based on their findings, raters perceived that the important qualities of effective EFL writing include rhetorical organisation (introductory statement, development, cohesion, and task fulfilment), idea expression (logical argumentation, clarity, uniqueness, and supporting points), accuracy and fluency of grammar and vocabulary, and the amount of produced written texts. Their findings also revealed that most raters perceived that their previous experiences in rating compositions and teaching English had influenced their criteria and processes for rating the compositions. Table 2.4 presents Cumming et al.'s (2002) framework showing 27 rater decision-making behaviours based on their findings. They proposed that the model accounts comprehensively and parsimoniously for the decision-making behaviour that experienced EFL/ESL raters reported during think-aloud protocols.

They suggested that the rating behaviours in the framework can be used as the schemes for scoring EFL writing compositions and guiding rater training.

Table 2. 4 *Model of Rater Decision-Making Behaviour (Cumming et al., 2002)*

Strategies	Self-monitoring	Rhetorical/ideational	Language
Interpretation Strategies	• Read or interpret prompt or task input or both • Read or reread composition • Envision personal situation of the writer	• Discern rhetorical structure • Summarise ideas or propositions • Scan whole composition or observe layout	• Classify errors into types • Interpret or edit ambiguous or unclear phrases
Judgment Strategies	• Decide on macro-strategy for reading and rating; compare with other compositions; or summarise, distinguish, or tally judgments collectively • Consider own personal response or biases • Define or revise own criteria • Articulate general impression • Articulate or revise scoring decision	• Assess reasoning, logic, or topic development • Assess task completion or relevance • Assess coherence and identify redundancies • Assess interest, originality, or creativity • Assess text organization, style, register, discourse functions, or genre • Consider use and understanding of source material • Rate ideas or rhetoric	• Assess quantity of total written production • Assess comprehensibility and fluency • Consider frequency and gravity of errors • Consider lexis • Consider syntax or morphology • Consider spelling or punctuation • Rate language overall

Table 2.5 presents a synthesis of the aspects of L2 writing ability or quality selected from L2 language and writing theories under review. The selected features are categorised into four main components: (1) discourse and text, (2) linguistics and language use, (3) socio-linguistics and pragmatics, and (4) idea and content. As can be seen, all the theoretical models provide similar and different features representative of the four components of L2 language and writing knowledge. However, only the model of writing knowledge (Grabe & Kaplan, 1996) does not include the aspects related to the content and idea knowledge. These features can be considered as the basis for describing the quality and construct of L2 written product. The reader is reminded that the components of the current construct, informed by the theory and existing scale, are presented in Chapter 3.

Table 2. 5 *Synthesis of Features Based on L2 Language and Writing Ability Theories*

Components	Communicative language ability	Text construction	Writing knowledge	Rater decision-making behaviour
Discourse and text	· Cohesion · Rhetorical organisation	· Coherent signalising · Genre and organisation structuring	· Cohesion · Syntactic parallelism · Informational structuring · Main topic · Genre structure · Organizing schemes · Inferring (bridging, elaborating)	· Rhetorical structure · Whole composition · Layout · Topic development · Coherence · Redundancy · Text organization · Use and understanding of source material
Linguistics and language use	· Vocabulary · Syntax Graphology	· Syntactic structure · Semantic senses and mappings · Lexical forms and relations	· Morphology · Vocabulary · Syntax or sentence · Intra-and inter sentential making devices (connectors)	· Error · Ambiguous or unclear phrases · Quantity of total written production · Comprehensibility · Fluency · Frequency and gravity of error · Lexis · Syntax · Morphology · Spelling · Punctuation · Overall language
Socio-linguistics and pragmatics	· Varieties · Registers · Natural or idiomatic expressions · Genres	· Style · Registers	· Registers	· Style · Registers · Discourse functions · Genre
Idea and content	· Ideational functions	· Non-linguistic knowledge or content	· n/a	· Ideas or propositions · Task completion · Task relevancy · Ideas or rhetoric · Reasoning · Logic

2.2.3 Types of Rating Scales

A diagnostic tool needs to target specific and discrete-point features of language ability, provide detailed and immediate diagnostic feedback, and be user-friendly and practical in an ongoing classroom (Alderson et al., 2015; Kunnan & Jang, 2009; Lee, 2015). In this section, I present the characteristics of three types of scales, holistic scale, analytic scale, and checklist, which are typically-used rating formats for scoring constructed-task performances or responses. I also highlight the type of rating scale that is potentially well suited for formative diagnostic assessment in the classroom. It should be noted that different scholars in different fields may define "*rating scale" and "rubric"* as different types of assessment instrument (e.g., Brookhart, 2013) or use the terms interchangeably.

2.2.3.1 Holistic Rating Scale

The holistic scale is designed to evaluate the overall performance (Brown, 2012; Davis, 2015). The holistic criteria are developed based on the assumptions that language ability develops in a hierarchical manner (Turner, 2013) and different aspects of ability develop at the same rate (Weigle, 2002). Although the holistic scale includes different attributes of language production, it requires raters to make holistic evaluation or global judgement of overall ability (Brown, 2012; Turner, 2013; Weigle, 2002) and thus assign a single global score on a particular language production (Brown, 2012; Davis, 2015; Turner, 2013; Weigle, 2002). In this way, it tends to be easy to judge and use (Brown, 2012).

Due to a single global judgement, the holistic scoring does not produce useful diagnostic information about a person's writing ability (Weigle, 2002). Obviously, the holistic scale is not useful for diagnostic assessment as it provides a single global score of overall ability or performance. The type of diagnostic rating scale needs to target various and separate aspects of ability in order to provide detailed and targeted feedback.

2.2.3.2 Analytic Rating Scale

The analytic scale is aimed at evaluating different aspects of performance (Brown, 2012). Like the holistic scale, the analytic criteria rest on the assumption of ability hierarchy (Turner, 2013). The analytic criteria consist of different attributes of language production (Brown, 2012; Turner, 2013; Weigle, 2002) and require raters to separately score or judge different aspects of ability and may also make a global or holistic judgement of overall

ability (Brown, 2012; Turner, 2013; Weigle, 2002). Accordingly, it is suitable for diagnostic assessment in the classroom (Brown, 2012).

It is, however, argued that the analytic scale may not well identify weaknesses at sufficiently specific levels and point to the underlying causes of the weaknesses (Kunnan & Jang, 2009). It is also more time-consuming to rate than the holistic scale (Brown, 2012; Weigle, 2002) and may not be practically suitable for multiple-round assessment (Alderson et al., 2015). Moreover, raters may find it difficult to independently judge and give scores on various domains (Brown, 2012; Davis, 2015; McNamara, 1996). Though the analytic scale can assess in-depth ability and provide detailed information about the quality of individual traits or skills, it may not be practical and time-efficient for multiple-round diagnosis and immediate feedback provision in an ongoing classroom.

2.2.3.3 Binary Checklist

The binary checklist is designed to judge very specific attributes or features of language production (Arter & McTighe, 2001; Brown, 2012; McMillan, 2014). The checklist does not require the assumption of ability hierarchy (Brown, 2012). Although the checklist may be designed to roughly capture the quality of specific features (Brown, 2012), each item on the checklist is typically judged dichotomously as "*present or non-present*" or "*yes or no*" (Arter & McTighe, 2001; McMillan, 2014). Scores on various checklist items may be summarised into different categories or attributes or even a single total score for judging the overall quality of language production (Brown, 2012). Thus, the checklist can provide focused and specific information about varying aspects of test-takers' ability and is thus appropriate for diagnostic assessment (Brown, 2012). The checklist is also very useful when there is limited time to make more deliberate judgements and when a performance requires specific steps to be carried out in a particular order (Arter & McTighe, 2001; McMillan, 2014).

However, the checklist often contains little description of what constitutes good or poor performance and thus provide little information on the relative importance of various items or descriptors (Brown, 2012). It may not also be appropriate to assess performances containing a wider range of qualitatively distinct performance levels (Arter & McTighe, 2001). While the checklist is not effective to capture the complex layer of specific traits or skills, they allow criteria to include numerous discrete and specific

descriptors which could provide more digestible and specific feedback for learners than that provided by the analytic scale. Thus, the checklist format has the potential for serving as a diagnostic tool in an ongoing classroom.

Table 2.6 summarises the key characteristics of the holistic scale, analytic scale, and binary checklist. In the next section, I review approaches to rating scale development and highlight the approach that is potentially suitable for diagnostic scale development in the classroom.

Table 2. 6 *Characteristics of Holistic Scale, Analytic Scale, and Binary Checklist*

Key characteristics	Holistic scale	Analytic scale	Checklist
Focus of evaluation	・Full complexity of performance and overall performance	・Different dimensions of performance weighed in order of relative importance	・Presence or absence or rough quality of very specific features or skills
Assumption underlying the construct	・Ability develops in a hierarchical manner	・Ability develops in a hierarchical manner	・No assumption of ability hierarchy
Focus of judgement	・Global judgement of overall ability	・Separate judgements of abilities	・Several judgements of specific discrete skills
Number of scores assigned	・Single global score for overall ability	・Separate scores for various attributes	・Several scores for various attributes
Rating practicality	・Tend to be less time-consuming to judge and use	・Tend to be more time-consuming to judge and use	・May or may not be time-consuming to judge and use, depending on the number of descriptors
Specificity of assessment information	・Overall global information on ability	・Detailed information on different aspects of ability	・Detailed information on very specific features of ability but little information on quality level

2.2.4 Approaches to Rating Scale Development

This section reviews scale development approaches employed in the field of language assessment. To date, approaches to rating scale development have been defined and termed somewhat differently in the literature (e.g., Davis, 2015; Fulcher, 2003, 2012; Fulcher et al., 2011; Jamieson & Poonpon, 2013; Montee & Malone, 2014: Turner, 2013; Turner & Upshur, 2002). The terms used to label each approach is typically associated with the key information source(s) used to inform the scale. The purpose and

context of assessment will determine what approach is suitable for developing a particular rating scale (Fulcher, 2012; Turner, 2013). In this section, I discuss the key characteristics, advantages, and criticisms of each approach and point to the potential approach that would be viable for development of a diagnostic rating scale for a classroom assessment.

2.2.4.1 Intuition-Based Approach

An intuition-driven approach is the first and most commonly-used method (Davis, 2015; Fulcher, 2012; Turner, 2013). It is otherwise referred to as a priori/armchair method (Fulcher, 1993). In this approach, experts rely mainly on their prior experience and knowledge to adapt existing descriptors or develop new descriptors into the scale, try out the scale with samples of language performances, and revise the scale and criteria over time until the criteria well match actual language performances and experts are satisfied with the scale properties and functions (Davis, 2015; Fulcher, 2012).

Although some scholars argued that the theory implicitly informs the scale through expert experience and knowledge (Wilds, 1975, as cited in Fulcher, 2012), the approach is criticised as lacking explicit theoretical basis underlying the hierarchical structure of scale descriptors (Fulcher, 2012; Fulcher et al., 2011; North & Schneider, 1998) and lacking empirical underpinning from language performance, hence resulting in inadequate and abstract descriptions of actual language performance (Turner, 2013; Fulcher, 1996, 2012; Fulcher et al., 2011).

In the classroom context, students' language performances are normally evaluated by teachers. Therefore, the intuition-based approach is of use to inform scale development. However, it is not adequate to serve the basis for scale development in the classroom.

2.2.4.2 Theory-Informed Approach

A theory-based approach draws upon theories of language ability, acquisition and/or development to describe observable behaviours underlying the construct of interest and develop and organise criteria descriptors assumed to represent the linear and hierarchical relationship of language acquisition or development (Fulcher, 2012; Jamieson & Poonpon, 2013). Informed by the theory, this approach is typically used to develop

proficiency assessment scales (North, 1996; North & Schneider, 1998) which can be generalisable across tasks and contexts (Montee & Malone, 2014: Turner, 2013).

Nevertheless, the theory-based construct and criteria may or may not be relevant to actual language performance and assessment tasks in a given context (Turner & Upshur, 2002). Though generalisable across different tasks, the theory-based scale has not been shown to be equally valid for different types of tasks (Chalhoub-Deville, 1997; Turner & Upshur, 2002) and lacks empirical validation (Turner & Upshur, 2002).

Since the theory is deemed as important to inform the construct (Bachman & Palmer, 2010), it should be considered as the basis for scale development. Like the intuition-based approach, the theory-based approach alone does not suffice to inform scale development in the classroom context.

2.2.4.3 Empirically-Derived Approaches

The intuitively- and theoretically-based approaches are mainly criticised as lacking the link between performance scores and actual language performances in a specific context. Accordingly, several empirically-derived approaches are proposed to solve such problems by drawing explicitly on empirical language performance to develop a rating scale: *(a) performance data-driven approach, (b) empirically derived, binary-choice, boundary definition approach, and (c) performance decision tree approach*. Each of these approaches is discussed in more detailed next.

One of the empirical approaches is a performance data-driven method (Fulcher, 1996, 2003, 2012). In this approach, a discourse analysis is used to extract salient features at varying levels of proficiency from a sample of empirical language performance produced in a specific language use domain (Fulcher, 2012). Then, statistical modelling may be used to estimate the accuracy with which these features can be used to separate performance levels, thus providing empirical evidence for the number of levels and the content of descriptors (Fulcher, 2012). Developed based on empirical language performance on a specific context, the performance-driven criteria thus provide rich and relevant descriptions of actual language performance and are authentic to an assessment context (Fulcher, 2012; Davis, 2015; Turner & Upshur, 2002). Nevertheless, this type of scale tends to be task-specific and is criticised as lacking theoretical underpinning for hierarchical descriptors (Fulcher et al., 2011). The approach is also time-consuming and

tends to generate level descriptors that are too complicated and difficult to rate in practice (Davis, 2015; Turner & Upshur, 2002).

Another empirical approach is an empirically derived, binary-choice, boundary definition (EBB) method (Turner & Upshur, 2002; Upshur & Turner, 1995, 1999). In this method, a group of scale developers, experts, or teachers individually rank a representative sample of empirical language performance produced in a specific context, together determine the specified number of performance quality levels, and identify and describe salient features distinguishing performances at adjacent levels on the basis of a series of repeated and branching binary decision (Turner, 2013; Turner & Upshur, 2002; Upshur & Turner, 1995). While the performance-driven approach relies on a discourse analysis of empirical language performance, the EBB approach draws on a series of experts' repeated and branching binary decision in order to generate questions that could separate the quality levels of language performance samples (Turner & Upshur, 2002; Upshur & Turner, 1995). In this way, the EBB scale yield criteria representative of actual language performance in a specific context (Jamieson & Poonpon, 2013; Turner & Upshur, 2002), does not place a heavy cognitive burden on raters' decision-making (Jamieson & Poonpon, 2013), is relatively easy to use in classroom practice (Jamieson & Poonpon, 2013; Turner & Upshur, 2002) and tends to provide high inter-rater reliability (Turner, 2013). Like the performance-driven scale, however, the EBB scale is driven by empirical language performance within a particular context and thus tends to be task-specific (Fulcher, 2012; Turner & Upshur, 2002). Despite building on empirical language performance, the EBB scale may not contain rich descriptions of actual language performance (Fulcher, 2012). Moreover, it lacks general applicability and the performance samples used to develop the scale and the scale developers may affect the criteria (Turner & Upshur, 2002; Upshur & Turner, 1999).

To enhance the richness of descriptors and practicality of a rating scale, Fulcher et al. (2011) build on the performance data-driven and EBB approaches to develop the performance decision tree method which draws on both a comprehensive analysis of empirical language performance and a series of experts' binary decision (Fulcher, 2012; Fulcher et al., 2011). This approach does not rest on the assumption of linear and hierarchical relationship of descriptor (Fulcher et al., 2011). In this approach, a detailed discourse analysis is used to identify salient features of a representative sample of

language performances in a specific context and then experts perform a series of repeated and branching binary decision regarding the presence of key discourse elements and features that represent the quality of language performance (Fulcher et al., 2011). In this way, descriptors are richly representative of actual language performance (Fulcher et al., 2011). Driven by empirical language performance from a specific context like the performance-driven and EBB approaches, the performance decision tree approach tends to be task-specific and context-bound (Davis, 2015) and is also time-consuming (Davis, 2015). A detailed account of the process involved can be found in Fulcher et al. (2011).

Overall, the three empirically-derived methods discussed above seem to draw heavily on empirical language performance without regard to the curriculum deemed as necessary to inform classroom assessment. In the classroom, assessment criteria are strictly tied to standards, curriculum or learning contents and what teachers expect students to master rather than student language performance. Thus, a detailed discourse analysis of language performance may not necessarily useful for scale development in this context. Moreover, existing rating scales based on the empirically-derived methods tend to be task-specific and comprise a small number of features or descriptors which may not target several specific and discrete-point skills and provide rich and targeted diagnostic feedback. Therefore, either of these methods may not be viable for developing task-generic and diagnostic rating scales.

2.2.4.4 Curriculum-Oriented Approach

Another scale development approach mentioned in the literature is the curriculum-based approach which draws primarily on learning objectives, syllabus, or curriculum to inform scale development (Montee & Malone, 2014; Turner, 2013; Turner & Upshur, 2002). Tied strongly to the learning curriculum in a specific educational context, the curriculum-driven scale provides useful information about how well students learn within one classroom and across classrooms where students are expected to progress at similar rates (Montee & Malone, 2014).

However, this type of scale may not be reliable, as instructors may be biased in their application of the scale and it is difficult or impossible to apply the scale outside of the context (Montee & Malone, 2014). The curriculum-driven approach is relevant and useful to inform a classroom-based rating scale as it draws on curriculum-related sources.

Like the intuition- and theory-based approaches, the curriculum alone is still not adequate to account for the construct of language and learning in the classroom. Table 2.7 summaries the key characteristics of the scale development approaches discussed so far.

Table 2. 7 *Characteristics of Rating Scale Development Approaches*

Approaches	Data sources	Advantages	Criticisms
Intuition-based	• Intuitive judgment • Existing scale	• Not time-consuming to develop	• Lacks theoretical underpinning for hierarchical scale descriptors. • Lack empirical language performance, thus yielding inadequate and abstract descriptions of actual language performance.
Theory-based	• Relevant theories	• Generalisable across tasks and context.	• Lacks empirical language performance, thus yielding inadequate and abstract descriptions of actual language performance. • May not be equally valid for various task types.
Performance-driven	• Empirical language performance • Statistical modelling	• Provides rich and relevant descriptions of actual language performance. • Provide stable measurement properties • Enhance practicality and authenticity to a given situation.	• Lacks theoretical underpinning for hierarchical scale descriptors. • Tends to yield too complex level descriptors difficult to use in real-time rating. • Take a great deal of time to develop.
Empirically-derived, binary-choice, boundary-definition	• Intuitive judgment • Empirical language performance	• Easy and practical to use in real-time rating. • Does not place a heavy burden on the memory of the raters. • Enhances reliability and validity in classroom assessment. • Provides rather rich and relevant descriptions of actual language performance.	• Lacks theoretical underpinning • Time-consuming to develop • Lacks general applicability • May not provide rich descriptions of language performance • May be affected by language performance samples and scale developers • Appropriate for specific task and context
Performance decision tree	• Intuitive judgment • Empirical language performance	• Easy and practical to use in real-time rating. • Provides rich and relevant descriptions of actual language performance.	• Appropriate for specific task and context
Curriculum-oriented	• Curriculum-related materials	• Provides useful information about how well students learn within one classroom and across classrooms, where students are expected to progress at similar rates.	• May not be reliable as instructors may be biased in their application of the scale • Appropriate for specific context

2.2.4.5 Multisource-Driven Approach

Multiple scale development approaches may be combined to create a hybrid approach (henceforward referred to as a "*multisource*" approach) to address some of the limitations that each of the earlier methods has (Banerjee et al., 2015; Montee & Malone, 2014). The multisource-driven approach is referred to in this study as a systematic and iterative process of utilising and triangulating data from multiple information sources to inform scale development and revision in order to optimally arrive at a final rating scale and rating criteria.

It is argued that the hybrid approach provides the benefits of triangulating information from multiple data sources in rating scale design and thus generates optimal outcomes (Banerjee et al., 2015). Over the years, more studies tend to draw on multiple sources to develop rating scales (Banerjee et al., 2015; Chan et al., 2015; Lallmamode et al., 2016) and diagnostic rating scales of writing in particular (Kim, 2010; Knoch, 2009b; Wagner, 2015).

As previously discussed, the intuition-based, theory-based, empirically-derived, and curriculum-oriented approaches each tend to draw more or less on one or a few of the following: intuition, theory, language performance, and curriculum. Each of these approaches is also aimed at different assessment purposes, different assessment tasks, and different assessment contexts. The approaches also differ in relation to descriptor richness, task and context generalisation, and real-life practicality.

Since the construct of L2 writing in the classroom assessment is multifaceted and tied to various sources and variables, such as the theory, stakeholder, and curriculum (Bachman & Palmer, 2010; Cumming, 2016; Hirvela et al., 2016; Weigle, 2002), a single approach may not be particularly well suited to the development of a classroom-based diagnostic rating scale. The multisource-driven approach could bridge the gaps in each method and provide the benefit of triangulating multiple sources of information to arrive at optimal outcomes for scale development. In light of this, the multisource-driven approach drawing on a combination of intuition, theory, existing scales, and/or curriculum sources offers the potential to serve as a well-grounded basis for classroom-based scale development.

2.3 Assessment Validity and Validation

This section reviews current conceptualisation of assessment validity and validation. In particular, the focus will be on Kane's argument-based approach to validation which is used as the theoretical framework for this research. After that, I will present the claims or purposes of the current scale and assessment, which in turn inform the current research questions.

2.3.1 Contemporary Perspectives of Validity and Validation

The concept of validity has long been refined and debated and various definitions of validity and validation have been put forwards in the literature (e.g., American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1985, 1999, 2014; Bachman & Palmer, 2010; Cizek, 2012, 2016; Cronbach, 1988; Kane, 1992, 2006, 2011; 2012; 2013, 2016a, 2016b; Messick, 1989, 1995; Sireci, 2013, 2016). Traditionally, validity is perceived as the degree to which a test measures what it purports to measure and it is commonly thought of as content validity, criterion-related validity (predictive and concurrent), and construct validity (Akbari, 2012; Sireci, 2016). Contemporarily, the conceptualisation of validity is extended beyond such traditional concept to consider validity as the degree to which test scores are meaningfully interpreted and used as intended by test developers (AERA, APA, & NCME, 2014; Bachman & Palmer, 2010; Cizek, 2012, 2016; Kane, 1992, 2006, 2013, 2016a, 2016b; Messick, 1989; Sireci, 2013, 2016).

There is, nevertheless, a debate in the field as to whether both the interpretation and the use of test scores can be validated separately (see Cizek, 2012, 2016; Sireci 2013, 2016, for further discussion). Some validity theorists posit that the interpretation and the use of test scores require different sources of evidentiary support and thus can be validated separately (Cizek, 2012, 2016; Messick, 1989). Others argue that the interpretation and the use of test scores are inextricably intertwined and validating score interpretation is a necessary but not sufficient component for supporting the use of test scores for a particular purpose (Kane, 2013, 2016a, 2016b; Sireci, 2013, 2016) Therefore, two validity arguments need to be constructed for both test score interpretation and use (Kane, 2013). The latest definition of validity in the Standards for Educational and Psychological Testing (AERA, APA, & NCME, 2014) appears to highlight the inseparable

connection of both interpretations and uses in educational and psychological test validation. According to the Standards, validity refers to "*the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests*" (AERA, APA, & NCME, 2014, p. 11).

To date, many validation frameworks have been proposed in accordance with the modern validity concept. Five validation frameworks have typically been used in language assessment and validation research: unitary theory of validity (Messick, 1989), assessment use argument (Bachman & Palmer, 2010), argument-based approach (Kane, 1992, 2006, 2011, 2012, 2013, 2016a, 2016b), evidence-centred design (Mislevy & Yin, 2012), and socio-cognitive framework (Weir, 2005). The commonly-known frameworks include the unified validation framework (Messick, 1989) and a line of the argument-based approach driven by Toulmin's argument model (Bachman & Palmer, 2010; Cronbach, 1988; Kane, 1992, 2006, 2013). The frameworks proposed by Kane and Messick lay the foundation for the validity concept stipulated in AERA, APA, and NCME's Standard for Educational and Psychological Testing (Sireci, 2013, 2016).

While Messick's unified framework considers construct validity as the most important validity covering other types of validity, Kane regards validity as the meaningful interpretation and use of test scores as specified in the interpretation and use argument instead of drawing entirely on the construct validity proper. Kane's approach also allows the construct to be investigated as part of the score interpretation and use (Kane, 2013). In addition, Messick's validation framework is seen as complex and abstract and provides little practical guidance for validation research (Knoch, 2016; Knoch & Elder, 2013; Xi & Davis, 2016; Xi & Sawaki, 2017), whereas Kane's argument-based approach to validation is flexible and provides clear general guidance allowing tests developers to propose the interpretations and uses of test scores as they see fit (Chapelle, 2011a, 2011b; Chapelle & Voss, 2014; Kane, 2013). This approach allows test validators to propose a wide range of sensible interpretations and uses while remaining rigorous in that the proposed interpretation and use need to be substantiated sensibly and sufficiently by evidence (Cumming, 2013; Chapelle, 2011a, 2011b; Chapelle & Voss, 2014; Kane, 2013; Knoch, 2016; Xi & Sawaki, 2017).

Compared to Messick's framework, Kane's argument-based approach has been widely used in language assessment and validation research across a wide range of

contexts by virtue of its flexibility and comprehensiveness (e.g., Chapelle, Chung, et al., 2010; Chapelle et al., 2015; Chapelle et al., 2008, 2010; Jang, 2009; Knoch & Chapelle, 2018; Knoch & Elder 2013). Although, it is argued that the argument-based approach is designed on the basis of high-stakes and standardised testing (Moss, 2013, 2016), it can generally be used across classroom assessment contexts (Chapelle & Voss, 2014). In light of this, Kane's argument-based approach has the possibility to serve the theoretical framework for the current classroom assessment research. In the next section, I discuss the basic concepts and characteristics of Kane's argument-based approach to validation.

2.3.2 Argument-Based Approach to Validation

In Kane's argument-based approach (Kane, 2006, 2013, 2016a, 2016b), validity is conceived as the property of the proposed interpretation and use of test scores, not the assessment instrument proper. Accordingly, validation is the process of validating the proposed score interpretation and use by way of accumulating and evaluating backing evidence to justify the plausibility of the proposed score interpretation and the appropriateness of the intended score use. The degree of validity depends on how well the collected evidence substantiates the proposed interpretation and use of test scores. Kane's argument-based approach involves two main activities of building two different yet interconnected arguments. The first activity is to develop the interpretive and use argument (IUA) where the proposed interpretation and use of test scores are stated. The second activity is to construct the validity argument for the proposed interpretation and use of test scores. The two arguments are discussed in detail in the next sections.

2.3.2.1 Interpretive and Use Argument

The specification of the proposed interpretation and use of test scores is the first step in an argument-based approach. The proposed interpretation and use of the test score may be otherwise conceived as the claims, decisions, consequences, or testing purposes (Sireci, 2013) that test developers propose for a particular assessment. Test developers typically have some purposes in mind, and such purposes can guide the development of the IUA (Kane, 2013). The score interpretations and uses are the arguments articulated as inferences with their underlying assumptions that are both interconnected and interdependent. The inference is a logical statement or warrant that

moves from one fact or proposition to another (Chapelle, 2011a, 2011b). The inferences are defined by test developers to underlie proposed interpretations and uses of test scores and the number of inferences depends on the number of the proposed score interpretations and uses. Each of the inferences has a warrant which is based on the underlying assumptions that point to the types of backing evidence needed to justify the plausibility of the warrant (Chapelle, 2011a, 2011b). The inferences and assumptions are framed through a network laying out statements starting from test performances to the conclusions, decisions, or consequences to be made based on test scores (Kane, 2013). As complicated interpretations and uses of test scores call for strong supportive evidence and demanding validation process, test developers should avoid unnecessary complexity of the score interpretation and use in the IUA (Kane, 2013).

In addition, the IUA framework needs to be clear, coherent, and complete for a particular assessment situation in order for test developers to precisely envisage how an assessment system should be developed and what kinds of evidence needed to examine and collect in justification of the proposed score interpretations and uses (Kane, 2013). The coherence and completeness of the IUA will direct how an assessment tool and an assessment system should be developed, what contents and properties of an assessment tool are suitable, what relevant facets or factors should be included or controlled in an assessment situation (Kane, 2006, 2013). The assessment tool, system, and IUA can subsequently be changed, revised and adjusted through an iterative process of development and revision until they are correspondent and considered acceptable (Kane, 2013). Although the focus of this development stage is on the development of the assessment tool, system, and IUA, much of the evidence needed for developing the validity argument is amassed during this stage. Test developers can commence gathering evidence during the processes of test development, IUA development, and research conducting (Chapelle, 2011a, 2011b). It is also important at this stage that test developers explicitly articulate what evidence should be accumulated, how much evidence is needed, and how the collected evidence should be merged to shed light on the validity argument (Chapelle, 2011a, 2011b).

2.3.2.2 Validity Argument

Once the test and the IUA are developed and all sources of required evidence are collected, the second activity is to construct the validity argument which needs to be provided for both the score interpretations and the score uses (Kane, 2013; Sireci, 2016). Again, validity is interpreted as the extent to which the proposed interpretations and uses of test scores are justified through (a) conceptual analysis of the coherence and completeness of the proposed interpretations and uses and (b) empirical analyses of the inferences and assumptions inherent in the proposed interpretations and uses (Kane, 2013, 2016a, 2016b). The proposed interpretations and uses that are sound and supported by appropriate and sufficient evidence are considered as valid, whereas those not adequately and appropriately substantiated by evidence are not regarded as valid (Kane, 2013). The validity argument thus relies essentially on the coherences and completeness of the IUA and the evidence collected in support of the proposed interpretation and use of test scores specified in the IUA (Kane, 2013). Kane (2013) argued that if the interpretation and use include a small number of inferences and assumptions that are plausible a priori, they would not require much empirical backing. If they consist of several inferences and assumptions that may not be sufficiently supported a priori, they require more empirical evidence.

In short, there are two types of arguments in Kane's argument-based approach that test developers need to develop during the test development and validation processes. The first argument is the IUA, which needs to be developed initially for proposing the interpretation and use of test scores, specified through a coherent and complete network of inferences and assumptions. The clearly-specified inferences and assumptions pave the way for test developers to envisage the kinds of evidence to be accumulated in support of the defined inferences and assumptions. The more ambitious the proposed interpretation and use of test score are, the more compelling the evidence is required, and hence the more demanding the validation process is.

2.3.3 Research on L2 Classroom Assessment Validation

As mentioned earlier, although Kane's argument-based approach is initially framed for high-stakes and standardised testing, it has been used in several studies as the framework for validating L2 classroom language assessments in different contexts.

Chapelle, Chung, Hegelheimer, Pendar, and Xu (2010) validated the piloted test of ESL productive grammatical ability for placement purposes in university classrooms. The test was delivered as a paper-based, single-administration, and standardised assessment. To provide validity evidence for the interpretation and use of the test, they proposed five inferences for validation: *domain description, evaluation, generalisation, explanation, and extrapolation*. The application of the argument-based approach in this study, nevertheless, was situated within a static, single-administered, and standardised assessment, for which the argument-based approach was particularly initiated. Such assessment is different from a non-standardised formative assessment, where assignment tasks are rather varied and performed by students under various conditions, and where assessment is continual to provide formative information to promote teaching and learning.

Studies applying the argument-based approach within formative classroom assessment scenarios were conducted by Chapelle, Cotos, and Lee (2015) and Ranalli, Link, and Chukharev-Hudilainen (2017). Chapelle et al. validated two online automated writing evaluation systems, the Criteria developed by ETS and the Intelligent Academic Discourse Evaluator (IADE) developed by Iowa State University. The Criteria system was used in less-standardised formative diagnostic assessment to assess ESL university students' grammar-focused writing performances and provide feedback for them to improve their academic writing. Six inferences were proposed to validate the interpretation and use of the Criterion: *domain description, evaluation, generalisation, explanation, extrapolation, utilisation, and ramification*. The IADE was implemented in less-standardised formative diagnostic assessment to assess ESL university students' research paper writing performances and provide feedback for them to revise their writing. Eight inferences were examined to provide validity evidence for the interpretation and use of the IADE, including *domain description, evaluation, generalisation, explanation, extrapolation, utilisation, consequence, and impact*. More recently, Ranalli et al., validated the online Criterion system, used in a less-standardised formative diagnostic assessment context to diagnose ESL university students' academic writing and provide diagnostic feedback for learners to revise their academic writing. In this study, they focused mainly on investigating the evaluation and utilisation inferences to provide backing evidence for the Criteria.

Although the validation frameworks developed by Chapelle et al. and Ranalli et al are potentially useful for guiding the argument-based validation of formative diagnostic

assessment in the language classroom, the focus of their validation was on the diagnostic power of the automatic scoring system and its diagnostic feedback on student learning. Yet, in most typical classrooms, teachers are normally the raters of student performance rather than the scoring technology which is not widely used and available in local classroom contexts. Therefore, the validation frameworks developed by Chapelle et al. and Ranalli et al., though specified for formative diagnostic assessment, were situated mainly within technology-driven assessment contexts. However, most typical formative classroom assessments are not let by such technology, tend to be teacher-mediated, and are more varied, multifaceted and influenced by various contextual factors than technology-driven assessment.

The study by Chapelle and Voss (2014) also highlighted certain challenges of the argument-based approach in the varied and dynamic nature of classroom assessments. In this study, they reviewed research using the argument-based approach and the assessment/use argument approach to validate low-stakes classroom language assessments (Chapelle, Chung, et al., 2010; Koizumi et al., 2011; Pardo-Ballester, 2010) and concluded that the argument-based validation frameworks can generally be used across classroom assessment contexts. Nevertheless, they observed that it is challenging to define the relevant domain of language use in classrooms or other contexts which tend to be dynamic and complex. In such contexts, learning/teaching materials regularly change over the course and thus it is necessary to redesign the test to support that the test scores represent the learning objectives across time, hence making it difficult to develop the test and interpret test results in a classroom assessment. Another challenging issue involves defining the learning construct linked to curriculum which could be interpreted and implemented differently by different teachers with various professional backgrounds and interests. Despite the common curriculum, different teachers may require students to do different tasks under different conditions so as to reach the same learning objectives.

In spite of several studies using the argument-based validation approach in L2 classroom assessment, there are still relatively few studies conducted in ongoing classroom contexts and in particular no studies applied the argument-based or other validation frameworks to validate a diagnostic writing rating scale situated within

formative classroom assessment and particularly in an EFL writing classroom context which is the focus of the current study.

2.4 Rationale for the Current Research

It is clear from the literature that development, implementation, and validation of a diagnostic rating scale in the ongoing classroom context encompass several activities and variables. This section highlights gaps in the literature and the potential areas that the current research could further investigate to extend previous studies and provide further insights into the field.

2.4.1 Gaps in the Literature

The review of the literature has highlighted certain limitations in previous research on development and validation of a diagnostic writing rating scale. It has been suggested diagnostic language assessment should be situated within formative assessment and incorporate such important stages as diagnostic assessment, diagnostic feedback, and remedial intervention (Alderson et al., 2015; Lee, 2015) and involve student self-assessment to optimise the effectiveness diagnostic assessment and feedback (Alderson et al., 2015; Kunnan & Jang, 2009; Lee, 2015). That being said, previous studies on diagnostic assessment of L2 writing focused primarily on development and validation of diagnostic rating scales, which were done on the basis of static and standardised assessment and outside of the classroom context (Kim, 2010; Knoch, 2009b). Although very few (e.g., Wagner, 2015) investigated the diagnostic feedback generated by a diagnostic scale and attempted to incorporate student self-assessment in diagnostic assessment process, the scale was used with standardised assessment tasks and was aimed at detecting learning progress in lieu of supporting day-to-day teaching and learning in an ongoing classroom. It can thus be argued that no research has thus far integrated the key elements, in particular self-assessment, involved in diagnostic assessment and truly implemented a diagnostic rating scale for formative assessment in the real-world classroom, particularly in an EFL classroom context. Alderson et al. (2015) also pointed out that more classroom-based assessment research is needed to examine the processes of diagnostic assessment in order to gain more insights into the interface between such processes and how each might best be actualised in routine classroom

contexts. Clearly, there is a need to further explore the key components of diagnostic assessment in the classroom context and examine how well a diagnostic rating scale works, along with other diagnostic assessment components, in promoting teaching and learning in an ongoing classroom.

In addition, Kane's argument-based approach is driven by the current conceptualisation of validity and validation and has widely been used in previous L2 classroom assessment research (e.g., Chapelle, Chung, et al., 2010; Chapelle et al., 2015; Ranalli et al., 2017). The assessment contexts in those previous studies were rather different from typical formative classroom contexts which are non-standardised, varied, and dynamic in nature. These formative assessment contexts are different form static, high-stakes, and standardised testing, for which the argument-based approach is initiated. With regard to the validation of diagnostic rating scales in previous research, Knoch (2007, 2009a) and Kim (2010) focused mainly on examining validity evidence based on scores obtained from standardised testing and did not consider students' perspectives in the validation. Wagner (2015) did not validate the rating scale as her study focused on investigating the impact of diagnostic feedback generated by the scale. Clearly, there is a need to explore how well the argument-based approach fits into the formative classroom assessment context, which is rather varied, dynamic, and non-standardised, and focuses on learning progression inferred not merely from the test score but also from other sources of information emerging in an ongoing classroom. In addition, there is a call for to elaborate, demystify, and make the approach more practical to wider local assessment practitioners (Cumming, 2013; Knoch, 2016; Sireci, 2016).

2.4.2 Aims of the Study

Building on the existing literature and research with emphasis on the diagnosis stage, this study follows a multisource-driven scale development approach and adopts a binary checklist, both of which are deemed to optimise the diagnostic power and implementation of a diagnostic scale in supporting teaching and learning in the classroom context. In addition, this study uses the terms "*binary rating scale*", "*binary checklist*", or "*binary rubric*" interchangeably to refer to the defined concept of dichotomous diagnostic scoring or judgement, in which "0" represent a non-mastery or unsatisfactory status and "1" represents a mastery or satisfactory status of a writing skill under diagnosis. To enrich

the impact of a rating scale in diagnostic assessment, student self-assessment is integrated alongside the formative diagnostic assessment with the focal aim of promoting students' self-regulated learning skills. The reader is also reminded that this research does focus on investigating teachers' diagnostic feedback and remedial intervention. Thus, the extent to which such factors influence the impact of the scale in the ongoing diagnostic assessment remain underexplored in this research. In terms of instrument validation, the present study follows Kane's argument-based approach to validation, which is driven by the current validity theory, viewed as a flexible framework, and widely used in educational and psychological fields and various assessment contexts. This study focuses the validation on the score-based interpretation and use of a diagnostic scale and partly on its impact, alongside student self-assessment in formative diagnostic assessment, on teaching and learning in the classroom. As well as the score-based evidence, both teachers' and students' perspectives are considered in validation of classroom assessment.

To this end, this mixed methods research set out to develop and validate a diagnostic writing rating scale for formative assessment in a Thai EFL university writing classroom. The scale is intended to identify students' strengths and weaknesses in academic writing products and to support teaching and learning. The present study could shed further light on (a) the design and implementation of a diagnostic language assessment system to bring about its optimal impact in the real-world language classroom, (b) the contribution of the multisource-driven approach and the binary checklist to the identification of the EFL writing construct and the diagnostic power of a diagnostic rating scale in a Thai EFL classroom context, and (c) the applicability of Kane's argument-based approach in the context of formative classroom assessment. Four research questions are addressed in this research.

1) To what extent does the diagnostic rating scale function appropriately for the formative diagnostic assessment in the EFL university writing classroom?
2) To what extent does the diagnostic rating scale function consistently for the formative diagnostic assessment in the EFL university writing classroom?
3) To what extent does the diagnostic rating scale support formative decisions about teaching and learning in the EFL university writing classroom?
4) To what extent does the formative diagnostic assessment have beneficial consequences for teaching and learning in the EFL university writing classroom?

The research questions addressed above are also linked with the argument-based approach in that findings for these questions are used as backing evidence for the overarching validity argument for the scale. The first question investigates the appropriateness of the scale functioning and hence seeks empirical findings to justify the evaluation, explanation, and extrapolation inferences. The second research question examines the consistency of the scale functioning and thus seek empirical findings in justification of the generalisation and explanation inferences. The third question examines the usefulness of the scale to support decisions about teaching and learning and search for empirical findings justifying the decision inference. The fourth question probed into the utilisation and consequence of the scale scores and formative diagnostic assessment for teaching and learning and therefore looks for empirical results for justifying the consequence inference. The first and second research questions receive more attention whereas the third and fourth questions attract somewhat less attention and indeed need further investigations, which is beyond the scope of this research. In the next section, I present the assessment claims or the intended interpretations and uses of the current assessment, which are laid out in the IUA framework. The IUA structure needs to be clearly stated at an initial stage as it directs the data collections and analyses which could be accomplished within the current research.

2.5 Framework of the Current Interpretive and Use Argument

In this study, the intended purposes of the scale are generally to support the formative diagnostic assessment in promoting ongoing learning and teaching in the EFL university writing classroom, and to provide scores which can be interpreted as strengths and weaknesses of the defined academic writing ability, and used to inform formative decisions leading to beneficial consequences or impacts on teaching and learning.

As displayed in Figure 2.3, the IUA structure, adapted from Chapelle et al. (2008), lays out the process, starting from the TLU domain to the observed performances, the observed performances to the scale scores, the observed scores to the defined construct and to the expected or universe scores, the universe scores to the target scores, and the scores to the decision and consequences. These steps are connected through seven inferences: *(1) domain description, (2) evaluation, (3) generalisation, (4) explanation, (5) extrapolation, (6) decision, and (7) consequence*. The domain description inference is

associated with the development and administration of the scale. The evaluation, generalisation, explanation, and extrapolation inferences are specifically related to the score interpretation. The decision and consequence inferences concern the score use.

Figure 2. 3 *The Current Interpretive and Use Argument Structure*

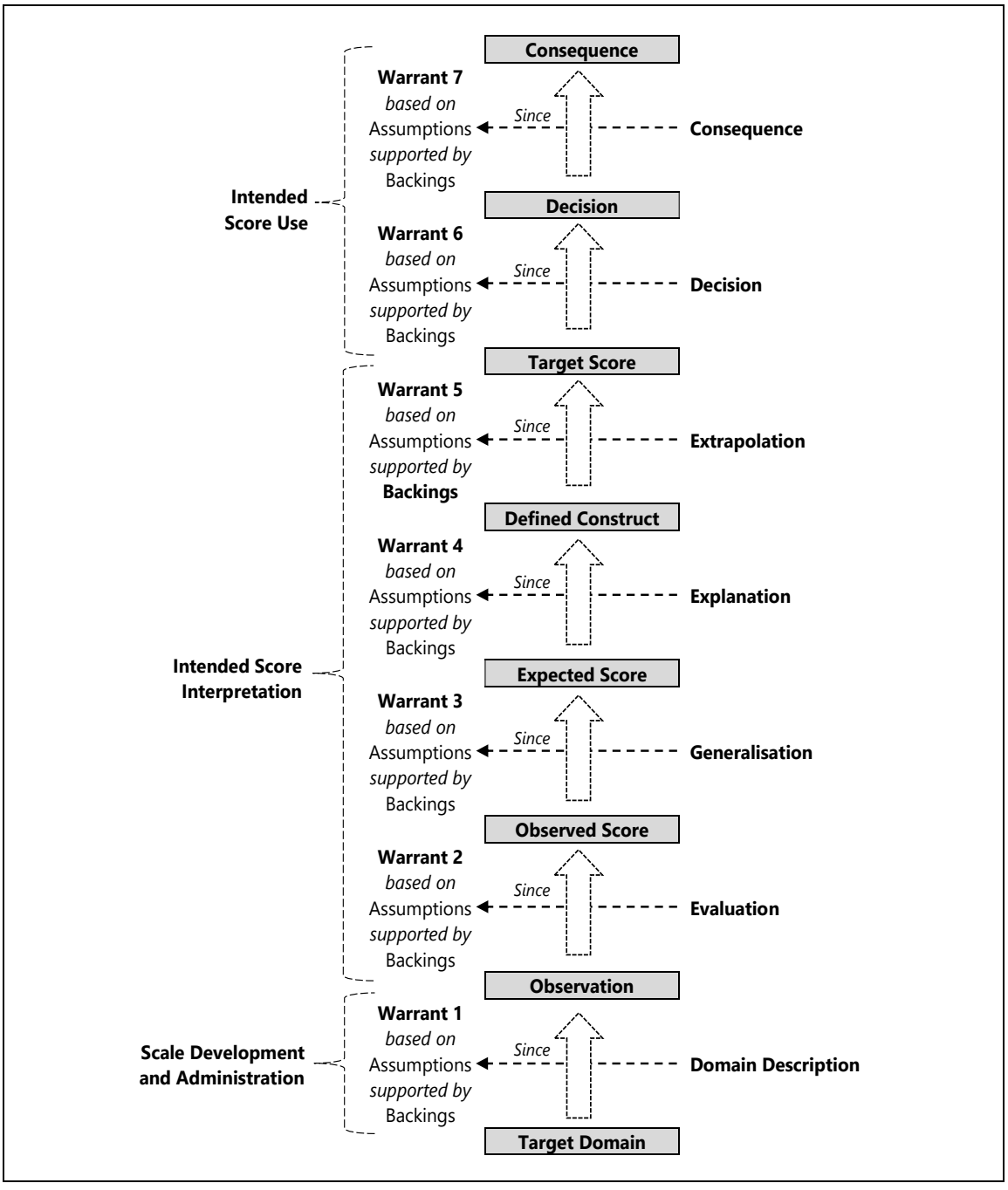


Table 2.8 summarises the inferences with their warrants each resting on the underlying assumptions pointing to the expected sources of empirical evidence needed to be investigated and collected in this research.

Table 2. 8 *Inferences, Warrants, Assumptions, and Expected Backing*

Inferences	Warrants	Assumptions	Backing
1) Domain description	The scale criteria represent academic writing ability and skills in student writing performances and learning contents in the TLU domain of EFL university classroom.	1. The expected academic writing quality features, writing skills, and learning contents in the classroom can be identified.	• Scale development
		2. The characteristics of writing assignment tasks in the classroom can be identified.	• Scale development and administration
2) Evaluation	The scale provides observed scores reflective of the academic writing ability and skills in student writing performances in the classroom.	1. The rating format is appropriate to assess the strengths and weaknesses of the student writing ability.	• QCA
		2. The scale shows acceptable psychometric properties to ensure accurate functioning.	• CTT and MFRM analyses
		3. The raters are positive about the scale functioning.	• QCA
		4. The raters go through appropriate rater training and rating procedures.	• Scale development and administration
		5. The raters show acceptable psychometric properties to ensure appropriate rating behaviours.	• MFRM analysis
3) Generalisation	The scale provides observed scores as estimates of the expected scores across raters and student writing performances in the classroom.	1. The scale shows acceptable psychometric properties to ensure consistent functioning.	• CTT and MFRM analyses
		2. The raters show acceptable psychometric properties to ensure consistent rating behaviour.	• CTT and MFRM analyses
4) Explanation	The scale provides observed scores as estimates of the expected scores attributed to the defined academic writing construct required in the classroom	1. The diagnostic scores are internally consistent with the defined writing construct	• CTT and MFRM analyses
		2. The diagnostic scores reflect the academic writing skills learned and assessed in the classroom	• QCA
5) Extrapolation	The scale provides diagnostic scores accounting for the quality of the student academic writing ability on other tasks in the classroom.	1. The diagnostic results distinguish between low-, mid-, and high achieving students.	• ANOVA analysis
		2. The diagnostic results have a positive relationship with student learning achievement.	• Correlation analysis
6) Decision	The scale is useful to support formative decisions about teaching and learning in the classroom.	1. The scale is practical for teachers and students in the ongoing classroom.	• QCA
		2. The scale provides diagnostic information meaningfully interpretable by teachers and students.	• QCA
		3. The scale provides useful diagnostic information to inform teachers' and students' formative decisions about teaching and learning.	• QCA
7) Consequence	The scale-driven assessment has beneficial consequences on teaching and learning in the classroom.	1. The scale provides diagnostic information to improve teacher instruction and feedback.	• QCA
		2. The scale supports self-assessment in promoting student self-regulated learning.	• ANOVA and correlation analyses and QCA
		3. The assessment system promotes student learning progression	• Scale score analyses and QCA
		4. The assessment system contributes to student learning achievement.	• Regression analysis
		5. The assessment system has potential positive impacts on teachers' and students' academic development.	• QCA

The domain description inference, linking the performance in the TLU domain to the sample of the criteria on the scale and the observed performances, rests on the warrant that the scale criteria represent academic writing ability and skills in student writing performances and learning contents in the TLU domain of EFL university classroom. This warrant depends on two assumptions: (1) the expected academic writing quality features, writing skills, and learning contents in the classroom can be identified, and (2) the characteristics of writing assignment tasks in the classroom can be identified. Backing evidence for these assumptions is tied to research procedures during the conceptualisation, development, revision, and administration of the scale.

The evaluation inference which claims that the scale provides observed scores reflective of the academic writing ability and skills in student writing performances in the classroom. This warrant depends on five assumptions: (1) the rating format is appropriate to assess the strengths and weaknesses of the student writing ability, (2) the scale shows acceptable psychometric properties to ensure accurate functioning, (3) the raters are positive about the scale functioning, (4) the raters go through appropriate rater training and rating procedures, and (5) the raters show acceptable psychometric properties to ensure appropriate rating behaviours. Evidentiary backing for these inferences will be partly associated with the scale development and largely obtained from CTT and MFRM analyses of scale scores and a qualitative content analysis (QCA) of user perceptions.

The generalisation inference, connecting the observed scores to the estimate of the expected score, states that the scale provides observed scores as estimates of the expected scores across raters and student writing performances in the classroom. This warrant depends on two assumptions: (1) the scale shows acceptable psychometric properties to ensure consistent functioning, and (2) the raters show acceptable psychometric properties to ensure consistent rating behaviour. Empirical evidence for these assumptions will be investigated through CTT and MFRM analyses of scale scores.

The explanation inference, linking the expected scores and the construct of the scale, claims that the scale provides observed scores as estimates of the expected scores attributed to the defined academic writing construct required in the classroom. This warrant relies on two assumptions: (1) the diagnostic scores are internally consistent with the defined writing construct, and (2) the diagnostic scores reflect the academic writing skills learned and assessed in the classroom. Empirical backing for the first and second

assumptions will be obtained from CTT and MFRM analyses of scale scores and a QCA of user perceptions respectively.

The explanation inference, linking the expected scores and the construct of the scale, claims that the scale provides observed scores as estimates of the expected scores attributed to the defined academic writing construct required in the classroom. This warrant relies on two assumptions: (1) the diagnostic scores are internally consistent with the defined writing construct, and (2) the diagnostic scores reflect the academic writing skills learned and assessed in the classroom. Empirical backing for the first and second assumptions will be obtained from CTT and MFRM analyses of scale scores and a QCA of user perceptions respectively.

The extrapolation inference, connecting the construct of the scale to other criteria of language ability, rests on the warrant that the scale provides diagnostic scores accounting for the quality of the student academic writing ability on other tasks in the classroom. This warrant rests on two assumptions: (1) the diagnostic results distinguish between low-, mid-, and high achieving students, and (2) the diagnostic results have a positive relationship with student learning achievement. Backing evidence for the first and second assumptions will be obtained from ANOVA and correlation analyses of scale scores respectively.

The decision inference, linking the scores to make meaningful decisions and beneficial consequences in the real use, claims that the scale is useful to support formative decisions about teaching and learning in the classroom. This warrant relies on three assumptions: (1) the scale is practical for teachers and students in the ongoing classroom, (2) the scale provides diagnostic information meaningfully interpretable by teachers and students, (3) 3.The scale provides useful diagnostic information to inform teachers' and students' formative decisions about teaching and learning. Empirical evidence associated with these assumptions will be examined through a QCA of user perceptions.

The consequence inference rests on the warrant that the scale-driven assessment has beneficial consequences on teacher instruction and student learning in the classroom. This warrant relies on five assumptions: (1) the scale provides diagnostic information to improve teacher instruction and feedback, (2) the scale supports self-assessment in promoting student self-regulated learning, (3) the scale promotes student learning progression, (4) the scale contributes to student learning achievement, and (5) the scale

has potential positive impacts on teachers' and students' academic development. Empirical evidence associated with these assumptions will be examined through psychometric and statistical analyses of score scores and a QCA of user perceptions.

Only the assumptions and evidence that could be investigated in this research are specified and the IUA framework is proposed and iteratively revised from the perspective of the scale developer or researcher. The assumptions and evidentiary sources presented in the IUA are adequate and appropriate to validate the proposed interpretation and use of the scale scores in this study. As can be seen, some inferences deserve more attention and are based on strong assumptions, thus requiring more and strong evidence, while others are not the focus of this research and based on basic assumptions and thereby may not require much and strong evidence in this research. The IUA not only points to several sources of validity evidence to be documented and empirically investigated but also serves as the blueprint for the present mixed methods research in developing the scale and formative diagnostic assessment system, while at the same time accumulating both evidentiary sources to support the overarching validity argument of the scale.

2.6 Chapter Summary

This chapter has presented the relevant literature underpinning the current scale development and validation research. I began by looking at principles and characteristics of diagnostic language assessment, which provides information as to how a diagnostic tool should be designed and developed and how a diagnostic language assessment in the classroom context should be framed and implemented to reach its optimal impact on learning. I then reviewed types of rating scales and approaches to scale development that are most suitable to inform the current scale development in order to achieve the intended diagnostic assessment purposes. This chapter then explored theoretical models of L2 writing ability to generate theoretically-informed aspects of L2 writing construct which partly inform the L2 writing construct in question. After that, I reviewed the current conceptualisation of validity and validation, which has shifted from a focus principally on test-internal psychometric evidence to a broader concern with the meaningful interpretation and use of assessment results. Particular attention was paid to Kane's argument-based approach to validation, which is deployed as the theoretical framework for the current research. After reviewing all the relevant literature, I then presented the

rationale for this research and the intended interpretations and uses of the scale in the current formative diagnostic assessment via the IUA framework, which directs the current study. In the next chapter, I will describe the current research methodology which includes the rationale for the use of the mixed methods research methodology, the scale development procedures, data collections, and data analyses over the three study stages with emphasis on the scale implementation stage.

Chapter 3: Methodology

This chapter presents the research methodology underlying the scale development and validation. In this chapter, I begin with an overview of the research objective and research design before highlighting why the mixed methods research methodology adds greater value to the current scale development and validation research. Afterwards, I present a brief overview of the multistage exploratory sequential design adopted in this mixed methods research before describing the scale construction, trialling, and implementation study stages. In particular, the focus of the methodology description is on the scale implementation stage which is the main study of interest. Following this, a brief review of the research questions is presented in the scale implementation stage to illustrate the interface between the mixed methods research methodology and the argument-based validation framework in which this study is situated.

3.1 Overview of the Research Objective and Research Design

This PhD research aimed to develop a rating scale for a formative diagnostic assessment in order to support teachers' instruction and students' teaching as key stakeholders or scale users, and explore to what extent the assessment results generated by the scale were interpreted and used appropriately, and the scale-driven formative diagnostic assessment leads to the intended beneficial consequences for teacher instruction, student learning, and academic achievement.

The scale was driven by a diagnostic language assessment approach (Alderson et al., 2015; Jang, 2012, Knoch, 2009b, 2011; Lee, 2015) and a multisource approach to scale development (Banerjee et al., 2015; Knoch, 2009b, 2011) and was designed specifically for a formative classroom assessment in a Thai EFL university setting. In the classroom, the scale was aimed at diagnosing students' strengths and weaknesses on teacher-made assignment tasks involving a five-paragraph essay. A student self-assessment was also incorporated in the assessment with a view to encouraging students to regularly engage with learning and acquire the independent learning skills and strategies necessary for them to progressively improve learning and achieve learning goals.

To this end, this research employed a mixed methods methodology with a multistage exploratory sequential design (Creswell & Plano Clark, 2018) drawing on a wide range of qualitative and quantitative data to inform the scale development and validation.

3.2 Rationale for the Mixed Methods Research Methodology

The mixed method research design adds more value to the current scale development and validation research for several reasons. To begin with, since there are no existing diagnostic scales used in the assessment context of interest, there is a need to develop a novel scale that meets the intended assessment purposes and fits with the language goals of the Thai university writing classroom. To develop a classroom diagnostic writing scale, the relevant literature suggests that multiple sources of data, including theoretical frameworks, expert intuition, and contextual data, can be used in a complementary fashion to ensure that the assessment scale and criteria are well grounded and comprehensive (Bachman & Palmer, 2010; Brown, 2012; Davis, 2015; Knoch, 2011; Lee, 2017; Weigle, 2002). These data sources are qualitative in nature and thus require a constructivism-driven qualitative approach to exploring, analysing and generating findings which can be subsequently built into the scale construction (Creswell & Plano Clark, 2018; Creswell & Shou, 2016; Greene, 2007; Ziegler & Kang, 2016).

In addition, both development and validation of scientific measurement instruments require a cyclical and iterative process through which several activities need to be carried out all the way through to guide and support one another in order to achieve a well-developed test and sound assessment (Guetterman & Salamoura, 2016; Saville, 2016). In addition, current validation requires an ongoing process by which test developers document test development procedures and accumulate empirical evidence over time to argue in support of validity (Kane, 2013). Therefore, both scale development and validation, in nature, call for a mutually inclusive and lengthy process that operates over an extended span of time, thereby necessitating multiphase research.

Also of necessity during a test development process is to trial and modify a test to maximise its comprehensibility and applicability prior to the actual operational stage (Guetterman & Salamoura, 2016; Kenyon & MacGregor, 2012). This pre-operational stage should thus be an integral part of test development research (Kenyon & MacGregor, 2012). Accordingly, this research incorporates the scale trialling stage in order to trial the

draft scale to ensure that it is ready to be operationally implemented. By adding the trialling stage, I am able to optimise the scale quality, cross-validate findings, and seek evidence to ensure sound development and validation procedures as well as effective implementation of the scale in the classroom.

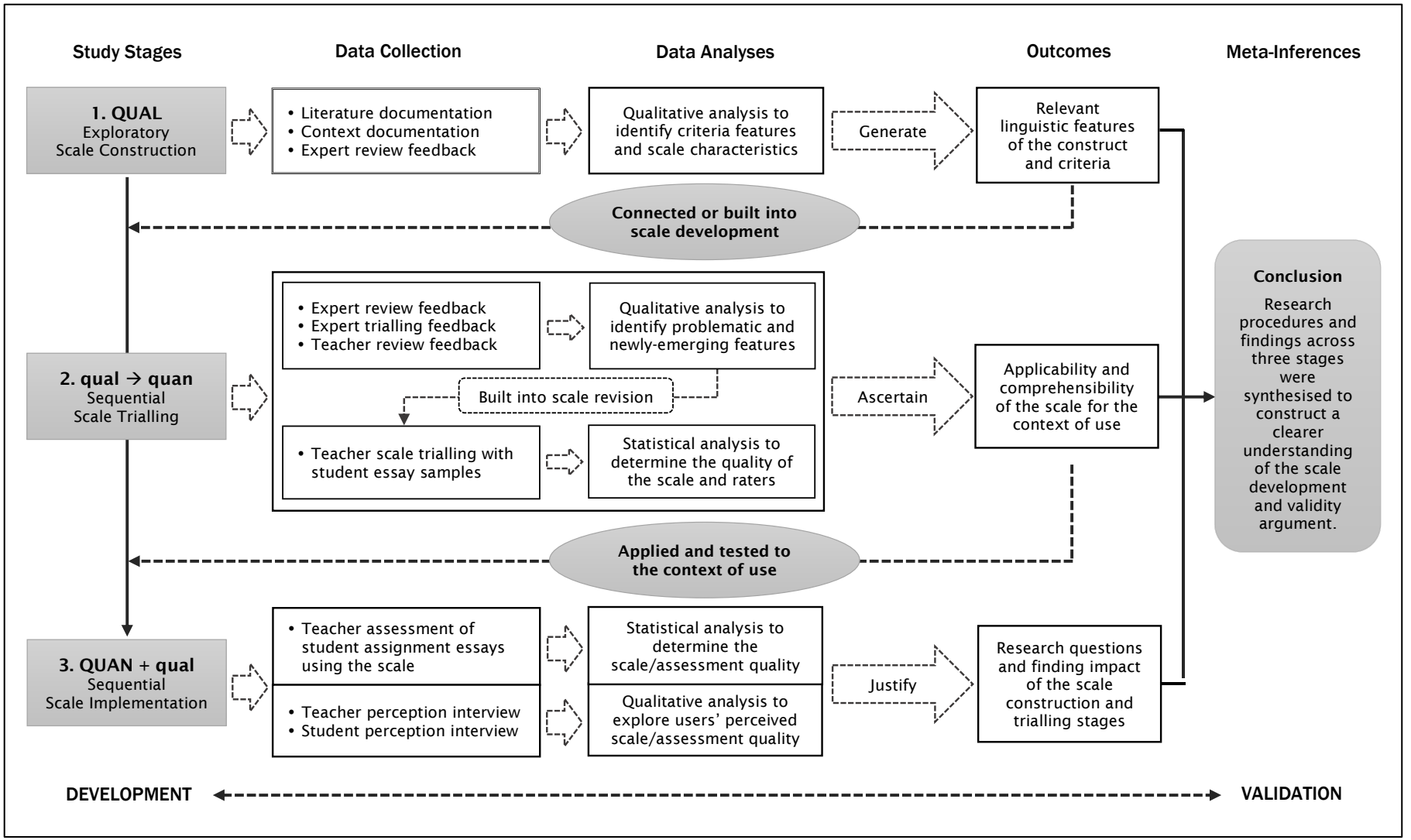
It is also well established that assessment validity requires not only score-driven quantitative evidence but also stakeholders' perceptions of an assessment tool to support meaningful interpretation and use of assessment outcomes (Kane, 2013). Clearly, neither qualitative nor quantitative methodology offer a sufficient basis for developing and validating the scale.

Propelled by a multisource-driven scale development and Kane's argument-based validation approaches, I take a pragmatist view that only through combining the constructivist qualitative and post-positivist quantitative ways of seeking knowledge am I able to gain more profound understanding and more solid conclusion of the phenomena under study. I thereby adopted a multistage mixed methods research, for it draws upon multiple complementary worldviews and is thus a pluralism-oriented methodology that leads to logical and convincing research answers as well as well-documented and triangulated validity evidence. Specifically, a multistage exploratory sequential design (Creswell & Creswell, 2018; Creswell & Plano Clark, 2018) is employed as it fits into the practicality and characteristics of this scale development and validation research where qualitative and quantitative data are collected and analysed sequentially over an extended period of time to arrive at sound and rigorous findings (Creswell & Creswell, 2018; Creswell & Plano Clark, 2018; Riazi, 2017).

3.3 Overview of the Current Research Design

Figure 3.1 portrays a procedural diagram of the multistage exploratory sequential design in this mixed methods research. In this study, capital- and small-letter abbreviations represent greater or lesser emphasis of a particular approach respectively. This three-stage research included (1) scale construction stage, (2) scale trialling stage, and (3) scale implementation stage, each of which will be described in more detail later.

Figure 3. 1 *Procedural Diagram of the Multistage Exploratory Sequential Mixed Methods Research Design*



The scale construction stage set out to extract salient features representing the writing construct in the classroom context. To achieve this aim, this stage adopted a qualitative approach to explore a representative range of data from theoretical, intuitive, and contextual sources for generating features subsequently used to determine the criteria domains and descriptors as well as the properties of the initial draft scale.

The subsequent trialling stage aimed to seek preliminary evidence to ascertain that the scale appropriately served its intended purposes without further modification and was ready to be operationally implemented. This trialling stage involved two sequential phases. The first phase was a qualitative trial aiming to explore areas for further modification of the initial draft scale from views of context-external experts and local teacher stakeholders. The second phase was a quantitative trial seeking psychometric evidence to confirm the functionality and comprehensibility of the revised draft scale with emphasis on teachers' rating agreement and consistency.

Central to this research was the scale implementation stage. This stage aimed to provide strong argument-based evidence to justify, based on the scale users' perception and solid psychometric evidence, that the operationalised scale provided assessment information that was meaningfully interpreted and used as it was intended. This stage involved quantitative and qualitative methods. A quantitative method aimed to ascertain the psychometric quality of the scale while a qualitative approach was intended to explore the scale quality, usefulness, and impact from the users' perspectives.

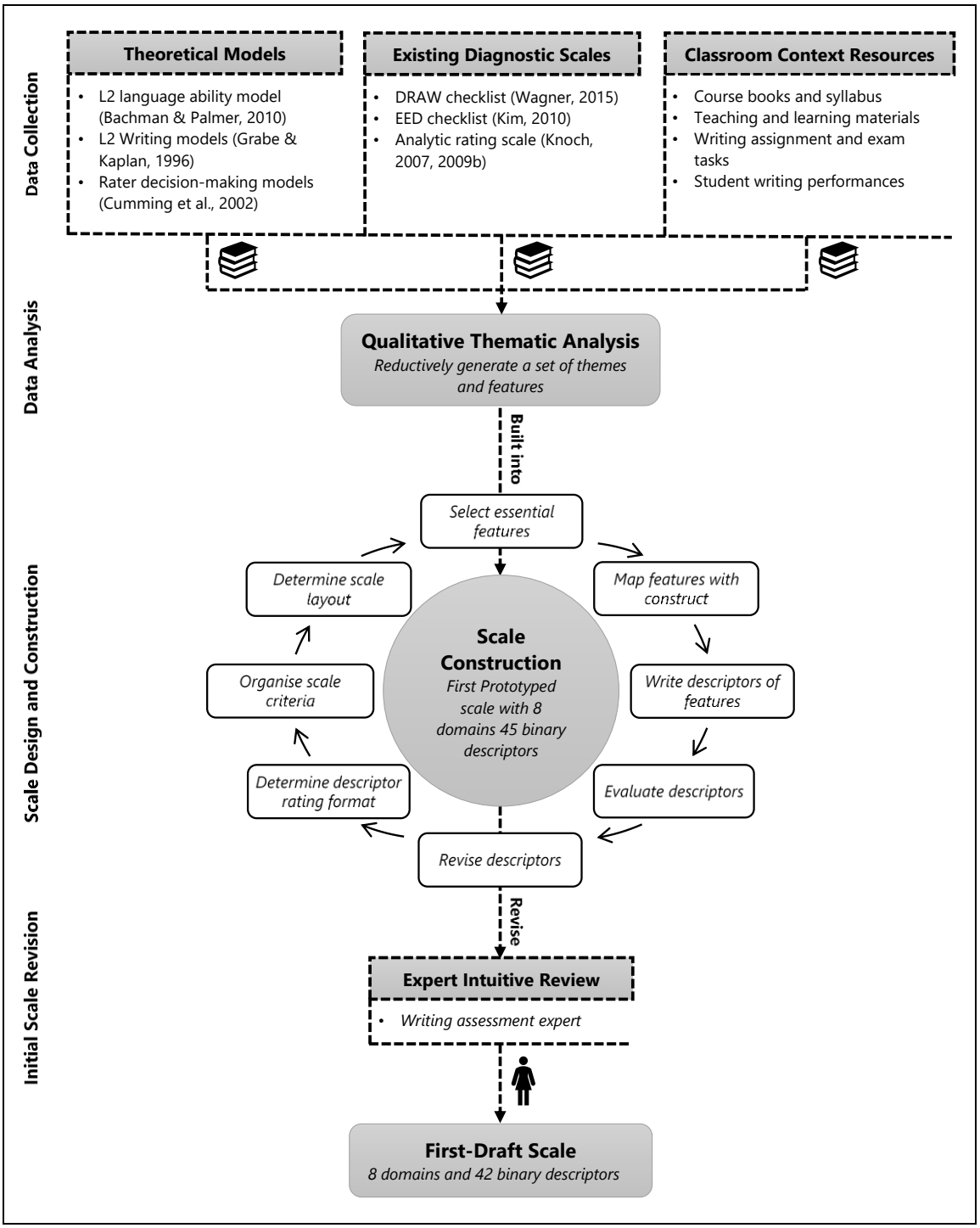
All procedural, quantitative, and qualitative evidence documented and aggregated across the study stages were ultimately integrated and synthesised in order to make meta-inferences or conclusions regarding the overall validity argument for the scale.

3.4 Scale Construction Stage

Figure 3.2 displays a procedural diagram of the scale construction stage. This stage aimed to design and construct the draft scale by drawing on a qualitative approach to collecting and reductively analysing an array of qualitative data to arrive at a small set of construct components and product-based writing quality features. In particular, this stage focused on conceptualising and constructing the draft scale to ensure that: (a) the writing construct in question was appropriately described and decomposed, (b) the scale criteria were appropriately extracted and structured to suit the intended diagnostic assessment

purposes and represent the contents and writing skills taught and assessed in the classroom context, (c) the scale incorporated appropriate properties regarding scale format, descriptor wording, scoring format, criteria organisation, and overall scale layout, and (d) the scale would be appropriately applied in correspondence with the learning, teaching, and assessment practices in the classroom context.

Figure 3. 2 *Procedural Diagram of the Scale Construction Stage*



3.4.1 Participant

The participant in the scale construction stage was an expert of language and writing assessment who was a professor of applied linguistics in a public Australian university. The expert was asked to review and comment on the prototyped scale. The participant was provided with plain language statement and consent forms.

3.4.2 Qualitative Data Collection

In the exploratory scale construction stage, I conducted a literature review to figure out how the construct in question should appropriately be defined and what sources of data should inform the construct, criteria, and scale development. As a result, four main sources of qualitative data were purposively compiled, namely (1) theories of L2 language and writing ability, (2) existing diagnostic writing scales, (3) curriculum resources, and (4) expert/teacher intuition. These sorts of data were selected to inform the construct of classroom diagnostic writing assessment underlying the scale (Bachman & Palmer, 2010; Knoch, 2011; Weigle, 2002). At this stage, the research ethic application was still in the process. Consequently, this study could not collect data from teachers in the context to inform the scale construction.

The theories included L2 language ability, text construction and L2 writing knowledge, and rater decision-making behaviour and these theoretical models are suggested to inform the construct of L2 writing (Knoch, 2011; Weigle, 2002). The existing diagnostic writing scales included the Empirically-derived Descriptor-based Diagnostic (EDD) checklist (Kim, 2010), the analytic diagnostic rating scale of writing (Knoch, 2007, 2009a, 2009b), and the Diagnostic Rubric for Assessing Writing (DRAW) (Wagner, 2015). These scales were chosen as they were developed for diagnostic writing assessment purposes and should thus provide useful information to inform the current scale development. The theories and existing scales were aimed at informing the construct components.

To optimally contextualise the scale, a range of materials in the classroom context, including the course syllabus, coursebooks, learning materials, writing tasks, assessment criteria, and student writing performances, was compiled to inform the scale construct, criteria, and implementation. To ensure appropriate application and interpretation of the scale before its trialling, an expert of language and writing assessment was asked to review

the contents and properties of the draft scale and provide intuitive written feedback to revise the draft scale.

3.4.3 Qualitative Data Analysis

Following a qualitative content analytic approach (Schreier, 2012, 2014), I comprehensively reviewed the theoretical, existing scale, and contextual data to identify salient themes and features that should inform the construct constituents and learning contents in the context. Having experience in teaching the writing course in the context, my professional intuition helped to identify the themes and features representing the writing skills and learning contents in the context. To avoid exercising to much of my intuition in the data interpretation, I tried, to the extent possible, to maintain a systematic, iterative, and reductive analytic process and compare and triangulate information from theoretical, contextual, and intuitive sources. In addition, specific criteria were set in the qualitative content analysis in order to arrive at the themes and features representing the target writing skills and learning contents. First, the target features of interest must be measurable from a single writing product. Second, the features should essentially represent those exiting in the actual assessment context, including for example, students' writing productions, learning contents, and rater' perceived criteria.

3.4.4 Findings and Scale Construction

Applying the qualitative content analysis described in the previous section, a number of features were systematically extracted to create a final set of construct components and writing quality features. A pool of themes and features generated in this initial exploratory qualitative stage were organised into sets of criteria domains and sub-skills of productive writing quality, which were in turn built into the initial draft scale. To ensure the comprehensibility and applicability of the scale, the scale was worded, organised, ordered, formatted, and revised by the researcher through an iterative process.

As it was impossible to include and measure all the linguistic aspects presented in the theories and models, I selected only the linguistic components that are appropriate to account for the L2 writing construct in question which should be rich, parsimonious, and well fit to the assessment purpose and the target assessment situation (Chalhoub-Deville, 1997; Fulcher & Davidson, 2007; Jamieson, 2014; McNamara, 1996).

The prototyped scale included eight construct domains or categories, measured by 45 writing quality indicators or descriptors which were driven by the contextual curriculum. As presented in Table 2.9, I decomposed the L2 writing construct into eight dimensions, including (1) organisation, (2) coherence or unity, (3) cohesion, (4) content, idea, or topic knowledge, (5) grammar, (6) sentence, (7) vocabulary, and (8) mechanics. The terms used to label the categories are consistent with the existing scales and should be understandable and interpretable for teachers and student in the assessment context.

Overall, the theoretical models and existing scales include virtually all categories of the current construct. However, Bachman and Palmer's (2010) communicative language ability model does not describe explicitly about the unity or coherence and mechanics of writing and Grabe and Kaplan's (1996) text construction and writing knowledge models do not have any features referring to writing mechanics. It appears that Cumming et al.'s (2002) rater decision-making behaviour model and two existing scales (Kim, 2010; Wagner, 2015) include features representing all of the classified categories, while Knoch's (2007, 2009b) rating scale does not have the organisation.

Table 3. 1 *Domains of the writing Construct Informed by Theories and Existing Scales*

Construct domains	Theories				Existing scales		
	CLA	TC	WK	RDMB	Knoch	Kim	Wagner
1. Organisation	✓	✓	✓	✓	-	✓	✓
2. Coherence	-	✓	✓	✓	✓	✓	✓
3. Cohesion	✓	✓	✓	✓	✓	✓	✓
4. Content	✓	✓	✓	✓	✓	✓	✓
5. Grammar	✓	✓	✓	✓	✓	✓	✓
6. Sentence	✓	✓	✓	✓	✓	✓	✓
7. Vocabulary	✓	✓	✓	✓	✓	✓	✓
8. Mechanics	-	-	-	✓	✓	✓	✓

Note. CLA = communicative language ability; TC = text construction; WK = writing knowledge; RDMB = rater decision-making behaviour

It should be noted that the definition and categorisation of the construct may differ from test purposes, testing context, and test developers' judgement. In this study, the organisation domain refers to the structure of the essay which includes the introduction, main body, and conclusion paragraphs and the ways in which the ideas are organised in each of these essay structures. The coherence domain refers to the ways in which the ideas are presented to support the main ideas of individual paragraphs and a

single topic of the essay. The cohesion domain is defined as the ways in which sentences, paragraphs, and essay are connected using appropriate linguistic devices (e.g., transition signals, conjunctions, or connectors). While the coherence focuses on the connection and unity of the ideas, the cohesion focuses on the use of linguistic devices to connect the ideas. Wagner (2015) included coherence and cohesion as part of the organisation category. Yet, separating coherence and cohesion as different categories from the organisation may provide more specific information on learners' writing abilities. The content category refers to the overall themes or ideas that are presented to the readers and in response to the writing topic or prompt of the essay. The grammar domain is defined as the ways in which students use specific grammatical rules (e.g., tense, passive voice, part of speech) to form texts in order to express ideas. The sentence domain refers to the construction and combination of various types of sentences to express ideas and the vocabulary domain refers to the use of various vocabulary to convey meaning appropriately in different contexts. The mechanics category refers to the technical and structural conventions that constitute the overall accuracy and meaning of a text.

The prototyped scale was then reviewed by a writing assessment expert, who was asked to consider the appropriateness and clarity of the scale criteria and properties and to write any comments on the scale paper. The comments were afterwards used to revise the prototyped scale. Ultimately, the final draft scale contained the same construct domains, while the descriptors were revised and reduced into 42 descriptors (see Appendix A).

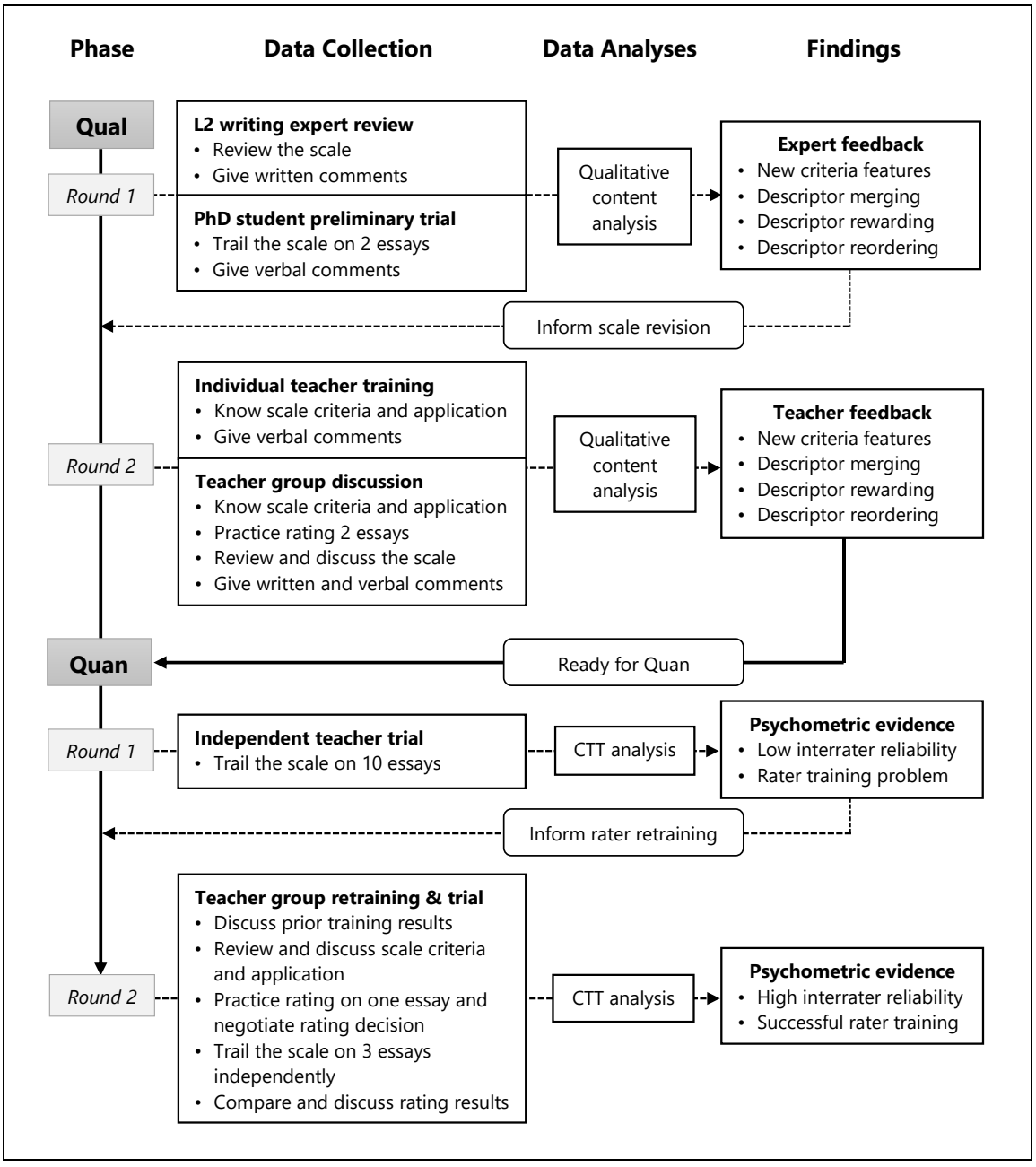
In short, throughout this qualitative exploratory stage, the scale was theoretically, intuitively, and contextually shaped to ensure sound diagnostic assessment in the classroom context. At the same time, initial validity evidence was accrued to support the assumptions underlying the warrant of the domain description inference which links the scale construct and criteria with those in the TLU classroom domain. The first-draft scale was subsequently trialled and modified in the scale trialling stage.

3.5 Scale Trialling Stage

The sequential scale trialling stage aimed to qualitatively and quantitatively trial and revise the first-draft scale to maximise its comprehensibility, comprehensiveness, and applicability for the assessment in the context. This stage involved two rounds of

qualitative trialling and two rounds of quantitative trialling. Findings of the scale trialling stage served to confirm the readiness of the scale for implementation. Figure 3.3 portrays a procedural diagram of the scale construction stage.

Figure 3. 3 *Procedural Diagram of the Scale Trialling Stage*



3.5.1 Participants

Participants in this stage consisted of an expert of L2 writing in Australia, two advanced EFL PhD students of applied linguistics in Australia, and five classroom teachers in the Thai classroom context. The L2 writing expert and PhD students were considered as

context-external experts. The L2 writing expert was a female lecturer of English as a second language with a specialisation in L2 writing. One PhD student was a Chinese female student who had several years of experience in teaching EFL students. The other student was a Saudi Arabian male teacher who had a wealth of experience in teaching English and specifically writing in a Saudi Arabian university. Five classroom teachers, who were to teach and apply the scale in the academic writing courses, were recruited through a quota-convenient sampling method. They also subsequently participated in the quantitative trialling phase. At the beginning of this stage, participants were provided with plain language statement and consent forms. The teachers' pseudonymous names and demographic information are presented in Table 3.2.

Table 3. 2 *Classroom Teachers' Demographic Information*

Name	Age	Gender	L1	L2	Highest education	Essay teaching experience
Sara	31	Female	Thai	English	MA English Teaching	2 semesters
Nana	33	Female	Thai	English	MA English	2 semesters
Ivy	48	Female	Thai	English	MA English Teaching	1 semester
Ken	41-45	Male	Thai	English	MA English Teaching	5 semesters
Cali	36-40	Female	Thai	English	MA Applied Linguistics	1 semester

3.5.2 Instruments

The instruments used in the scale trialling stage included the first-draft diagnostic rating scale (see Appendix A) and the scale evaluation form (see Appendix D). The scale evaluation form was developed to evaluate whether the wording of the descriptors is clear and meaningful, the categories and descriptors cover essential aspects of the L2 writing construct, and the ordering of the categories and descriptors is easy to follow. The form also provided a space for experts to write additional comments. Participants were encouraged to provide any comments they could think of.

3.5.3 Qualitative Trialling Phase

The qualitative trialling phase consisted of two round trails and was aimed at eliciting external and local experts' perceptions of the scale comprehensibility, comprehensiveness, and applicability to inform the scale revision.

3.5.3.1 First-Round Data Collection, Data Analysis, and Findings

The first-round qualitative trial was to elicit context-external experts' feedback to improve the scale quality. In this round, the L2 writing expert used the scale evaluation form to evaluate the first-draft scale in terms of language clarity, construct coverage, and criteria organisation, and then provided written feedback. Meanwhile, I and the PhD students of applied linguistics used the first-draft draft scale to rate two samples of student essays and then together discussed the applicability and comprehensibility of the draft scale. I myself made notes on the scale and two PhD students gave verbal comments which were recorded and noted. This session took about one hour. Verbal comments were transcribed into texts and combined with written comments.

Following the qualitative analytic approach, I manually reviewed all the qualitative textual comments to identify themes and features for revising the first-draft scale. The L2 writing and PhD student experts generally perceived that the scale layout and application were appropriate and the language was clear. Yet, there were some problems with the descriptor wording and criteria. All the experts thought that it was subjective and difficult to decide if an introduction paragraph was attractive or interesting and well-written texts do not necessarily contain this feature. They also perceived that "*main idea uniqueness*" may not be necessary and "*concluding sentence*" is not necessary since not all paragraphs necessarily have a concluding sentence. The PhD student experts also added that "*content redundancy*" may not be necessary.

In addition, all the experts generally suggested that certain descriptors which looked similar and assessed similar skills should be combined into a single descriptor and some descriptor that may not be necessary should be excluded. They also added that it was at times difficult to judge "*supporting idea adequacy*" when students provided sufficient supporting ideas for some paragraphs but not for others in an essay. Furthermore, words, such as "*logically*", "*final thought*", "*accurately*", "*various words*", and "*sentence type*", on the descriptors were subjective to define and difficult to judge. Two PhD students also suggested adding more rating options for the grammar descriptors and perceived that "*sentence types*", "*passive voice*", and "*collocations*" were not always necessarily used in all essays, depending on the topic and context. They also proposed that "*text length*" and "*spelling*" were important skills and thus should be assessed in writing.

Based on the three context-external experts' comments, four main decisions were made on the scale revision: descriptor deletion, descriptor combining, descriptor rewording, and new feature inclusion. First, five descriptors (*introduction paragraph interestingness, paragraph concluding sentence, concluding sentence paraphrase, main idea uniqueness, and content redundancy*) were excluded from the scale since the experts suggested they were not necessary. Second, voice and tense descriptors were merged into the same descriptor as they are highly related. Third, some descriptors including subjective words (*logically, accurately, various words, and sentence type, etc.*) were reworded for greater clarity and more objective interpretation. Finally, two more descriptors were developed to assess the text length and spelling skills, as proposed by the PhD student experts. After the first-round qualitative trial, 42 descriptors were revised and reduced to 37 descriptors in the revised draft scale (see Appendix B).

3.5.3.2 Second-Round Data Collection, Data Analysis, and Findings

In the second-round qualitative trial, individual teachers first participated in a 45-minute rater training. During this session, I first familiarised teachers with the revised draft scale and asked them to comment on the scale properties and criteria and I noted their comments.

Following the individual rater training, two groups of the teachers separately participated in a 2-hour group discussion carried out at different points in time to elicit their feedback on the scale. During each session, the teachers were trained to use the revised draft scale before trialling it on two student essays and then filling out the scale evaluation form. The teachers then discussed and provided written comments on the scale evaluation forms, the paper-based scales, and the marked essays. One teacher who was not able to participate in any group discussions later participated in an individual session.

The elicited comments, including written comments on the scale evaluation forms and oral comments, were transcribed and thematically analysed in order to generate themes to refine the revised draft scale. I manually reviewed and compared the comments to examine what aspects of the scale were considered problematic and which problematic features were frequently mentioned by the teachers.

Following the qualitative analytic approach, I was able to identify some aspects of the scale that should be refined to ascertain the appropriate functionality,

comprehensibility, and applicability of the scale from the perspectives of the local classroom teachers. In general, all teachers perceived that the scale criteria covered the teaching contents, the rating format and scale layout were applicable and easy to use, and the language was understandable. Most teachers made frequent comments on descriptors related to coherence, content, grammar, sentence, and vocabulary.

Regarding the coherence domain, the teachers suggested some descriptors related to essay main ideas were highly relevant and thus should be merged in order to make the descriptors clearer and more distinct from one another. Further, essay length was deemed to be more related to the organisation domain and thus this descriptor was moved to the organisation domain. Some teachers proposed that all paragraphs should be well balanced in terms of content and thus a paragraph balancing descriptor was added into the content domain.

As for the grammatical descriptors, all teachers found it difficult to judge grammatical descriptors and perceived that the term "accurately" was subject to individual raters' interpretation. Accordingly, the teacher proposed that the counting of errors would help them to objectively score grammar descriptors. After the discussion, we came to the agreement that if students made more than three errors for a particular feature, then they would be judged as weak for that skill. Some teachers also proposed adding the phrase "*a few errors*" in the descriptors to make them more specific and objective. Therefore, this phrase was added to the descriptors. This rating rule and descriptor rewording were also applied for the descriptors measuring sentence accuracy and the use of mechanics. The transition signal descriptor in the grammar domain was deleted as some teachers thought it was already assessed in the cohesion domain and thus may not be necessary. Some teachers proposed that parallel structure could be excluded from the grammar domain to reduce the number of descriptors for practical purposes.

In terms of sentence use, all teachers perceived that the counting of errors for each type of sentence would consume too much time. Accordingly, they proposed that the sentence variety feature could be assessed by checking whether students used each and all types of sentence. Doing so would indicate that they used a variety of sentences. Sentence accuracy could be assessed by counting the sentence errors or problems.

In regard to vocabulary use, while all teachers thought that collocations were very important in writing and vocabulary knowledge, some teachers argued that the use of

collocations depended heavily on the writing topic and context and that some topics or context might not facilitate students' use of collocations. They also considered that the many types of collocations in English made it difficult to identify which chunks of words were or were not collocations. Consequently, the collocation descriptor was excluded from the vocabulary domain.

In addition, there were interesting findings emerging from the teacher discussion. All teachers agreed that a common problem found in the student trial essays was Thai-like expression, resulting from Thai learners' tendency to think in and translate from their L1. However, all teachers perceived that it takes a great deal of time and experience for students to use English in a more native-like manner and it is impossible for teachers to remedy this problem over a short course of time or a one semester course. Precisely for this reason, we decided not to diagnose the Thai-like expression problem. After the second qualitative trialling step, the 37 descriptors were reworded and reduced to 33 descriptors on the finalised scale (see Appendix C) The finalised scale was then returned to all teachers to confirm its appropriateness for the quantitative trialling phase.

3.5.4 Quantitative Trialling Phase

The post-hoc quantitative trial were taken over two rounds to investigate, based on the teachers' ratings of student essay samples, whether the teachers were acceptably consistent in applying the scale criteria. The quantitative findings could confirm the impact of the preceding qualitative trial and the scale comprehensibility. It should be noted that the quantitative trial focused on examining the teachers' interrater agreement.

3.5.4.1 First-Round Data Collection, Data Analysis, and Findings

In the first step, five teachers independently used the scale to pilot-rate the same set of 10 essays with different genres and qualities, purposively selected and written by the students in the context. After the teachers returned the scales and student essays, I then checked the scales for missing data. If there were any unmarked descriptors on any scales, I would immediately return the scales to teachers to complete the unmarked descriptors. The score data were analysed based on the CTT reliability statistics to evaluate the teachers' interrater consistency reliability in order to ensure the teachers' rating consistency and criteria comprehension. Since the data were complete without missing

data and each rater' rating data were normally distributed, the Cronbach's alpha method was appropriate for analysing the dichotomous rating scores and thus was used to investigate the scale internal consistency and interrater reliability (Stemler & Tsai, 2008). Based on the first quantitative trialling step, statistical results revealed that the teachers were not acceptably consistent in their rating performances, with a Cronbach's interrater reliability coefficient below the acceptable value of 0.7 as expected in a low-stake classroom assessment (Nunnally & Bernstein, 1994). Accordingly, two separate extensive sessions of rater retraining were conducted to maximise the teachers' rating consistency and criteria comprehension.

3.5.4.2 Second-Round Data Collection, Data Analysis, and Findings

Since not all the teachers were able to participate in the same retraining session. The retraining was divided into two sessions, in which three teachers participated in the first session and two attended the subsequent session. The rating results in the first session was used to guide and norm the teacher's judgement in the second session. During the retraining sessions, I and the teachers discussed the results from the first round of the quantitative trial and we found that during the rating practice, the teachers did not have the opportunity to compare and negotiate their rating decisions in order to norm their judgement and comprehension of the descriptors. This may well have been the reason for the unsatisfactory interrater reliability. After that, we discussed the scale criteria in detail before practicing rating one student essay. During the rating practice, the teachers discussed, compared, and negotiated their decisions for each descriptor rating.

After the practice session, each teacher independently rated the same set of three student essays which were also used in the first-round trial. The teachers were reminded to make sure all descriptors were marked. Each group rater retraining session took around 3 hours and 30 minutes. In the rater retraining, I also rated the same essays in order to ensure that the teachers and I was applying the scale in a similar manner. It should be noted that the number of student essay samples used in the second round was very small. This was inevitably due to the fact that the teachers had limited time to pilot-rate a representative number of essays as the writing courses were about to start and the scale needed to be ready for actual classroom implementation.

The quantitative scores consisted of the five teachers' ratings and the researcher's ratings. The score data were analysed via the IBM SPSS programme to examine the Cronbach's alpha interrater reliability and via the RStudio programme to estimate the percentage of interrater agreement. Statistical results showed that the alpha coefficient of interrater consistency was 0.89 and the average percent interrater agreement for the scale was 82%, thereby confirming the teachers' acceptably high rating consistency and agreement. Given the time constraint and the low-stakes nature of classroom assessment, the Cronbach's alpha coefficient and percent interrater agreement could be deemed an adequate indication of the scale quality and teachers' rating quality at this small-scale trialling stage.

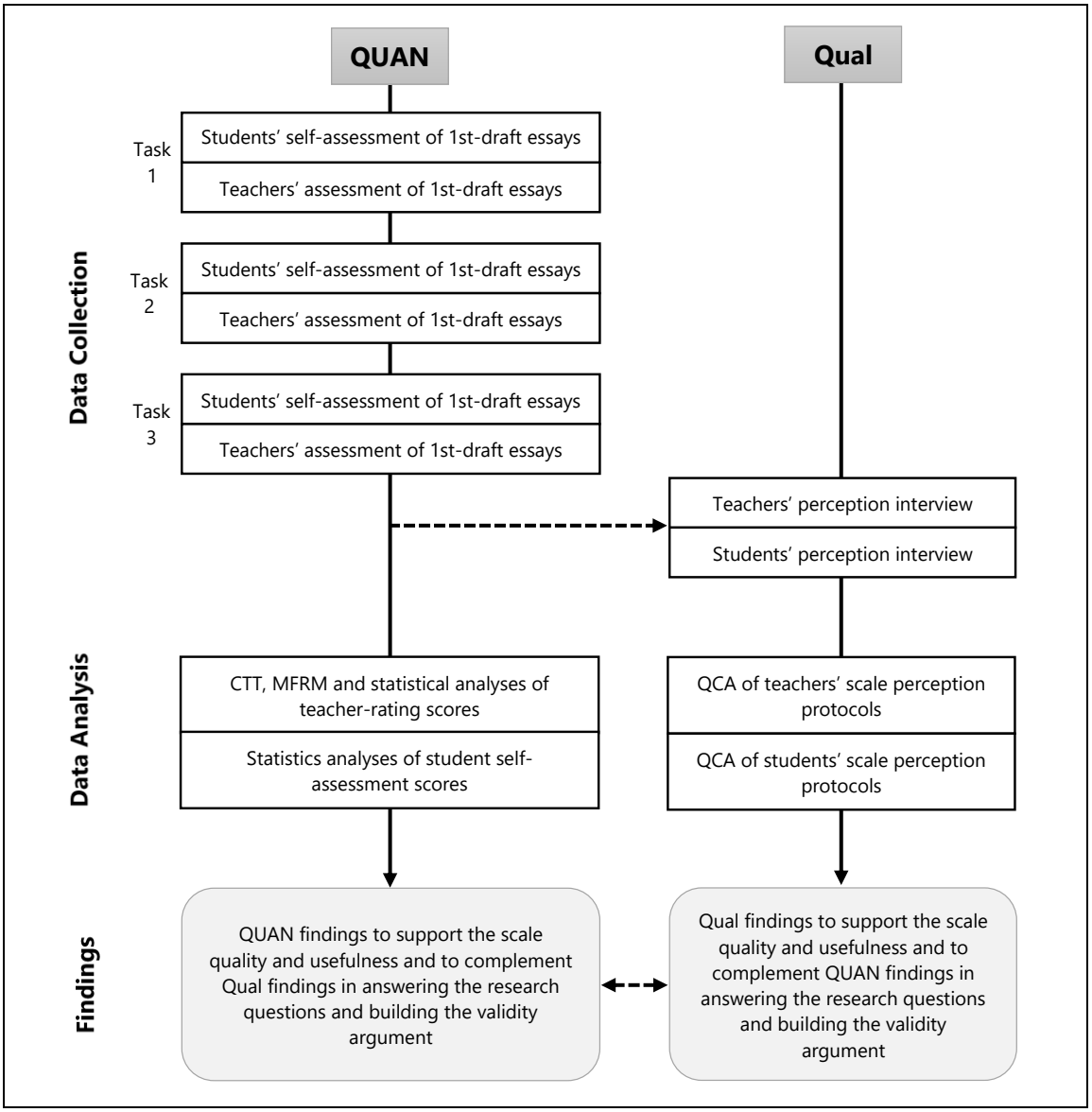
The findings from the quantitative stage were very helpful in two main aspects. Firstly, the statistical results derived from the first-round quantitative trial showed that the alpha interrater reliability was not robust enough, signposting problems in relation to the rater training process. Secondly, the statistical results from the second round showed an overall interrater reliability of almost 0.9 and a percent interrater agreement of over 80% for the scale, thereby demonstrating that all raters were sufficiently consistent in applying the scale. The quantitative findings could be taken as a sign that the rater training was successful and the scale criteria were interpretable by the teacher users. All in all, the first-round findings resulted in the rater retraining of the teachers and the improvement of the rater training procedures, whereas the second-round findings confirmed that the teachers inter-consistently applied the scale and thus were ready to move to the scale implementation stage.

In the scale trialling stage, both qualitative and quantitative data informed and confirmed each other in contributing to the scale quality enhancement. Based on the findings of the scale trialling stage, it could be concluded that (a) the scale criteria were appropriate to capture the construct assessed in the context, (b) the scale criteria were clear and interpretable by the teacher users in the context, and (c) the scale application and rating procedures were suitably aligned with the classroom teaching and learning process. The sequential and connected process of this qualitatively-informed, quantitatively-confirmed study thus added more value to the scale trialling process. The findings resulting from this stage led to the decision that the scale was ready to be operationally implemented.

3.6 Scale Implementation Stage

The final scale implementation stage was intended to apply and justify the scale in order to confirm, from the perspectives of psychometrics and stakeholders, that the scale serves it intended purposes. Figure 3.4 displays the procedural diagram of the scale implementation stage, where quantitative and qualitative methods were used to sequentially collect, separately analyse, and complementarily interpret both types of data. The findings from this stage were aimed at responding to the research questions and providing evidence to evaluate the validity argument for the scale. The findings could also corroborate insights from the scale construction and trialling stages and the generalisability of the scale to the population and context under study.

Figure 3. 4 *Procedural Diagram of the Scale Implementation Stage*



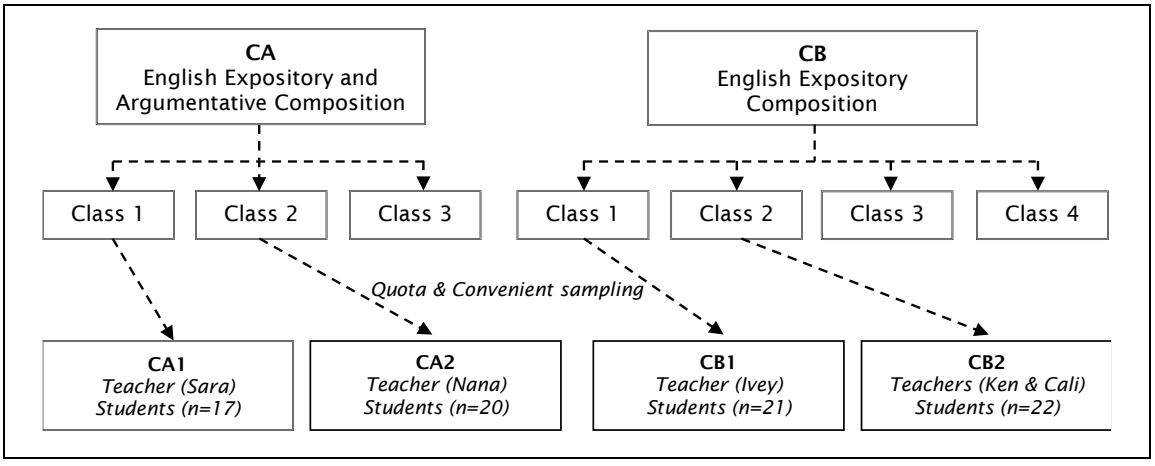
3.6.1 Participants and Context

Participants in the scale implementation stage included the five teachers, who also participated in the scale trialling stage (see Table 3.2 for more detail), 80 third-year TESOL and English-major undergraduates, and a Thai lecturer of English who was an additional coder in the trial of the coding frame. Participants were provided with plain language statement and consent forms to participate in this research. The scale was applied in two academic English writing classrooms, selected following a quota-convenient sampling strategy (see Figure 3.5), over a period of one semester in a Thai public university.

The first course (CA), comprised three classrooms in total, was intended to teach the students how to write English expository and argumentative writing compositions with emphasis on a classic five-paragraph essay format. Only two classrooms (CA1 and CA2) were selected. The CA1 classroom included 18 students and was taught by Sara and the CA2 classroom consisted of 20 students and was taught by Nana. The students enrolled in this course were English-major students at the faculty of humanities and social sciences. One student in the CA1 classroom did not participate in this research and thus only 17 students participated from this class.

The second course (CB), divided into four classrooms in total, was designed to teach English expository compositions with a focus on a classic five-paragraph essay to third-year TESOL students. Only two classrooms (CB1 and CB2) were selected. The students enrolled in this course were TESOL students at the faculty of education. The CB1 classroom included 21 students taught by Ivey and the CB2 classroom consisted of 23 students taught by Ken and Cali.

Figure 3. 5 *Procedural Diagram of Participant Sampling*



In terms of L2 language education, the Thai Ministry of Education has set the policy that educational institutions at all levels follow the Communicative Language Teaching (CLT) approach and adapt the Common European Framework of Reference for Languages (CEFR) as the frameworks for language curriculum, syllabus, and assessment in Thailand. Accordingly, more and more educational institutions have started to embrace the policy and attempted to base L2 language curriculum, syllabus, and assessment on CEFR. At the tertiary education level, the Ministry has set the target CEFR level for undergraduate students at B2 (vantage or upper-intermediate). However, higher CEFR levels are expected for English-major students and those enrolling in international programmes.

In the context under study, the existing course curriculum, syllabus, and assessment were not yet driven by CEFR at the time the current data collection was undertaken in 2018. Therefore, the design and development of the curriculum, syllabus, and assessment in the writing courses of interest were based primarily on the existing curriculum and the group of teachers responsible for the courses. In relation to classroom assessment, students were evaluated through grading system on the percentage scale which was assessed mainly from midterm and final examinations, and partly from classroom participation and assignment.

Table 3.3 summarises the characteristics of the writing assignment tasks (see also Appendix K), students, and teachers in the classrooms. The teachers were responsible for designing their own writing tasks and instruction methods as they saw fit. Over the semester, the writing courses lasted around 15 weeks and students attended a three-hour class a week, making up about 45 hours in total. Throughout the courses, students were mainly taught how to write a basic five-paragraph academic essay mainly on expository and argumentative genres and were required to write four assignment tasks altogether and write two drafts for each assignment task. Students were assigned to write on the same topic, or choose one preferred topic from many topics provided, or create their own topic related to the assigned genre. Apart from the assignment, students had to take midterm and final exams developed by teachers. It is important to note that this research did not cover the design of the writing assignment tasks and instruction methods. Yet, the scale was designed to fit the teaching practice to the extent possible.

Table 3. 3 *Characteristics of Assignment Tasks, Students, and Teachers*

Classes	Students	Task 1	Task 2	Task 3	Teachers
CA1	• n = 17 • Males = 4 • Females = 13 • English major	• Cause and effect *(write the same topic)*	• Problem and solution *(choose one from five optional topics)*	• Argument *(choose one from five optional topics)*	• Sara
CA2	• n = 20 • Males = 2 • Females = 18 • English major	• Cause and effect *(choose one from three optional topics)*	• Problem and solution *(choose one from four optional topics)*	• Argument *(choose your own topic)*	• Nana
CB1	• n = 21 • Males = 21 • TESOL major	• Description *(write the same topic)*	• Process *(choose your own topic)*	• Compare and contrast *(choose your own topic)*	• Ivey
CB2	• n = 22 • Males =3 • Females = 22 • TESOL major	• Description *(choose your own topic)*	• Cause and effect *(choose your own topic)*	• Compare and contrast *(write the same topic)*	• Ken • Cali

3.6.2 Instruments

In the scale implementation stage, data collection instruments included (1) the diagnostic binary rating scale, (2) a teacher perception interview, (3) a student self-assessment interview, (4) a student perception interview, (5) a teacher background questionnaire, and (6) a student background questionnaire.

The perception and self-assessment interview questions were driven by the IUA framework and hence intended to seek qualitative responses in contributing to the research questions and validity argument. Apart from the questions appearing on the interview instruments, there were other questions or prompts directed by the researcher in order to follow up participants' responses during the interview sessions and these questions are not included in the instruments. All the perception responses were thematically analysed to generate findings in response to Research Question (RQ)1, RQ3, and RQ4 and in justification of relevant validity assumptions.

The finalised diagnostic rating scale (see Appendix C) included 33 binary descriptors, grouped into eight domains: organisation, coherence, cohesion, content, grammar use, sentence use, vocabulary use, and mechanic use. The reader is reminded that the terms "*binary rating scale*" and "*binary checklist*" are used interchangeably to refer

to the diagnostic binary scoring or judgement, where "0" represents a non-mastery or unsatisfactory status and "1" indicates a mastery or satisfactory status of a writing skill.

The teacher perception interview (see Appendix F), adapted from Kim (2010) and Wagner (2015), included semi-structured questions used to elicit feedback from all teachers generally about the quality of the scale and the usefulness of the scale with regard to writing instruction, assessment, diagnostic feedback, student-self-assessment, and their reflection on participating in this research.

The student self-assessment interview (see Appendix G), modified from Kim (2010) and Wagner (2015), consisted of structured and semi-structure questions used to look into 20 volunteer students' self-assessment strategies.

The student perception interview (see Appendix H), adapted from Kim (2010) and Wagner (2015), consisted of semi-structured questions asking 20 volunteer students' perception regarding the quality of the scale and the usefulness of the scale related to their learning, diagnostic feedback, self-assessment implementation, and their reflection on participating in this research.

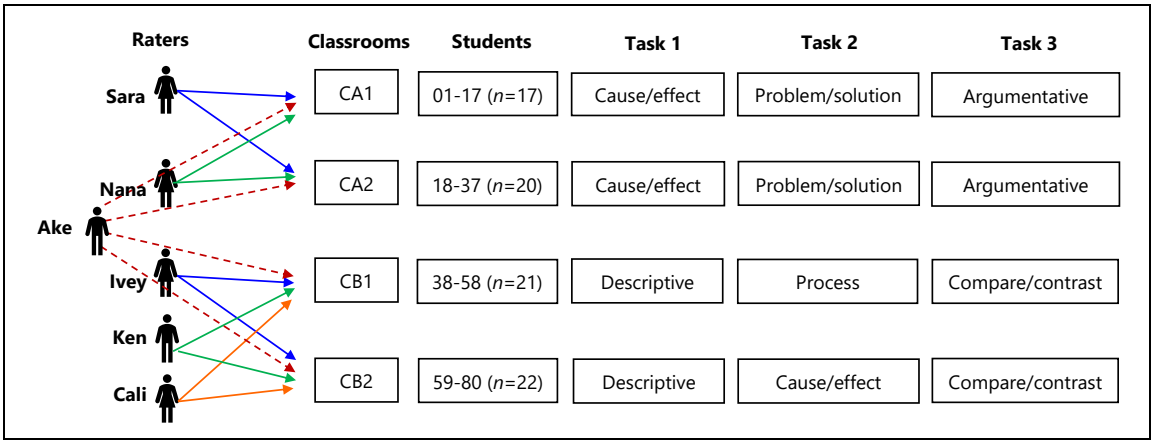
The teacher and student background questionnaires were adapted from previous research (Barkaoui, 2008; Cumming et al., 2002; Douglas & Hegelheimer, 2007; Kim, 2010; Wagner, 2015). The teacher questionnaire (see Appendix I) was used for all teachers to fill out their personal and educational information as well as their writing teaching and assessment experience. The student questionnaire (see Appendix J) was used for all students to fill out their personal and educational background and their experience in English and writing learning.

3.6.3 Formative Diagnostic Assessment Procedures

In this implementation stage, the teachers were not given formal rater training since they had been substantially trained in the preceding scale trialling stage. As displayed in Figure 3.6, each teacher began rating the student essays in his/her classroom first before rating the student essays from the other intact classroom. In order for the data to be linked for the MFRM analysis, I (Ake) as the researcher and the teacher in the context also randomly rated about half of the students in each class on three tasks. Thus, only 41 randomly-selected students' essays were rated by the researcher. It is important to note

that the rating conditions of Ake were rather different from those of the classroom teachers as he was not involved in the teaching and learning processes.

Figure 3. 6 *Procedural Diagram of Rating Design*



During the classes, the scale was primarily utilised as a diagnostic tool in the formative diagnostic assessment aiming to support ongoing teaching and learning processes. As it was difficult for the teachers to diagnose two essay drafts for each task on all four tasks delivered over the semester, they suggested that the scale be applied only to the first-draft essays on the first three assignment tasks so that they could provide feedback to the students to further revise the second-draft essay. The teachers were also encouraged to use the scale as part of their feedback delivery.

As regards the student self-assessment, I was allowed to spend approximately 90 minutes in each classroom at the beginning of the courses to familiarise the students with the assessment purposes, scale characteristics, criteria interpretation, and self-assessment procedures. The students were informed regarding the purpose of doing the self-assessment and how it could support their writing learning in the classrooms. In particular, the students were advised to use the scale to guide, self-evaluate, and revise their assignment writing draft until they were satisfied with their writing essays before submission. The students were also encouraged to compare their self-assessment results with their teachers' feedback. All students were required to email their first draft essays of each assignment to me and I then arranged the student essays before giving the same sets of the essays to each teacher responsible for each classroom in each writing course.

The students were informed that the diagnostic scores would not be used as part of their summative achievement assessment.

3.6.4 Overview of the Research Questions

In this study, the research questions are driven by the mixed methods research paradigm and the argument-based approach to validation. As the validity argument requires sound evidentiary backing, complete answers to the research questions need to be shaped by mixed qualitative and quantitative findings in the scale implementation stage in order to achieve the findings which are linked to the IUA and validity argument in the present study.

The first question- *To what extent does the diagnostic rating scale function appropriately for the formative diagnostic assessment in the EFL university writing classroom?* - examined the appropriateness of the scale functioning in order to provide evidentiary findings justifying the assumptions of the warrants underlying the evaluation, explanation, and extrapolation inferences. The appropriateness of the scale functioning is reflected through (1) the psychometric properties of the scale and rater behaviour, examined via descriptive, the CTT, MFRM, and correlation analyses, and (2) the teachers' and students' perceptions of the scale functioning, investigated via the qualitative content analysis (QCA).

The second question – *To what extent does the diagnostic rating scale function consistently for the formative diagnostic assessment in the EFL university writing classroom?* – investigated the consistency of the scale functioning with a view to providing evidence justifying the generalisation inference. The consistency of the scale functioning is illuminated through the psychometric properties of the scale and rater behaviour, investigated via the CTT and MFRM analyses.

The third question – *To what extent does the diagnostic rating scale support formative decisions about teaching and learning in the EFL university writing classroom?* – probed the usefulness of the scale in supporting the teacher and student formative decisions in teaching and learning processes. The usefulness of the scale is reflected solely though the QCA of the teacher and student perceptions of the scale. The evidentiary findings are used to justify the decision inference.

The fourth question – *To what extent does the formative diagnostic assessment have beneficial consequences for teaching and learning in the EFL university writing classroom?* – looked at the consequences of the scale-driven assessment to provide evidence pertaining to the consequence inference. The consequences are investigated via descriptive, MFRM, ANOVA, and regression analyses and the QCA of the teacher and student perceptions.

3.6.5 Quantitative Data

There were two main sets of quantitative data in the scale implementation stage. The first dataset was the students' diagnostic scores assigned by the teachers on the three assignment tasks. The second dataset was the student self-assessment scores on the three assignment tasks. For investigation purposes, the students were grouped into different ability levels based on the total achievement exam (midterm and final exams) developed by the classroom teachers.

As shown in Table 3.4, Course A (CA) and Course B (CB) used different criteria for evaluating the student achievement which had different total scores for the midterm and final exams. Since the total exam scores are different in both courses, the exam scores were converted into percentages for ability group classification and data analysis. Of the total exam score, students receiving 75% and above were grouped as high-ability students, those receiving between 65% and 74% were classified as mid-ability students, and those receiving less than 65% were classified as low-ability students.

Table 3. 4 *Criteria for Grouping Student Ability Level Based on Total Exam Scores*

Criteria	CA (Totall score = 30)	CB (Total score = 60)	Ability Levels
≥ 75%	22.50 – 30	45.00 – 60	High (*n* = 15)
≥ 65%	19.50 – 22.49	39.00 – 44.99	Mid (*n* = 33)
< 65%	0 – 19.49	0 – 38.99	Low (*n*= 32)

3.6.5.1 Diagnostic Scores Based on Teacher Ratings

As shown in Table 3.5, the first quantitative dataset was obtained from the formative diagnostic assessment of the student first-draft essays, scored by five teachers on the three tasks and it was analysed to investigate rater behaviour, scale functioning, and student writing ability. The CA scores were averaged from the ratings of Sara and

Nana, both of whom fully cross-rated 37 students' essays and the CB scores were averaged from the ratings of Ivy, Ken, and Cali, all of whom fully cross-rated 43 students' essays. There were no missing data in the teachers' ratings since individuals were followed up to re-score the unmarked descriptors. The CA and CB datasets were used for descriptive, ANOVA, correlation, regression, and CTT analyses.

Table 3. 5 *Characteristics of CA and CB Datasets*

Assessment conditions	CA	CB
Number of teachers	2 (Sara & Nana)	3 (Ivy, Ken, & Cali)
Number of students	37 (ID01-37)	43 (ID38-80)
Number of tasks	3	3
Number of essays	111	129
Number of ratings	222	387
Number of scores	7,326	12,771

The MFRM analysis was based on the connected dataset shown in Tables 3.6 and 3.7, in which I (Ake) randomly rated the essays of 41 randomly-selected CA and CB students in order to link the CA and CB datasets together for the MFRM analysis.

Table 3. 6 *Rating Design in the Connected Dataset*

Class	Raters	Student ID	Number of essays	Number of descriptors
CA1	Sara	01-17	51	1,683
	Nana	01-17	51	1,683
	Ake	02, 05, 07, 08, 10, 11, 13, 15, 16	27	891
CA2	Sara	18-37	54	1,980
	Nana	18-37	54	1,980
	Ake	18, 19, 20, 21, 22, 28, 29, 31, 32, 33	30	990
CB1	Ivy	38-58	63	2,079
	Ken	38-58	63	2,079
	Cali	38-58	63	2,079
	Ake	38, 39, 44, 45, 48, 50, 51, 52, 54, 56, 58	33	1,089
CB2	Ivy	59-80	66	2,178
	Ken	59-80	66	2,178
	Cali	59-80	66	2,178
	Ake	59, 61, 63, 65, 68, 70, 72, 74, 76, 78, 80	33	1,089

Table 3. 7 *Characteristics of the Connected Dataset*

Raters	N of students	Number of tasks	Number of essays	N of scores
Sara	37 (ID01-37)	3	111	3,663
Nana	37 (ID01-37)	3	111	3,663
Ivy	43 (ID38-80)	3	129	4,257
Ken	43 (ID38-80)	3	129	4,257
Cali	43 (ID38-80)	3	129	4,257
Ake	41 random students	3	123	2,970

3.6.5.2 Diagnostic Scores Based on Student Self-Assessment

The second quantitative dataset was the self-assessment scores on three tasks. Due to some missing data, 12 students (ID: 01, 03, 12,14, 38, 41, 45, 48, 50, 52, 55, 70) who did not self-rate all descriptors and all tasks were excluded from the self-assessment data. Accordingly, only 68 students were found to rate all descriptors and all tasks and thus were included in the dataset. The self-assessment scores were analysed to examine how well the students self-assessed their own essays in comparison to their teachers' ratings. Although the purpose of the current self-assessment was to promote students' learning engagement and self-learning skills rather than the self-assessment accuracy, the degree of correspondence between the self-ratings and the teacher ratings could indicate to what extent the students were attentive to the self-assessment, thus offering partial evidence of the student self-regulated learning development. The self-assessment scores from the CA and CB courses were combined for ANOVA, correlation, and regression analyses. Due to no double-ratings for the self-assessment dataset, the self-assessment scores were not appropriate for the MFRM analysis.

3.6.6 Quantitative Data Analyses

Table 3.8 shows the quantitative analyses, analytic purposes, expected results, and research questions (RQ) in the scale implementation stage. Descriptive, CTT, ANOVA, correlation, and regression analyses were conducted via the IBM SPSS programme (Version 25) and only the analysis of percent interrater agreement was conducted using the RStudio programme. The MFRM analysis was run via the FACETS programme (Version 3.80.4; Linacre, 2018). Each analytic method is presented in more detail in the next sections.

Table 3. 8 *Quantitative Analyses, Analytic Purposes, and Research Questions*

Statistics	Analytic purposes	RQ
Descriptive	・Describe characteristics of rater ratings and diagnostic scores in order to examine appropriacy of scale functioning and rater behaviour.	1
	・Describe average diagnostic scores over tasks in order to examine student learning progression.	4
CTT	・Determine indices of appropriacy of scale functioning.	1
	・Determine percent interrater agreement in order to examine consistency of rater behaviour.	2
MFRM	・Determine indices of appropriacy and consistency of rater behaviour.	1, 2

Statistics	Analytic purposes	RQ
	• Determine indices of appropriacy and consistency of scale functioning.	1, 2
	• Determine indices of student writing performance in order to examine appropriacy and consistency of scale functioning and rater behaviour.	1, 2
	• Determine diagnostic logits over tasks in order to examine student learning progression.	4
ANOVA	• Compare diagnostic scores/logits between ability groups in order to examine alignment between scale function and achievement assessment.	1
	• Compare self-rating severity/leniency between student ability groups in order to examine student severity/leniency difference.	4
	• Compare student-ratings with teacher-ratings across tasks in order to examine self-assessment behaviour and self-regulated learning.	4
Correlation	• Explore relationship between diagnostic scores/logits and percent exam in order to examine alignment between formative assessment and summative achievement	1
	• Explore relationship between self-assessment scores and percent exam in order to examine impact of self-assessment on summative achievement.	4
	• Explore relationship between student-ratings and teacher-ratings across tasks in order to examine self-assessment behaviour and self-regulated learning development.	4
	• Explore relationship between rater agreement, descriptor difficulty, and essay quality in order to examine consistency of rater behaviour.	2
Regression	• Explain relationship between diagnostic scores/logits and percent exam in order to examine impact of formative assessment on summative achievement.	4
	• Explain relationship between self-assessment scores and percent exam in order to examine impact of self-assessment on summative achievement.	4

3.6.6.1 Descriptive Statistics

The descriptive analysis was conducted to (1) describe the characteristics of the teachers' ratings and students' diagnostic scores through measures of centrality (mean), variability (standard deviation and range), and distribution (skewness and kurtosis) of the quantitative data, and (2) describe the changing patterns of the student scores over the tasks based on the score means. The skewness and kurtosis indices should fall within the acceptable range of ±2 in order to show an acceptably univariate normal distribution of the data (George & Mallery, 2018). Descriptive results indicate (1) the appropriateness of the scale functioning and rater behaviour, thus yielding information relevant to RQ1, and (2) the student learning progression over the tasks, hence responding to RQ4.

3.6.6.2 Classical Test Theory

The CTT analysis was run to (1) determine the psychometric indices of the scale functioning, including item difficulty, corrected item-total correlation, Cronbach's alpha

reliability, and (2) determine the percent absolute interrater agreement of descriptors and students. CTT results indicate the appropriateness of the scale functioning, thereby answering RQ1, and the consistency and homogeneity of the rater behaviour, hence responding to RQ2.

The dichotomous item difficulty is estimated by dividing the total number of correct options on the descriptor by the total number of students (Clauser & Hambleton, 2018). The higher the average score on a descriptor, the easier the descriptor and vice-versa (Clauser & Hambleton, 2018).

The corrected item-total (CIT) correlation demonstrates to what extent the descriptors align with one another to measure the same construct of interest (Johnson & Morgan, 2016). The correlation ranges from -1 to +1 and desirable values should be positive or above 0.20 (Johnson & Morgan, 2016). A high CIT correlation, at least over 0.2, implies that students diagnosed as "*weak*" on a descriptor have a low total scale score and the students diagnosed as "*strong*" on the descriptor have a high total scale score (Johnson & Morgan, 2016).

The Cronbach's alpha internal consistency reliability method is used as a measure of internal consistency and homogeneity to examine how well all items on a test are consistent or homogeneous in targeting the same construct under measure (DeVellis, 2017). Nunnally and Bernstein (1994) suggest that reliability coefficients of at least 0.70 is required for low-stakes assessments.

The percent interrater agreement (absolute interrater consensus) was employed to determine the extent to which the raters assign the same exact rating on a particular descriptor or student (Eckes, 2015). The concept for interpreting a percent interrater agreement is similar to a reliability coefficient. The percentage of interrater agreement should thus be over 70% to show acceptable interrater agreement for the current less-formal formative diagnostic assessment, where low-stakes decisions are made on the basis of the assessment to adjust and improve learning and teaching.

3.6.6.3 Many-Facets Rasch Model

Introduced by Linacre (1989) as part of the family of the Rasch psychometric theory introduced by Georg Rasch in 1960, the MFRM is capable of investigating multiple sources of measurement error, such as raters, scoring criteria, and tasks in a single analysis and

thus it provides more precise and reliable results related to assessment quality and to rater-mediated performance assessments in particular (Eckes, 2015; 2019; Engelhard & Wind, 2018; McNamara et al., 2019). This study took advantage of the MFRM framework to investigate students' writing ability, raters' judgement quality, and scale quality in order to ensure that the scale-generated diagnostic information could be interpreted and used appropriately. The MFRM analysis was conducted via the FACETS programme which employs the Joint Maximum Likelihood method to estimate facet parameters, logits, or measures (Linacre, 2018).

The MFRM analysis was used to analyse the connected dataset for four main purposes. Firstly, the MFRM analysis examined the behaviour of six raters in order to discover the appropriacy and consistency of their rating behaviour, hence answering RQ1 and RQ2. Secondly, the MFRM analysis investigated the appropriacy and consistency of the scale functioning, thereby responding to RQ1 and RQ2. Thirdly, the MFRM analysis investigated 80 students' writing ability and the student-based Rasch indices could be used to imply the appropriacy and consistency of the scale functioning and rater behaviour, thereby contributing to RQ1 and RQ2. Finally, the MFRM analysis scrutinised the ability of student writing performance over the tasks in order to uncover the student learning progression, thereby yielding information related to RQ4.

In the MFRM analysis, each descriptor on the scale was dichotomously scored and thus the dichotomous Rasch model was employed to examine the five facets identified as the systematic sources of the score variability in the current assessment. The five facets were rater (N = 6), student (N = 80), writing tasks (N = 6), descriptor (N = 33), and domain (N = 8). It should be noted that the domains were simply groupings of the descriptors and thus was specified as a dummy facet in this analysis.

In the FACETS specification file (see Appendix L), the object of the diagnostic assessment was the student facet and thus it was allowed to float on the logit scale. The student and task facets were positively oriented whereas the others were negatively oriented. The MFRM analysis was performed based on the connected dataset to globally and locally examine the raters' rating performance, the scale functioning (descriptors and domains), and the student diagnostic writing outcomes (student and task performance). Important Rasch assumptions (global data-model fit, psychometric unidimensionality, and local independence) were preliminarily examined to ensure meaningful interpretation of

MFRM results (Linacre, 1989; Rasch, 1960, 1980). Rasch statistical and graphical indicators at group and individual levels were used to investigate each of the facets under scrutiny.

Table 3.9 lists the various Rasch indicators which provided meaningful interpretation across the facets under the current investigation. The Rasch indicators were classified into group-level and individual-level indicators. The group-level indicators included (a) visual variable map, (b) fixed chi-square homogeneity test, (c) all separation statistics, (d) percent exact agreement, and (e) Rasch-Kappa. The individual-level indicators focused on (a) logit or measure, (b) standard error of estimate, (c) fit statistics, (d) point-measure (PTM) correlation, and (e) percent exact agreement.

Table 3. 9 *Rasch Indicators sand Analytic Purposes*

Rasch indicators	Purposes
Group-level indicators	
・Visual variable map	・Examine and compare distributions of elements within a facet on the logit scale.
・Fixed chi-square test	・Examine homogeneity of elements within a facet.
・Separation ratio	・Examine homogeneity of elements within a facet.
・Separation strata	・Examine homogeneity elements within a facet.
・Separation reliability	・Examine homogeneity of elements within a facet.
・Percent exact agreement	・Examine interrater consensus and independency.
・Rasch Kappa	・Examine interrater consensus agreement and independency.
Individual-level indicators	
・Logit or measure	・Examine individual elements' locations on the logit scale.
・Standard error of estimate	・Examine precision of individual elements' locations on the logit scale.
・Fit statistics	・Examine accuracy and consistency of rating patterns of elements within each facet.
・Point-measure correlation	・Examine internal consistency of descriptors.
・Percent exact agreement	・Examine intrarater consistency.

3.6.6.4 Analysis of Variance

The one-way independent ANOVA was conducted to compare the student diagnostic scores and logits between low-, mid, and high-ability groups in order to determine how consistently the scale reflected the achievement assessment, hence answering RQ1. Further, the ANOVA was used to compare the self-assessment scores with the teacher-assessment scores across the tasks so as to examine whether the self-assessment was more or less severe than the teacher-assessment. Finally, the ANOVA was used to compare self-rating severity/leniency between the student ability groups. The degree of correspondence between the self-assessment scores and the teacher-

assessment scores could partially imply to what extent the students were engaged in the self-assessment process and thus were developing self-regulated learning skills. Therefore, the ANOVA-based self-assessment results could partly provide information related to RQ4.

3.6.6.5 Correlation

The bivariate Pearson correlation analysis was conducted to explore the relationship between the diagnostic outcomes (raw scores and Rasch logits) and achievement exam percentages in order to discover how well the scale was aligned with the student achievement assessment, thus responding to RQ1. Moreover, the correlation was used to explore the relationship between the self-assessment scores and exam percentages in order to investigate how well the self-assessment related to the student summative achievement, partly responding to RQ4. The correlation was used to explore the relationship between the self-assessment scores and the teacher-assessment scores across the tasks so as to determine how well the student self-ratings corresponded to the teacher ratings, hence contributing to answering RQ4. Finally, the correlation was used to explore the relationship between the rater agreement percentages, the descriptor difficulty indices, and the diagnostic scores in order to investigate the consistency of the raters' behaviour, thus partly responding to RQ2.

3.6.6.6 Regression

The simple linear regression analysis was conducted via SPSS to analyse the combined CA and CB dataset in order to explain the predictive relationship between the diagnostic scores and the exam percentages, the diagnostic logits and the exam percentages, and the self-assessment scores and the exam percentages. The regression results indicate to what extent the formative diagnostic assessment and self-assessment contributed to the student summative achievement, thus partly responding to RQ4. In the regression analysis, the R-squared (R^2) coefficient or the coefficient of determination is the percentage of variance effect size used to represent the proportion of the variance for the achievement percentage that is explained by the formative diagnostic score in the regression model (George & Mallery, 2018).

3.6.7 Qualitative Data Analysis

The qualitative content analysis was conducted with the assistance of the NVivo programme (Version 12 Plus) to analyse the teacher and student interview datasets separately in order to investigate their perceptions of the scale functioning, usefulness, and impact as well as the student self-assessment practices or strategies.

Table 3.10 shows the purposes of the qualitative content analyses and the research questions that need to be answered by the qualitative findings. In both teacher and student perception analyses, the findings about the scale functioning will reveal what scale descriptors and features are perceived as effective or ineffective in the current assessment, thus responding to RQ1. The findings related to the scale usefulness reveals whether the scale and diagnostic information contribute to teaching and learning in the classroom, thus responding to RQ3 and RQ4. The findings pertaining to the scale impact illuminates to what extent the scale has benefits for teachers and students. The findings from the student perception analysis also reveal some of the strategies that the students used during their self-assessment, thus responding to RQ4.

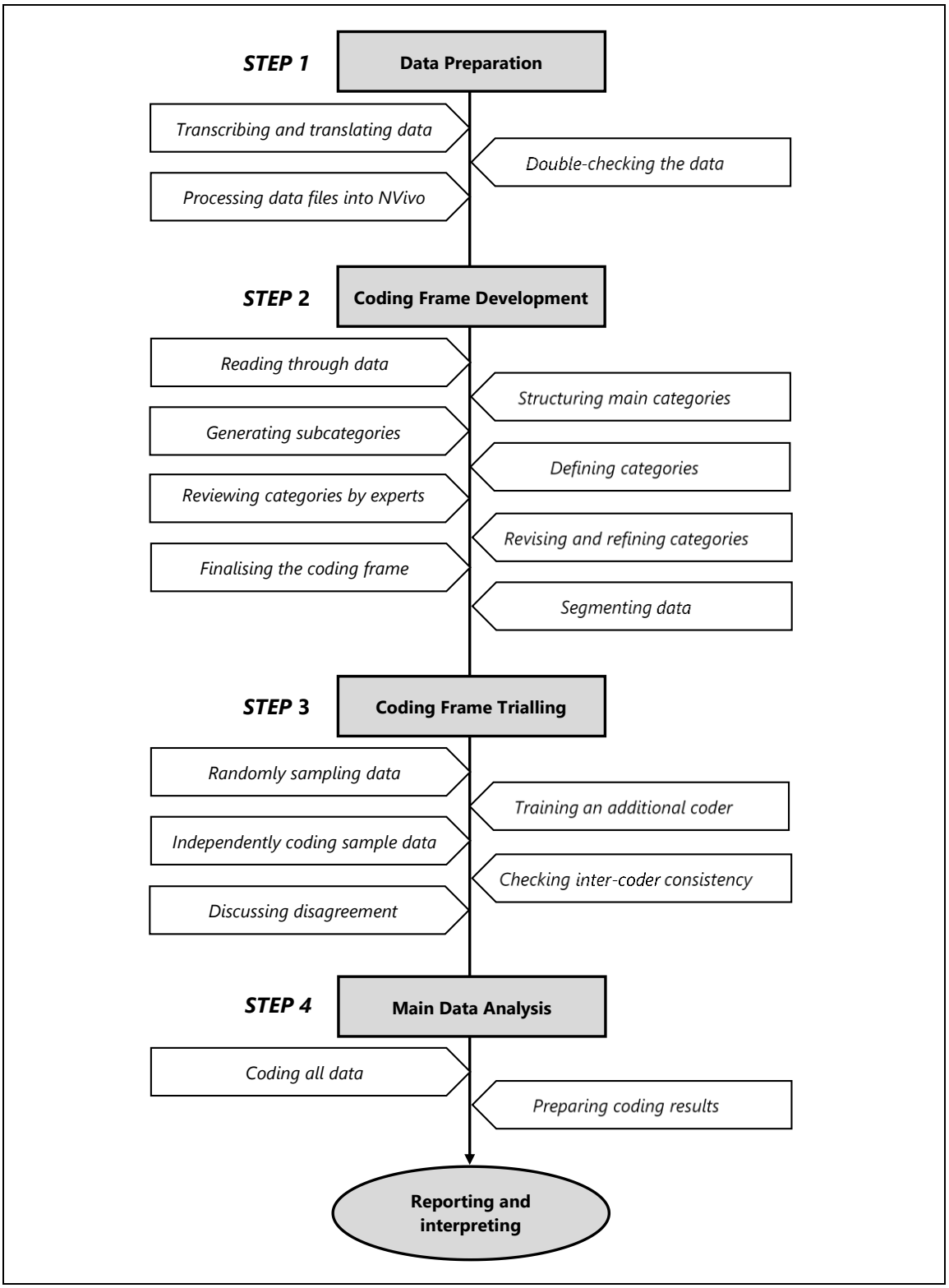
Table 3. 10 *Qualitative Analyses, Analytic Purposes, and Research Questions*

Analyses	Purposes	RQ
Qualitative content analysis of teacher perception	• Explore teacher perception of scale functioning	1
	• Explore teacher perception of scale usefulness and impact	3, 4
Qualitative content analysis of student perception	• Explore student self-assessment and self-regulated learning behaviours	4
	• Explore student perception of scale functioning	1
	• Explore student perception of scale usefulness and impact	3, 4

As displayed in Figure 3.7, the qualitative content analysis process comprised four main steps: (1) data preparation, (2) coding frame development, (3) coding frame trialling, and (4) main data analysis. The primary goal of the qualitative approach is to evaluate the validity of the interpretations and uses of the scale scores. To this end, this study retrospectively explored the teachers' and students' perceptions related to the functionality, usefulness, and impact of the scale via a semi-structured interview method. It should be noted that the focus was particularly upon the teachers' scale perceptions since the scale was designed specifically for the teacher user. The interview questions were conceptually predetermined by the IUA in turn driving the researcher questions. The

teacher and student perceptions were compared across the themes and the qualitative findings were used to supplement the quantitative findings in the evaluation of the validity argument.

Figure 3. 7 *Procedural Diagram of the Qualitative Content Analysis*



In the qualitative data analysis, the qualitative content analytic approach (Schreier, 2012) was adopted as it draws upon a constructivism paradigm and a Grounded Theory approach (Glaser & Strauss, 1967) to analyse qualitative data inductively and/or deductively through a systematic, flexible, and reductive process. In this way, the vast array of textual data was reduced deductively and inductively into aspects, themes and patterns of interest, in turn transformed into the main categories and subcategories in the coding frame (Schreier, 2012). The qualitative content analytic approach may otherwise be referred to as the thematic analysis (e.g., Braun & Clarke, 2006; Boyatzis, 1998) or applied thematic analysis (Guest et al., 2012).

Two essential aspects were checked to ensure the quality of the qualitative content analysis: coding consistency and coding frame validity (Ericsson & Simon, 1993; Green, 1998; Mackey & Gass, 2016; Schreier, 2012, 2014). A quantitative intercoder reliability method was used to ensure the consistency of the coding. In addition, experts were asked to review the coding frame and appropriate processes of data collection, data translation, data transcription, and coding frame development were undertaken to ensure the validity of the coding frame (Schreier, 2012, 2014). More details of the qualitative analytic procedures are presented next.

3.6.7.1 Data Preparation

In the data preparation step, I transcribed and translated all the audio-recorded interview data following the transcription conventions as shown in Table 3.11. Tables 3.12 and 3.13 summarise the characteristics of the five teachers' and 20 volunteer students' qualitative data respectively. The interviewees were given the option to respond in either Thai or English. Interview data in English were transcribed verbatim whereas the data in Thai were translated as thoroughly as possible to ensure the interviewees' intended meanings. All transcripts were double checked to ensure the completeness, legibility and accuracy of the transcription and translation. Non-verbal communication, such as pause, were not noted since they do not affect the analysis and interpretation of the data in question. Interview protocols were textually transcribed carefully and accurately and then kept in the researchers' computer and in NVivo for further qualitative content analysis.

Table 3. 11 *Interview Transcription Conventions*

Transcription conventions	Transcription elements
plain texts	Teacher's verbal report
[parentheses]	Researcher' prompts or speaking
(italic in parentheses)	Verbatim or almost verbatim mention of descriptor
Numeric (1, 2, 3)	Descriptor ID
ZERO, ONE, NO, YES	Rating score

Table 3. 12 *Characteristics of Teacher Interview Transcripts*

Teachers	Response language	Approximate interview Length	Approximate text Length
Sara	English	49 mins	4,350 words
Nana	English	49 mins	4,973 words
Ivy	Thai	28 mins	2,756 words
Ken	English	37 mins	3,875 words
Cali	English	35 mins	3,664 words

Table 3. 13 *Characteristics of Student Interview Transcripts*

Student ID	Response language	Approximate interview length	Approximate text length
03	Thai	20 mins	879 words
04	Thai	19 mins	1,021 words
09	Thai	39 mins	2,024 words
12	Thai	15 mins	1,406 words
17	Thai	14 mins	1,136 words
21	Thai	28 mins	1,315 words
23	Thai	30 mins	1,375 words
26	Thai	42 mins	1,294 words
31	Thai	28 mins	1,210 words
32	Thai	27 mins	1,500 words
39	Thai	32 mins	1,181 words
46	Thai	27 mins	879 words
54	Thai	25 mins	1,156 words
56	English	42 mins	1,691 words
57	Thai	28 mins	1,005 words
61	Thai	33 mins	1,577 words
62	Thai	29 mins	2,189 words
63	Thai	25 mins	1,346 words
64	Thai	39 mins	1,937 words
65	Thai	33 mins	1,382 words

3.6.7.2 Coding Frame Development

A linear and cyclic sequence of steps was used to develop and refine the coding frame based on a mixed concept-informed and data-driven method to ensure that the coding frame fit the data. Through a concept-driven method, I read carefully through the data to deductively review concepts and ideas that are related specifically to the research questions driven by the IUA. Using a data-driven method, I also read carefully throughout

the data to inductively explore newly-emerging concepts and ideas related to the research questions. This was to make sure that all the derived concepts informing the coding categories match the elicited data in the context and represent the framework-informed or predetermined concepts, thus ensuring the validity of the coding frame.

The predetermined concepts were hierarchically categorised from the most abstract highest-level main categories to the most concrete lowest-level subcategories. In this study, the research questions structured the general conceptual main categories, whereas the data generated the specific concrete subcategories, subsumed under each main category. Eventually, the coding frame comprised three key conceptual themes or dimensions serving as the main categories, each of which consisted of a set of two hierarchical level subcategories. The main categories are concerned with the teachers' perceptions of the scale with emphasis on three key dimensions: functionality, usefulness, and impact. In this study, the terms "*main categories*" and "*main themes*" are synonymous and used interchangeably. The subcategories were partly driven deductively by the interview concepts and partly driven inductively by the interview data.

The coding frame (see Appendix E) shows the teacher perception categories in a hierarchical three-level coding frame which includes the names, definitions, and quote examples of the coding categories. As indicated in the coding guidelines, the three most abstract, first-order categories include: (1) the scale functioning, (2) the scale usefulness, and (3) the scale impact. The quality of the scale properties comprised a set of three more specific, second-order subcategories, including scale comprehensibility, scale comprehensiveness, and scale organisation. The usefulness of the scale utilisation included a set of two more specific, second-order subcategories, including ongoing teaching and ongoing learning. The impact of the scale utilisation comprised a set of three more specific, second-order subcategories, including teacher awareness and future plan. Each of the second-order subcategories contained a set of most specific, third-order subcategories, each of which was clearly described for coding purposes.

The draft coding frame was repeatedly revised via NVivo based on cyclic reading of the transcripts and experts' review feedback to ensure that the main categories were well structured and the subcategories were well generated. After the coding frame was drafted, all interview transcripts were carefully segmented with emphasis on distinct

meaningful themes or contents rather than formal linguistic units such as worlds, phrases, and sentences (Schreier, 2012, 2014).

3.6.7.3 Coding Frame Trialling

Two teachers' transcripts were randomly selected for the coding frame trial. The coding frame generated from NVivo was modified for the purpose of the manual coding trialling. The additional coder was a female Thai lecturer of English with several years of EFL teaching experiences in a public Thai university. During the coding training, she was introduced to the data analysis purposes, the category definitions, and how to assign coding units to each of the categories. Then she practiced coding one of the data transcripts. After an about one-hour training session, I and the other coder independently and manually coded the two sampled transcripts for about one hour. Of all 78 codings judged by the two coders, 65 were similar and 13 were different, resulting in the percentage of absolute agreement of 83%. Mismatched codings were also compared and discussed until consensus was reached. There was no further revision of the coding frame for the main data coding.

3.6.7.4 Main Data Coding Analysis

All the coding units or segments were carefully checked again prior to the main coding process. In this stage, I coded all the interview data in NVivo. To begin with, I read carefully each of the teachers' transcript and assigned a coding segment to a coding category. After all the teachers' data were coded, I carefully coded each of the student transcripts. For the qualitative content analysis of the student perception data, not all coding categories matched the student data since the students were interviewed using fewer questions than the teacher interview. For this reason, some of the perception coding categories were reduced and the coding frame was modified to fit the student data. As well as the student perception data, the student interview included questions about the self-assessment practice and thus the responses to these questions were coded in a separate coding frame. To support the presentation of the qualitative findings, clear, concise and complete quotes of the participants were selected to illustrate the coding categories.

3.7 Chapter Summary

This chapter has presented the rationales underlying this mixed methods research and the interface between an argument-based approach to validation, the research questions and a mixed methods research methodology. A description of the scale construction and trialling was then provided, followed by a detailed description of the scale implementation. In addition, the quantitative and qualitative data collection and analytic methods used to generate the findings in response to each of the research questions were described and the connection between the IUA, the research questions, the backing data analyses were outlined in order to direct the development of the validity argument. The research results, first quantitative, then qualitative, will be presented in the next two chapters.

Chapter 4: Quantitative Results

This chapters presents the results of the quantitative data analyses. To begin with, descriptive results are presented to describe the characteristics of the teachers' ratings and the students' diagnostic scores on the three sequential assignment tasks. CTT results are then presented to demonstrate the quality of the scale functioning and the teachers' rating behaviour. Following this, MFRM results are presented to indicate the quality of the raters' rating behaviour, the scale functioning, and the students' writing ability. Finally, ANOVA, correlation, and regression results are presented to illustrate the relationship between the formative diagnostic assessment and the summative achievement assessment, and the correspondence between the student self-assessment and the teacher-led assessment.

4.1 Descriptive Results

Descriptive statistics were used to analyse Course A (CA) and Course B (CB) datasets separately. Descriptive results demonstrate (a) how appropriately the diagnostic rating scale and teachers assessed the student essays, thereby partly contributing to the results for Research Question (RQ) 1, and (b) whether the students showed any patterns of learning progress over the tasks, thereby partly responding to RQ4.

Table 4.1 shows descriptive statistics of the descriptor and student scores assigned by individual teachers across the three tasks. As can be seen, the mean (M) and standard deviation (SD) values suggest a noticeable variability of the descriptor and student scores across the tasks. The skewness (SK) indices are all negative yet within the acceptable range of ±2, suggesting rather negatively-skewed distribution of the scores. The kurtosis (KU) indices are, in large part, over 0 but within the acceptable range of ±2, indicating a rather leptokurtic distribution of the scores.

Table 4.2 shows descriptive statistics of the student scores, averaged from the ratings of the teachers within each course. Of 33 points in total, the CA students showed a high score on Task 1 (M = 25.62, SD = 4.74), a slightly decreased score on Task 2 (M = 25.09, SD = 4.14), and finally the highest on Task 3 score (M = 26.38, SD = 4.61). The CB students demonstrated the lowest score mean on Task 1 (M = 24.63, SD = 4.37), a

substantially increased score on Task 2 (*M* = 27.38, *SD* = 2.60), and finally a markedly decreased score Task 3 (*M* = 24.80, *SD* = 3.04). Overall, the students' score means are over 24 out of 33 points across the tasks, suggesting acceptable mastery of most writing skills. The range values from 11 to 22 and the SD values between 2.5 and 4.7 indicated a noticeably wide and varied range of the student ability captured by the scale. As the SK and KU statistics are largely within the acceptable range, it is thus assumed that the student scores were acceptably normally distributed.

Table 4. 1 *Descriptive Statistics of Descriptor and Student Scores*

Task	Rater	Descriptor scores					Student scores					
		N	M	SD	SK	KU	N	M	SD	Range	SK	KU
1	Sara	1221	0.75	0.43	-1.17	-0.64	37	24.81	4.26	21	-1.02	1.72
	Nana	1221	0.80	0.40	-1.51	0.28	37	26.43	5.59	21	-0.92	0.68
	Ivey	1419	0.77	0.42	-1.28	-0.36	43	25.40	4.93	22	-1.15	0.91
	Ken	1419	0.72	0.45	-0.99	-1.02	43	23.81	6.08	21	-0.29	-1.05
	Cali	1419	0.75	0.43	-1.14	-0.70	43	24.67	4.25	23	-0.96	2.58
2	Sara	1221	0.74	0.44	-1.07	-0.85	37	24.30	4.17	19	-0.74	0.58
	Nana	1221	0.78	0.41	-1.39	-0.08	37	25.89	5.23	25	-1.47	3.20
	Ivey	1419	0.85	0.35	-2.01	2.04	43	28.19	3.59	15	-1.01	0.79
	Ken	1419	0.79	0.41	-1.40	-0.03	43	25.98	3.58	18	-0.50	0.93
	Cali	1419	0.85	0.36	-1.94	1.76	43	27.98	2.92	15	-1.32	2.58
3	Sara	1221	0.76	0.43	-1.22	-0.50	37	25.11	4.59	17	-0.35	-0.75
	Nana	1221	0.84	0.37	-1.84	1.37	37	27.65	5.79	27	-1.91	4.58
	Ivey	1419	0.73	0.44	-1.04	-0.92	43	24.09	4.14	16	-0.58	-0.50
	Ken	1419	0.75	0.43	-1.18	-0.62	43	24.86	3.73	15	-0.36	-0.55
	Cali	1419	0.77	0.42	-1.29	-0.33	43	25.44	3.53	18	-1.18	2.18
All	Sara	3663	0.75	0.43	-1.15	-0.67	111	24.74	4.32	22	-0.64	0.28
	Nana	3663	0.81	0.39	-1.56	0.44	111	26.66	5.54	27	-1.34	2.23
	Ivey	4257	0.78	0.41	-1.38	-0.08	129	25.89	4.56	22	-0.88	0.39
	Ken	4257	0.75	0.43	-1.18	-0.61	129	24.88	4.66	21	-0.59	-0.05
	Cali	4257	0.79	0.41	-1.42	0.00	129	26.03	3.85	23	-1.12	2.30

Table 4. 2 *Descriptive Statistics of Student Scores Averaged from Teachers' Ratings*

Course	Task	N	M	SD	Min	Max	Range	SK	KU
CA	1	37	25.62	4.74	11.50	32.50	21.00	-0.98	1.22
	2	37	25.09	4.14	11.50	32.50	21.00	-1.25	2.27
	3	37	26.38	4.61	11.00	33.00	22.00	-1.43	2.43
	All	37	25.70	3.52	15.50	30.50	15.00	-0.99	0.84
CB	1	43	24.63	4.37	11.00	30.67	19.67	-0.87	0.85
	2	43	27.38	2.60	20.00	31.67	11.67	-0.71	0.52
	3	43	24.80	3.04	16.33	30.33	14.00	-0.74	0.61
	All	43	25.60	2.59	17.22	30.67	13.44	-0.56	1.22

All in all, descriptive results revealed that by using the scale, the teachers generally diagnosed the students as having acceptable mastery of the writing skills under diagnosis and targeted the relatively wide and varied range of the student writing ability.

Furthermore, the students showed a fluctuating pattern of scores over the tasks with a slight increase in the scores on the final task.

4.2 Classical Test Theory Results

The CTT approach was employed to analyse the CA and CB datasets separately to investigate the scale functioning via item statistics, and the five teachers' rating behaviours using the percentage of absolute agreement. CTT results indicate (a) whether the scale appropriately assessed the student essays, partly responding to RQ1, and (b) whether the scale was applied consistently across the teacher raters, hence partly contributing to RQ2.

4.2.1 Descriptor Item Statistics

Table 4.3 shows item statistics of the descriptors estimated based on the three tasks in each course. The item statistics include the frequency (N) of the teacher ratings of the three-task essays on each rating option (0 and 1), the difficulty index (P), and the corrected item-total (CIT) correlation. As can be seen, the teachers in both courses assigned far more ratings on the 1-point (strong) option than the 0-point (weak) option for almost all descriptors.

The item difficulty (P) index is the frequency of the 1-point ratings of all essays divided by all the ratings on the descriptor (Clauser & Hambleton, 2018). The higher the index of a descriptor is, the easier the descriptor, and hence the stronger the student ability on such descriptor. For a norm-referenced assessment, the index over 0.8 is considered as too easy whereas the index below 0.2 is deemed as too difficult (Clauser & Hambleton, 2018). The easiest descriptor (P = 1.00) for the CA and CB students was D25 (*use of simple sentences*), suggesting that the students showed the strongest mastery on this descriptor. The hardest descriptor (P = 0.23) for the CA student was D28 (sentence accuracy) and the hardest descriptor (P = 0.13) for the CB students was D30 (word variety).

The CIT correlation illustrates if each descriptor functions consistently with the others in measuring the focal construct (Johnson & Morgan, 2016). For the CA course, D25 (*use of simple sentences*) was not included in the estimation as it was assigned only a score of 1 and thus was treated as an extreme item in the SPSS software. All the descriptor CIT correlations are positive and mostly over 0.2, overall suggesting that all descriptors collaborated well together to assess the defined construct (Johnson & Morgan, 2016).

The Cronbach's alpha coefficients for both courses are over the minimum requirement of 0.70 for low-stakes assessment (Nunnally & Bernstein, 1994), suggesting that the descriptors on the scale were internally consistent and homogenous in assessing the focal construct (Johnson & Morgan, 2016).

Table 4. 3 *CTT Item Statistics of Descriptors in Each Course*

Descriptors	CA				CB			
	N (0)	N (1)	P	CIT	N (0)	N (1)	P	CIT
Organisation	*338*	*1660*	*0.83*	*0.43*	*856*	*2627*	*0.75*	*0.39*
01. Essay topic introduction	12	210	0.95	0.13	4	383	0.99	0.25
02. Thesis statement relevance	24	198	0.89	0.30	59	328	0.85	0.30
03. Topic sentence relevance	28	194	0.87	0.49	70	317	0.82	0.43
04. Topic sentence specificity	46	176	0.79	0.56	81	306	0.79	0.50
05. Supporting idea relevance	31	191	0.86	0.56	52	335	0.87	0.38
06. Thesis restatement	50	172	0.77	0.57	204	183	0.47	0.43
07. Main idea summarisation	64	158	0.71	0.51	224	163	0.42	0.46
08. Essay ending	63	159	0.72	0.52	133	254	0.66	0.39
09. Essay length	20	202	0.91	0.23	29	358	0.93	0.36
Coherence	*161*	*505*	*0.76*	*0.48*	*287*	*874*	*0.75*	*0.36*
10. Supporting idea unity	25	197	0.89	0.38	20	367	0.95	0.29
11. Supporting idea logic	111	111	0.50	0.53	198	189	0.49	0.35
12. Main idea unity	25	197	0.89	0.53	69	318	0.82	0.44
Cohesion	*226*	*662*	*0.75*	*0.36*	*288*	*1260*	*0.81*	*0.43*
13. Supporting idea arrangement	65	157	0.71	0.33	48	339	0.88	0.48
14. Supporting idea connector	91	131	0.59	0.36	99	288	0.74	0.38
15. Main idea arrangement	23	199	0.90	0.44	50	337	0.87	0.45
16. Main idea connector	47	175	0.79	0.30	91	296	0.76	0.40
Content	*107*	*559*	*0.84*	*0.33*	*261*	*900*	*0.78*	*0.35*
17. Content comprehension	35	187	0.84	0.53	106	281	0.73	0.29
18. Content fulfilment	65	157	0.71	0.34	59	328	0.85	0.41
19. Content distribution	7	215	0.97	0.13	96	291	0.75	0.34
Grammar	*320*	*790*	*0.71*	*0.23*	*343*	*1592*	*0.82*	*0.20*
20. Part of speech	103	119	0.54	0.31	83	304	0.79	0.03
21. Subject-verb agreement	50	172	0.77	0.14	71	316	0.82	0.17
22. Tense and voice	55	167	0.75	0.26	72	315	0.81	0.31
23. Article	89	133	0.60	0.24	75	312	0.81	0.23
24. Pronoun	23	199	0.90	0.22	42	345	0.89	0.24
Sentence	*430*	*458*	*0.77*	*0.29*	*320*	*1228*	*0.73*	*0.05*
25. Use of simple sentence	0	222	1.00	n/a	1	386	1.00	0.12
26. Use of compound sentence	2	220	0.99	0.19	36	351	0.91	0.02
27. Use of complex sentence	35	187	0.84	0.19	21	366	0.95	0.01
28. Sentences accuracy	171	51	0.23	0.49	262	125	0.32	0.12
Vocabulary	*211*	*233*	*0.53*	*0.36*	*418*	*356*	*0.46*	*0.22*
29. Word choice	129	93	0.42	0.32	83	304	0.79	0.12
30. Word variety	82	140	0.63	0.40	335	52	0.13	0.31
Mechanics	*50*	*616*	*0.92*	*0.19*	*90*	*1071*	*0.92*	*0.11*
31. Punctuation	35	187	0.84	0.11	63	324	0.84	0.11
32. Capitalisation	9	213	0.96	0.25	18	369	0.95	0.10
33. Spelling	6	216	0.97	0.21	9	378	0.98	0.10
Alpha coefficients	0.84				0.79			

By and large, CTT results indicated that the students performed very weakly on the descriptors associated with sentence accuracy and vocabulary skills, and remarkably strongly on the descriptors associated with mechanical skills. Furthermore, the diagnostic rating scale was internally consistent and homogeneous in measuring the student ability.

4.2.2 Percentage of Interrater Agreement

The percent interrater agreement represents the degree of the raters' rating consensus. Two sets of the rater agreement indices were estimated based on the ratings assigned solely by the classroom teachers and the ratings given by both the classroom teachers and the researcher (Ake) as an additional rater. Thus, the rater agreement estimates in the CA course were calculated from: 1) the ratings of Sara (S) and Nana (N), and 2) the ratings of Sara, Nana, and Ake (A). In the CB course, the rater agreement percentages were estimated from: 1) the ratings of Ivey (I), Ken (K) and Cali (C), and 2) the ratings of Ivey, Ken, Cali, and Ake. The reader is reminded that the researchers' rating conditions differed significantly from those of the classroom teachers. Therefore, the focus is on the classroom teachers' agreement which is expected to represent the interrater reliability that could be achieved in the real-world classroom assessment.

Table 4.4 presents the percent interrater agreement on 33 descriptors. As shown in the final column, based on the average agreement across the tasks and both courses, 27 descriptors (82%) showed an average agreement over 70, indicating the teachers' congruent judgements on these descriptors. Six descriptors (D11, D14, D20, D23, D28, D29) exhibited the average agreement below 70%, indicating that the teachers were more variable in judging these descriptors. In general, the teachers were in accordance in interpreting the descriptors associate with mechanics and judged those associated with vocabulary and cohesion less homogenously.

Tables 4.5 and 4.6 lay out the percent interrater agreement on CA and CB students respectively. The tables also present individual students' total scores assigned by the teachers in each course. On the whole, the teachers were congruent in judging most of the students but showed significant disagreement with respect to only six students (ID: 10, 13, 18, 22, 28, 60). It can also be observed that the teachers generally judged higher-score students more homogeneously than lower-score students. In other words, higher-quality essays were rated more congruently than lower-quality essays.

Table 4. 4 *Percentage of Interrater Agreement on Descriptors*

Diagnostic Criteria	CA								CB								Average agree across tasks and courses	
	Task 1		Task 2		Task 3		All		Task 1		Task 2		Task 3		All			
	SN	*SNA*	SN	*SNA*	SN	*SNA*	SN	*SNA*	IKC	*IKCA*	IKC	*IKCA*	IKC	*IKCA*	IKC	*IKCA*	SNIKC	*SNIKCA*
01. Essay topic introduction	92	*89*	86	*86*	100	*100*	93	*92*	97	*95*	100	*100*	98	*98*	98	*97*	96	95
02. Thesis statement relevance	89	*82*	78	*72*	84	*82*	84	*79*	92	*90*	91	*91*	83	*78*	89	*86*	87	83
03. Topic-thesis relevance	86	*72*	86	*75*	84	*72*	86	*73*	85	*78*	75	*80*	72	*77*	77	*78*	82	76
04. Topic sentence specificity	78	*75*	73	*61*	78	*65*	77	*67*	77	*73*	71	*73*	66	*61*	71	*69*	74	68
05. Supporting idea relevance	78	*75*	81	*72*	95	*75*	85	*74*	88	*78*	77	*81*	78	*80*	81	*80*	83	77
06. Thesis restatement	86	*82*	76	*75*	78	*65*	80	*74*	71	*62*	67	*63*	67	*54*	68	*60*	74	67
07. Main idea summarisation	76	*72*	73	*58*	70	*51*	73	*60*	72	*65*	61	*60*	64	*60*	66	*62*	70	61
08. Essay ending	78	*75*	76	*58*	78	*65*	77	*66*	83	*73*	75	*67*	69	*58*	76	*66*	77	66
09. Essay length	92	*86*	86	*86*	100	*89*	93	*87*	92	*82*	97	*89*	92	*86*	94	*86*	94	87
10. Supporting idea unity	76	*61*	81	*51*	86	*44*	81	*52*	88	*70*	94	*74*	94	*68*	92	*71*	87	62
11. Supporting idea logic	62	*68*	68	*65*	59	*54*	63	*63*	57	*57*	60	*55*	69	*63*	62	*58*	63	61
12. Main idea unity	92	*72*	89	*72*	84	*51*	88	*65*	80	*72*	78	*73*	80	*77*	79	*74*	84	70
13. Supporting idea arrangement	76	*68*	54	*54*	76	*44*	68	*56*	81	*64*	86	*52*	81	*61*	83	*59*	76	58
14. Supporting idea connector	59	*65*	41	*47*	51	*47*	50	*53*	66	*66*	77	*58*	67	*60*	70	*61*	60	57
15. Main idea arrangement	89	*68*	70	*79*	89	*47*	83	*65*	81	*65*	85	*84*	78	*77*	81	*76*	82	71
16. Main idea connector	81	*68*	68	*79*	57	*51*	68	*66*	63	*55*	78	*70*	71	*54*	71	*60*	70	63
17. Content comprehension	81	*79*	84	*65*	84	*72*	83	*72*	78	*73*	58	*53*	66	*67*	67	*65*	75	69
18. Content fulfilment	57	*54*	59	*65*	73	*54*	63	*58*	81	*73*	81	*88*	71	*73*	78	*78*	71	68
19. Content distribution	97	*86*	92	*86*	97	*75*	96	*82*	69	*68*	88	*86*	77	*67*	78	*74*	87	78
20. Part of speech	54	*54*	68	*58*	62	*61*	61	*58*	75	*72*	74	*64*	77	*64*	75	*67*	68	63
21. Subject-verb agreement	70	*79*	70	*65*	62	*65*	68	*70*	69	*73*	94	*86*	86	*80*	83	*80*	76	75
22. Tense and voice	73	*75*	70	*61*	57	*61*	67	*66*	66	*66*	88	*81*	80	*77*	78	*75*	73	71
23. Article	54	*72*	62	*61*	51	*61*	56	*65*	61	*61*	89	*64*	72	*55*	74	*60*	65	63
24. Pronoun	89	*79*	84	*75*	86	*79*	86	*78*	71	*70*	94	*87*	95	*95*	87	*84*	87	81
25. Use of simple sentence	100	*100*	100	*100*	100	*100*	100	*100*	98	*98*	100	*100*	100	*100*	99	*99*	100	100
26. Use of compound sentence	97	*93*	100	*89*	97	*86*	98	*89*	81	*76*	81	*72*	83	*69*	82	*72*	90	81
27. Use of complex sentence	65	*72*	70	*72*	76	*79*	70	*74*	98	*98*	92	*93*	78	*84*	90	*92*	80	83
28. Sentence accuracy	62	*79*	76	*82*	46	*58*	61	*73*	66	*74*	55	*55*	58	*52*	60	*61*	61	67
29. Word choice	49	*47*	65	*65*	59	*68*	58	*60*	66	*58*	85	*69*	69	*53*	73	*60*	66	60
30. Word variety	84	*72*	62	*51*	62	*51*	69	*58*	77	*71*	75	*69*	92	*75*	81	*72*	75	65
31. Punctuation	57	*54*	81	*65*	89	*68*	76	*63*	72	*59*	88	*62*	85	*67*	81	*63*	79	63
32. Capitalisation	95	*93*	92	*93*	100	*96*	96	*94*	86	*86*	97	*98*	97	*97*	93	*94*	95	94
33. Spelling	95	*100*	97	*96*	97	*96*	96	*98*	97	*95*	98	*95*	100	*100*	98	*97*	97	98
Organisation (D01-D09)	84	*79*	79	*71*	85	*74*	83	*75*	84	*77*	79	*78*	77	*72*	80	*76*	82	76
Coherence (D10-D12)	77	*67*	79	*63*	76	*50*	77	*60*	75	*66*	77	*67*	81	*69*	78	*68*	78	64
Cohesion (D13-D16)	76	*67*	58	*65*	68	*47*	67	*60*	73	*63*	82	*66*	74	*63*	76	*64*	72	62
Content (D17-D19)	78	*73*	78	*72*	85	*67*	81	*71*	76	*71*	76	*76*	71	*69*	74	*72*	78	72
Grammar (D20-D24)	68	*72*	71	*64*	64	*65*	68	*67*	68	*68*	88	*76*	82	*74*	79	*73*	74	70
Sentence (D25-D28)	81	*86*	87	*86*	80	*81*	82	*84*	86	*87*	82	*80*	80	*76*	83	*81*	83	83
Vocabulary (D29-D30)	67	*60*	64	*58*	61	*60*	64	*59*	72	*65*	80	*69*	81	*64*	77	*66*	71	63
Mechanics (D31-D33)	82	82	90	85	95	87	89	85	85	80	94	85	94	88	91	85	90	85
Overall scale	78	75	76	71	78	68	77	71	78	73	82	76	79	72	80	74	79	73

Table 4. 5 *Percentage of Interrater Agreement on CA Students*

ID	Task 1					Task 2					Task 3					Mean agreement across tasks	
	Assigned score			Agreement		Assigned score			Agreement		Assigned score			Agreement			
	Sara	Nana	Ake	SN	SNA	Sara	Nana	Ake	SN	SNA	Sara	Nana	Ake	SN	SNA	SN	SNA
01	19	24	-	73	-	22	31	-	73	-	27	31	-	82	-	76	-
02	19	24	26	67	72	24	26	29	76	76	31	33	22	94	78	79	75
03	23	30	-	73	-	22	31	-	73	-	33	33	-	100	-	82	-
04	30	33	-	91	-	27	32	-	85	-	28	33	-	85	-	87	-
05	27	26	32	79	82	25	30	31	73	80	18	22	18	76	76	76	79
06	25	29	-	76	-	24	28	-	76	-	26	33	-	79	-	77	-
07	21	25	32	82	76	28	26	30	82	78	28	26	29	82	76	82	76
08	29	33	22	88	74	28	24	26	76	78	31	30	17	85	62	83	71
09	26	30	-	76	-	27	24	-	79	-	28	32	-	88	-	81	-
10	25	22	24	67	70	21	13	21	58	60	21	32	10	67	56	64	62
11	24	23	17	73	68	17	22	21	67	64	24	32	26	76	76	72	69
12	27	31	-	76	-	22	27	-	73	-	25	33	-	76	-	75	-
13	17	12	16	67	76	21	25	20	76	78	18	30	15	58	58	67	70
14	28	28	-	76	-	27	24	-	79	-	26	30	-	76	-	77	-
15	26	25	15	79	58	20	27	19	61	60	24	28	14	82	56	74	58
16	11	12	11	73	78	24	31	19	79	64	25	27	15	76	66	76	69
17	22	21	-	61	-	30	30	-	88	-	25	29	-	70	-	73	-
18	23	25	18	70	60	15	8	14	67	60	16	6	6	64	72	67	64
19	28	31	31	91	90	24	31	24	73	74	26	31	15	79	64	81	76
20	25	26	32	79	80	28	28	30	82	80	30	25	32	85	84	82	81
21	32	33	32	97	96	32	33	27	97	86	28	25	28	85	82	93	88
22	27	25	26	70	76	20	24	13	58	54	26	30	19	76	68	68	66
23	28	29	-	73	-	27	23	-	82	-	19	17	-	58	-	71	-
24	27	29	-	88	-	27	33	-	82	-	27	28	-	79	-	83	-
25	25	27	-	82	-	30	27	-	79	-	29	29	-	82	-	81	-
26	29	28	-	79	-	28	26	-	76	-	29	32	-	85	-	80	-
27	28	33	-	85	-	28	29	-	79	-	30	33	-	91	-	85	-
28	19	17	14	70	68	19	18	12	67	64	19	15	16	70	70	69	67
29	23	28	20	73	64	24	26	12	82	64	19	28	13	73	64	76	64
30	30	32	-	88	-	26	25	-	79	-	24	32	-	76	-	81	-
31	28	33	28	85	86	27	28	23	91	82	33	28	17	85	60	87	76
32	22	26	11	76	60	25	19	17	76	74	27	25	16	82	66	78	66
33	30	33	33	91	94	25	28	21	85	78	18	28	15	58	60	78	77
34	23	18	-	79	-	22	22	-	82	-	27	24	-	79	-	80	-
35	27	33	-	82	-	27	30	-	79	-	25	30	-	79	-	80	-
36	22	22	-	76	-	13	23	-	70	-	18	22	-	76	-	74	-
37	23	22	-	79	-	23	26	-	73	-	21	21	-	76	-	76	-
All	24.81	26.43	25.22	78.11	79.33	24.30	25.89	19.89	76.57	72.89	25.11	27.65	19.00	78.11	68.67	77.59	73.44

Table 4. 6 *Percentage of Interrater Agreement on CB Students*

ID	Task 1						Task 2						Task 3						Mean agreement across tasks	
	Assigned score				Agreement		Assigned score				Agreement		Assigned score				Agreement			
	I	K	C	A	IKC	IKCA	I	K	C	A	IKC	IKCA	I	K	C	A	IKC	IKCA	IKC	IKCA
38	21	14	17	9	67	65	28	28	27	22	88	78	26	19	26	14	78	63	77	69
39	22	12	22	17	66	59	26	23	29	24	86	78	26	19	23	22	76	68	76	68
40	26	23	22	-	86	-	30	25	31	-	82	-	26	23	25	-	82	-	83	-
41	24	18	26	-	76	-	28	26	29	-	80	-	27	25	27	-	80	-	78	-
42	16	20	21	-	76	-	31	25	30	-	84	-	27	21	26	-	78	-	79	-
43	28	18	24	-	78	-	32	25	28	-	80	-	24	19	22	-	74	-	77	-
44	24	19	22	13	74	58	28	24	26	24	80	75	27	22	25	19	76	65	76	66
45	29	25	25	27	84	83	25	25	29	22	84	73	22	25	17	17	82	76	83	77
46	11	12	10	-	80	-	22	25	23	-	74	-	15	17	20	-	84	-	79	-
47	15	19	23	-	72	-	31	26	27	-	84	-	28	28	26	-	86	-	80	-
48	28	24	25	28	74	76	30	30	31	23	92	83	31	27	29	23	84	77	83	79
49	25	22	24	-	82	-	28	24	25	-	78	-	24	22	25	-	78	-	79	-
50	15	15	19	12	78	73	25	22	27	12	82	69	26	27	25	9	80	63	80	68
51	18	21	20	8	72	62	26	26	21	18	78	71	27	24	27	15	82	69	77	68
52	28	28	30	17	82	67	30	25	27	24	84	76	28	27	28	27	86	82	84	75
53	20	16	25	-	68	-	29	19	30	-	74	-	30	25	26	-	80	-	74	-
54	26	17	25	23	80	75	28	30	27	26	82	75	28	24	29	21	82	73	81	74
55	30	19	23	-	76	-	30	22	28	-	80	-	27	29	26	-	88	-	81	-
56	25	20	26	23	80	75	28	24	24	26	84	81	18	17	14	15	72	70	78	75
57	29	15	24	-	66	-	25	22	18	-	80	-	25	21	22	-	82	-	76	-
58	28	25	24	11	82	61	29	21	28	25	70	71	22	26	18	15	80	71	77	68
59	32	28	27	28	78	80	22	15	23	10	74	68	20	27	27	16	72	68	74	72
60	29	30	16	-	60	-	18	24	27	-	72	-	18	21	24	-	68	-	66	-
61	28	32	24	28	76	79	29	33	30	13	90	65	21	29	25	24	72	69	79	71
62	27	32	33	-	88	-	33	31	29	-	90	-	30	27	31	-	84	-	87	-
63	28	30	27	30	86	84	32	30	29	25	88	81	25	23	29	16	70	65	81	77
64	26	30	27	-	82	-	30	28	29	-	88	-	26	31	26	-	86	-	85	-
65	28	30	27	28	78	80	29	27	28	20	82	74	27	29	29	26	88	81	82	78
66	19	23	21	-	74	-	32	23	32	-	82	-	21	23	22	-	76	-	77	-
67	26	23	24	-	82	-	25	28	29	-	78	-	15	20	24	-	72	-	77	-
68	28	31	27	26	82	80	27	23	30	18	82	72	28	26	25	25	80	78	81	76
69	24	27	27	-	76	-	29	27	30	-	84	-	25	27	27	-	80	-	80	-
70	29	30	32	30	86	89	33	24	31	21	78	72	22	30	28	29	82	78	82	80
71	21	27	27	-	72	-	33	29	33	-	92	-	21	28	29	-	78	-	80	-
72	24	28	24	19	72	65	30	28	29	24	80	73	17	27	27	26	76	69	76	69
73	30	29	28	-	88	-	29	29	27	-	80	-	28	27	28	-	82	-	83	-
74	33	33	25	32	84	86	31	32	31	32	92	92	27	32	32	33	86	89	87	89
75	31	26	27	-	82	-	20	26	26	-	78	-	23	26	27	-	80	-	80	-
76	28	18	26	15	78	66	21	28	29	19	78	69	21	23	27	29	82	79	79	71
77	29	30	30	-	88	-	33	28	31	-	88	-	17	26	24	-	74	-	83	-
78	28	26	25	30	84	79	30	26	29	25	84	78	28	28	27	28	84	79	84	79
79	27	33	31	-	86	-	27	31	27	-	84	-	22	29	27	-	78	-	82	-
80	29	26	29	14	86	69	30	30	29	32	86	88	20	23	23	10	78	62	83	73
All	25.40	23.81	24.67	21.27	78.30	73.23	28.19	25.98	27.98	22.05	82.23	75.55	24.09	24.86	25.44	20.86	79.49	72.45	79.67	73.73

It was also observed that the raters' agreements were relatively stable over the tasks, suggesting that they did not became more reliable over time. The raters agreements involving the researcher' ratings were mostly lower than those based merely on the classroom teachers' ratings, indicating that the researcher overall assigned lower ratings than the classroom teachers. This is probably because the researcher's judgement was not influenced by contextual factors (e.g., teaching workload) and thus might have more time look at student essays, in particular linguistic errors.

4.3 Many-Facets Rasch Results

The MFRM approach was employed to analyse the connected dataset, connecting the CA and CB datasets by the researcher (Ake)'s ratings, to investigate the psychometric properties of the six raters' rating behaviour, the scale's descriptor functioning, and the students' writing ability at both group and individual levels. MFRM results indicate: (a) whether the raters appropriately and consistently applied the scale and assessed the student ability, thereby answering RQ1 and RQ2 respectively, (b) whether the scale functioned and was applied appropriately and consistently, hence contributing information relevant to RQ1 and RQ2 respectively, and (c) how the students were diagnosed by the scale and raters over the tasks, thus contributing to RQ1, RQ2, and RQ4. In this section, results of Rasch assumption investigation are presented first, followed by group-level and individual-level Rasch statistics.

4.3.1 Rasch Assumptions

The global data-model, unidimensionality and local independence assumptions were first investigated to ensure meaningful interpretation of the Rasch results (Linacre, 1989; Rasch, 1960, 1980). The data-model fit also implies the unidimensional assessment of the scale and the unidimensionality in turn implies the local independence and vice versa (Fan & Bond, 2019). The global data-model fit was investigated through the Log-likelihood chi-square test, the standardised residuals, and the fit statistics. Psychometric unidimensionality was examined through the proportion of the variance explained by the Rasch model, the descriptor fit index, the descriptor point-measure (PTM) correlation, and the data-model fit indicators. Local independence was investigated through the fit statistics of the raters and descriptors as well as the unidimensionality indicators.

4.3.1.1 Global Data-Model Fit

Table 4.7 shows global data-model fit indicators. As shown in the table, the chi-square test of goodness of fit was not statistically significant ($\chi^2(24030) = 21188.5098$, $p = 1.0000$), suggesting the satisfactory global fit of the observed data to the expected data generated by Rasch model (Linacre, 2018). The percentage of unexpected standardised residuals outside ±2 (4.7%) was less than the suggested maximum of 5 and the percentage of unexpected standardised residuals outside ±3 (1.3%) was slightly over the suggested maximum of 1, indicating a satisfactory level of global data-model fit (Linacre, 2018).

The mean and standard deviation of sample standardised residuals were very close to the expected values of 0 and 1 respectively, indicating a satisfactory data-model fit (Linacre, 2018). The means of the weighted mean square (Infit MS) fit statistics of the raters, descriptors, and students were within the productive or acceptable range of 0.50 - 1.50 (Linacre, 2018), thus indicating a satisfactory global data-model fit. The data-model fit indicators indicate that the differences between the observed scores and the expected scores generated by the Rasch model were acceptably small, indicating the scores yielded in the assessment conditions satisfactorily matched the Rasch expectations (Eckes, 2015).

Table 4. 7 *Rasch Indicators of Global Data-Model Fit*

Indicators	Value
Chi-square goodness-of-fit test	
• Log-likelihood chi-square test	21188.5098
• Approximate degrees of freedom	24030
• Significance probability	1.0000
Standardised residual indices	
• Valid responses used for estimation	24156
• Unexpected standardised residuals outside ±2	1146 (4.7%)
• Unexpected standardised residuals outside ±3	333 (1.3%)
• Mean of sample standardised residuals	0.00
• Sample standard deviation of standardised residuals	0.99
Fit statistics	
• Mean of rater Infit MS fit indices	1.00
• Mean of descriptor Infit MS fit indices	1.00
• Mean of student Infit MS fit indices	1.00

4.3.1.2 Psychometric Unidimensionality

As presented in Table 4.8, the Rasch model accounted for about 24% of the total variance, greater than the suggested minimum of 20% to ensure acceptable calibration and unidimensional measurement of the focal construct (Reckase, 1979, p. 227-228). All

descriptors showed Infit MS indices within the acceptable range, suggesting there were not underfitting or misfitting descriptors targeting dimensions other than the prime dimension of the construct (Linacre, 2018). The mean observed PTM correlation of the descriptors was over the suggested minimum of 2.0, suggesting that the scale internally targeted the focal construct (Schumacker, 2004).

Table 4. 8 *Rasch Indicators of Unidimensionality and Local Independence*

Indicators	Value
Proportions of variance	
• Responses used for estimation	24156 (*M* = 0.75, *SD* = 0.43)
• Raw-score variance of observations	0.19 (100%)
• Variance explained by Rasch measures	0.04 (23.6%)
• Residual variance (systematic and random error)	0.14 (76.4%)
Descriptor statistics	
• Number of fitting descriptors (Infit 0.50 - 1.50)	33 (100%)
• Number of underfitting descriptors (Infit > 1.50)	0
• Number of overfitting descriptors (Infit < 0.50)	0
• Mean of descriptor Infit MS fit indices	1.00
• Mean of observed descriptor PTM correlations	0.27
Rater statistics	
• Number of fitting raters (Infit 0.50 - 1.50)	6 (100%)
• Number of underfitting raters (Infit > 1.50)	0
• Number of overfitting raters (Infit < 1.50)	0
• Mean of raters Infit MS fit indices	1.00
• Rater Rasch-Kappa	0.15

4.3.1.3 Local Independence

With regard to local independence, no raters or descriptors showed Infit MS indices under 0.50, indicating no overfitting raters and descriptors generating redundant or dependent scores (Linacre, 2018). The Rasch-Kappa of 0.15, slightly over 0, indicates that the six raters did not exhibit an overly high degree of rating dependence (Eckes, 2015). All in all, the local independence indicators suggest that the scores generated by each rater or descriptor did not influence the scores generated by the others. In other words, the raters and descriptors were independent of one another in generating the observed scores under the current assessment (Fan & Bond, 2019).

In short, the acceptable global data-model fit, psychometric unidimensionality, and local independence together support that the scale and raters provided diagnostic scores which corresponded to the expected Rasch model, captured the defined construct without significantly targeting any other irrelevant dimensions, and were sufficiently

independent of one another in generating the diagnostic scores. Accordingly, the Rasch statistics in this MFRM analysis can be meaningfully interpreted and the parameter estimates can be generalised to other samples (Andrich, 1988; Rasch, 1960, 1980).

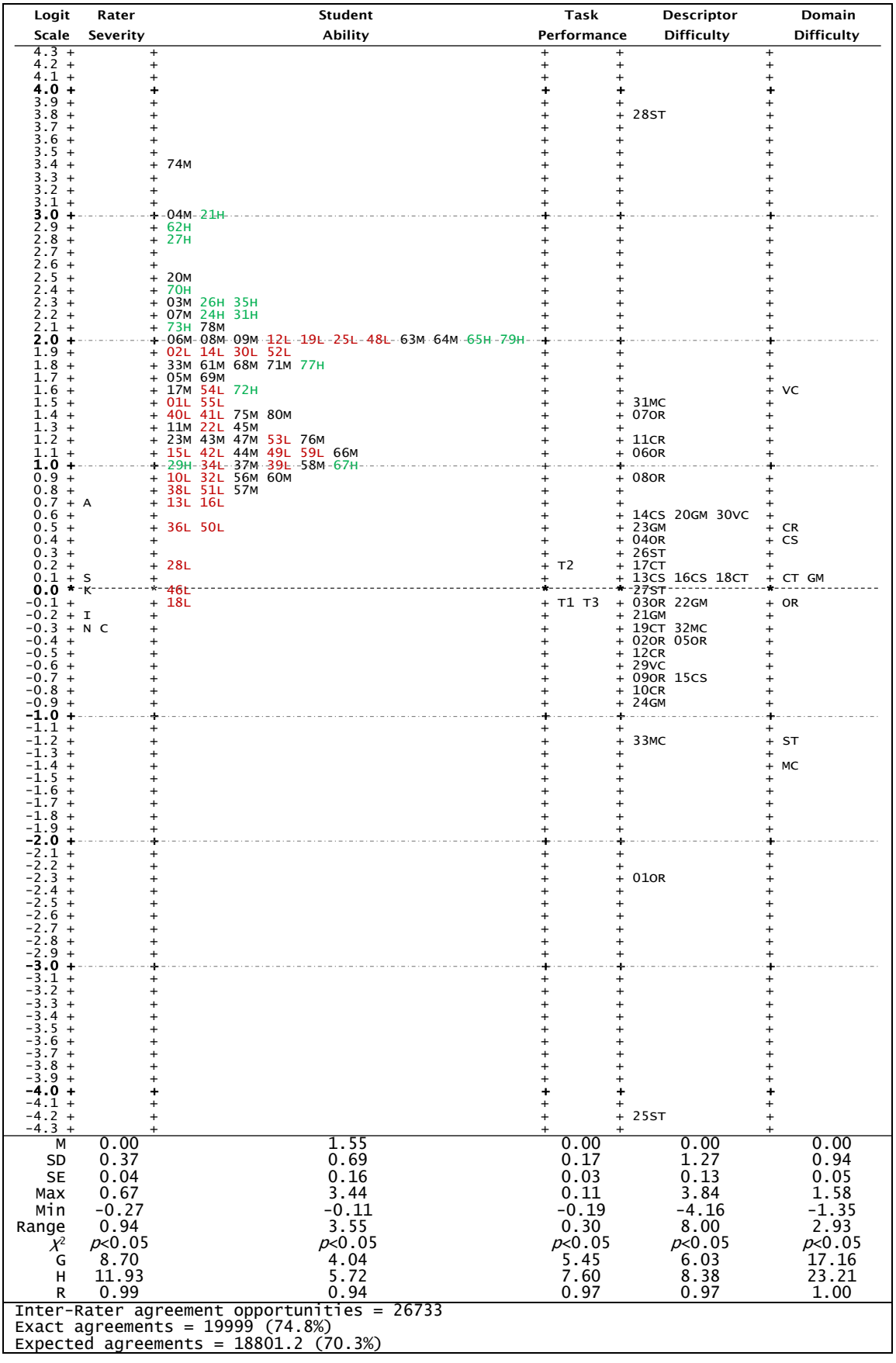
4.3.2 Group-Level Rasch Results

The variable map in Figure 4 provides a summary of how the elements within each facet are calibrated onto the common standardised log-odds unit (logit) scale in the first column, allowing for comparisons within and between facets (Linacre, 2018). In particular, this information provides a global pecture of the rater behaviour, scale functioning, and student writing ability. The logit scale is centred at 0 and ranges from the highest logit of 4.3 at the top to the lowest logit of -4.3 at the bottom. The student facet was allowed to float along the logit scale and the rater facet was negatively oriented. The higher the logits are from 0, the more severe the raters, the harder the descriptors and domains, and the higher the student ability. Conversely, the lower the logits are from 0, the more lenient the raters, the easier the descriptors and domains, and the lower the student ability. If a student shows the exact same logit as a particular rater or descriptor, this means that the student has the 50/50 probability of being judged by the rater as mastering the descriptor.

4.3.2.1 Rater Group Behaviour

The rater logits, distributed in the second column in Figure 4.1, represent individual raters' severity on the logit scale. Overall, the rater logits were noticeably scattered from one another, indicating substantial variability in severity within this group of raters (Linacre, 2018). Since the rater facet was centred, the average severity was 0 on the logit scale. The standard error mean of the severity logit (SE = 0.04) was very close to the expected value of 0, indicating a precise estimation of rater logits (Linacre, 2018). The standard deviation (SD = 0.37) was somewhat greater than the average severity, indicating substantial variability in rater severity (Linacre, 2018). The significant fixed chi-square test of the rater homogeneity ($\chi^2(5) = 432.7$, $p < 0.05$) indicates that at least two raters' severity logits were statistically different (Linacre, 2018).

Figure 4. 1 *Visual Variable Map Displaying Logit Locations of Individual Facets*



M	0.00	1.55	0.00	0.00	0.00
SD	0.37	0.69	0.17	1.27	0.94
SE	0.04	0.16	0.03	0.13	0.05
Max	0.67	3.44	0.11	3.84	1.58
Min	-0.27	-0.11	-0.19	-4.16	-1.35
Range	0.94	3.55	0.30	8.00	2.93
χ^2	p<0.05	p<0.05	p<0.05	p<0.05	p<0.05
G	8.70	4.04	5.45	6.03	17.16
H	11.93	5.72	7.60	8.38	23.21
R	0.99	0.94	0.97	0.97	1.00

Inter-Rater agreement opportunities = 26733
Exact agreements = 19999 (74.8%)
Expected agreements = 18801.2 (70.3%)

The high rater separation strata (H_R = 11.93) indicates that the six raters were separated into about 12 statistically distinct classes of severity, too far from the expected strata of 1 for homogenous severity (Eckes, 2015). The high rater separation ratio (G_R = 8.7) indicates that the severity variability was almost nine times larger than their measure precision or standard error mean, thereby suggesting a wide spread of the severity (Eckes, 2015). The high rater separation reliability (R_R = 0.99), over the expected value of 0 for homogeneous severity, suggests that the six raters were reliably different in severity when judging the descriptors and student essays (Eckes, 2015).

In addition, the raters exhibited the acceptable rating consensus. Of 26,730 interrater agreement opportunities, the observed exact agreement was 19,999 (74.7%) slightly higher than the expected exact agreement of 18,801.2 (70.4%) by about 4%. This suggests that the six raters tended to rate the student essays as trained expert raters should, trying to achieve congruent judgements, while at the same time maintaining independence to a certain extent (Linacre, 2018). The Rasch-Kappa of 0.15 was slightly over 0, indicating that the raters did not exhibit an overly high degree of interrater agreement and rating dependency and thus their ratings were in line with the local independence assumption (Fan & Bond, 2019).

By and large, the global variable map and Rasch indicators largely confirm that the six raters differed significantly and reliably in their rating severity, thereby suggesting that they interpreted the descriptors differently and were not behaving interchangeably in the current assessment context (Engelhard, 2013; Myford & Wolfe, 2003, 2004).

4.3.2.2 Student Group Ability

The student logits, scattered on the logit scale in the third column in Figure 4.1, represent individual students' writing ability, which is the latent variable of interest in the present diagnostic assessment. The student writing ability was calibrated over the three sequential writing assignments. Individual students were labelled as high (H), mid (M), and low (L) achieving levels based on their total scores on summative achievement exams made and judged by their classroom teachers.

On the whole, the locations of the student logits were consistent with the students' achievement levels; that is, those students who achieved at a high level on the summative exam were generally located higher whereas lower-achieving students were positioned

lower on the logit scale. This contributes to the predictive validity (Linacre, 2018) of the formative diagnostic assessment and suggests the differential correspondence between the formative diagnostic assessment and the summative achievement assessment. Further, the student logits were closely spaced and widely scattered, suggesting that the raters and the scale targeted the substantial variability of the student writing ability (Aryadoust, 2009; Baghaei, 2008).

The significant fixed chi-square test indicated that there were statistically significant differences at least between two student mastery logits ($\chi^2(79) = 1363.7$, $p < 0.05$) (Eckes, 2015). The high separation ratio ($G_S = 4.03$), strata ($H_S = 5.71$), and reliability ($R_S = 0.94$) indices showed that this group of students were differentially and reliably assessed by the raters and the scale (Eckes, 2015). When the student differentiation is of interest in an assessment, the examinee separation ratio should be at least 2 with the separation reliability of at least 0.8 to show that an assessment can reliably distinguish test-takers into two distinct (high and low) performers (Eckes, 2015).

The average ability logit of the students was markedly high ($M = 1.55$, $SD = 0.69$), about 1.5 higher than the average severity logit. This indicates that this group of students was largely assigned high ratings by the raters. The student writing ability spanned the entire range of 3.55 logits. The highest-ability student (74M) showed a logit value of 3.44 and the lowest-ability student (18L) exhibited a logit value of -0.11. This indicates that the raters and the scale could target different levels of ability within the writing construct under measure.

The task logits, shown in the fourth column and Table 4.9, were centred at 0 and deemed to represent the average writing ability of the students over the tasks. Although the students were expected to make progress over the tasks, the teacher-made tasks were rather different in terms of the prompt characteristics and genres. The current study design was also not actually suitable to trace learner development. It is, therefore, not clear whether differences between the task logits represent changes in writing ability or variations in task difficulty.

Table 4. 9 *Rasch Statistics of Teacher-Made Writing Assignment Tasks*

Task	Ave Obs	Ave Fair	Rasch Logit	Logit SE	Infit MS	Outfit MS	PTM Obs	PTM Exp	D Index
1	0.74	0.81	-0.08	0.03	1.01	1.12	0.41	0.42	0.97
2	0.78	0.85	0.19	0.03	1.00	1.05	0.40	0.40	0.99
3	0.74	0.81	-0.11	0.03	0.98	0.90	0.44	0.43	1.04

4.3.2.3 Scale Functioning

The descriptor logits, portrayed in the fifth column in Figure 4.1, represent individual descriptors' difficulty on the logit scale. Each descriptor is labelled with its associated domain. Descriptors located higher on the logit scale were associated with more severe and lower ratings, whereas those located lower on the logit scale were associated with less severe and higher ratings. The mean difficulty logit was centred at 0 with a wide range of up to about 8.0 logits. Without the outlying descriptors, including D28 (*sentence accuracy*), D01 (*topic introduction*), and D25 (*use of simple sentence*), the difficulty logit would range between 1.53 (D31) and -1.2 (D33), hence narrowing the difficulty range to about 2.73 logits. This suggests that the scale contained the range of descriptor difficulty needed to identify different levels of student writing ability (Aryadoust, 2009; Baghaei, 2008). The distribution of the difficulty logits was lower than that of the student logits, suggesting that most of the students were diagnosed as having satisfactorily mastered most of the descriptors over the three tasks.

The significant fixed chi-square test ($\chi^2(32) = 3274.1$, $p < 0.05$) indicates that there were significant differences at least between two difficulty logits (Eckes, 2015). The high separation ratio ($G_D = 6.03$) suggests that the variability of the descriptor difficulty logits was almost six times larger than the precision ($SE = 0.13$) of the logits (Eckes, 2015). The high separation strata ($H_D = 8.38$) illustrates that there were almost six statistically distinct classes of difficulty logits as highly reliably differentiated ($R_D = 0.97$) by the six raters and student performances (Eckes, 2015).

The domain logits, shown in the final column, show the difficulty levels of the eight criteria domains: organisation (OR), coherence (CR), cohesion (CS), content (CT), Grammar (GM), sentence (ST), vocabulary (VC), and mechanics (MC). The domain facet was treated as a dummy facet given that domains were simply groupings of the associated descriptors. Vocabulary appeared to be the most difficult domain (*Logit* = 1.58) whereas mechanics appeared to be the easiest domain (*Logit* = -1.35) for this group of students. All in all, the

variable map and global Rasch indicators reveal that the scale has varying levels of descriptor difficulty.

4.3.3 Individual-Level Rasch Results

The micro Rasch statistics of the five facets demonstrated individual raters' behaviours, individual students' writing ability, and individual descriptors' functioning. The rater statistics are presented first, followed by the student and descriptor statistics.

4.3.3.1 Individual Rater Behaviours

Table 4.10 lays out the micro Rasch statistics of each rater arranged from the highest to the lowest logits. As can be seen, the rater fair average indices were different, suggesting differences in their severity. All rater Infit MS fit indices fell within the acceptable range, indicating that the ratings assigned by each rater satisfactorily matched the expected ratings generated by the Rasch model (Linacre, 2018). Accordingly, the rater severity logits were highly accurate and each rater was self-consistent in his or her ratings across the students and descriptors in the current assessment. The good rater fit indices also suggested that each rater consistently interpreted the descriptors (Engelhard, 2013).

All in all, individual raters demonstrated varying levels of rating severity but each rater was self-accurate and self-consistent in applying the scale and in diagnosing the students. As the raters were trained to achieve high interrater agreement, they tended to rate the student essays in a relatively machine-like fashion while maintaining a certain degree of independent expert rating.

Table 4. 10 *Rasch Statistics of Individual Raters*

Rater	Ave Obs	Ave Fair	Rasch Logit	Logit SE	Infit MS	Outfit MS	PTM Obs	PTM Exp	Agree Obs	Agree Exp	D Index
Ake	0.64	0.71	0.67	0.04	1.04	0.96	0.44	0.46	67.80	66.90	0.95
Sara	0.75	0.80	0.15	0.04	0.99	1.07	0.43	0.42	75.30	70.90	1.01
Ken	0.75	0.83	-0.02	0.04	1.02	1.07	0.38	0.40	78.10	71.40	0.96
Ivy	0.79	0.86	-0.25	0.04	0.95	0.87	0.41	0.38	76.30	70.80	1.07
Nana	0.81	0.86	-0.27	0.05	1.09	1.26	0.34	0.39	73.30	71.00	0.88
Cali	0.79	0.86	-0.27	0.04	0.91	0.94	0.42	0.38	77.40	71.40	1.10

4.3.3.2 Individual Student Ability

Table 4.11 shows the Rasch statistics of individual students ordered from the highest to the lowest ability logits. As the ability logits are corrected to accommodate the raters' severity differences, they provide more accurate, reliable, and fair diagnostic results than raw scores (Linacre, 2018).

Table 4. 11 *Student Exam Scores, Diagnostic Scores, and Rasch Indices*

ID	Achievement exam scores					Diagnostic scores				Rasch estimates		
	Mid	Final	Total	(%)	Level	Task1	Task2	Task3	M	Logit	SE	Infit
74	24.58	20.00	44.58	74.30	M	30.33	31.33	30.33	30.67	3.44	0.23	1.09
21	13.00	13.00	26.00	86.67	H	32.50	32.50	26.50	30.50	3.04	0.22	1.09
04	11.40	10.80	22.20	74.00	M	31.50	29.50	30.50	30.50	2.99	0.28	0.94
62	23.75	24.00	47.75	79.58	H	30.67	31.00	29.33	30.33	2.89	0.23	0.96
27	13.00	13.00	26.00	86.67	H	30.50	28.50	31.50	30.17	2.83	0.27	1.05
20	9.50	12.00	21.50	71.67	M	25.50	28.00	27.50	27.00	2.49	0.18	1.17
70	25.83	26.00	51.83	86.38	H	30.33	29.33	26.67	28.78	2.36	0.16	1.09
03	9.90	12.30	22.20	74.00	M	26.50	26.50	33.00	28.67	2.29	0.23	1.19
26	11.50	13.00	24.50	81.67	H	28.50	27.00	30.50	28.67	2.29	0.23	1.10
35	12.00	12.00	24.00	80.00	H	30.00	28.50	27.50	28.67	2.29	0.23	1.06
24	11.50	11.00	22.50	75.00	H	28.00	30.00	27.50	28.50	2.24	0.22	1.04
07	10.50	9.90	20.40	68.00	M	23.00	27.00	27.00	25.67	2.16	0.17	1.13
31	12.00	13.00	25.00	83.33	H	30.50	27.50	30.50	29.50	2.16	0.17	1.09
78	23.33	21.00	44.33	73.88	M	26.33	28.33	27.67	27.44	2.08	0.15	0.98
73	27.08	26.00	53.08	88.47	H	29.00	28.33	27.67	28.33	2.06	0.18	0.95
09	10.95	10.10	21.05	70.17	M	28.00	25.50	30.00	27.83	2.05	0.21	1.03
19	8.00	11.00	19.00	63.33	L	29.50	27.50	28.50	28.50	2.05	0.16	1.05
25	9.00	10.00	19.00	63.33	L	26.00	28.50	29.00	27.83	2.05	0.21	1.05
65	22.08	24.00	46.08	76.80	H	28.33	28.00	28.33	28.22	2.04	0.15	1.01
64	20.00	21.00	41.00	68.33	M	27.67	29.00	27.67	28.11	2.03	0.18	0.89
79	24.17	24.00	48.17	80.28	H	30.33	28.33	26.00	28.22	2.03	0.18	0.94
08	10.20	10.20	20.40	68.00	M	31.00	26.00	30.50	29.17	2.02	0.16	1.07
48	16.00	20.63	36.63	61.05	L	25.67	30.33	29.00	28.33	2.01	0.15	0.96
06	9.90	9.90	19.80	66.00	M	27.00	26.00	29.50	27.50	1.96	0.21	0.95
12	8.40	10.20	18.60	62.00	L	29.00	24.50	29.00	27.50	1.96	0.21	0.98
63	20.83	21.00	41.83	69.72	M	28.33	30.33	25.67	28.11	1.95	0.14	1.01
02	8.10	11.10	19.20	64.00	L	21.50	25.00	32.00	26.17	1.87	0.16	1.04
14	9.00	10.20	19.20	64.00	L	28.00	25.50	28.00	27.17	1.87	0.20	0.99
30	9.00	10.00	19.00	63.33	L	31.00	25.50	28.00	28.17	1.87	0.20	1.04
52	19.08	19.13	38.21	63.68	L	28.67	27.33	27.67	27.89	1.85	0.14	0.94
71	22.08	22.00	44.08	73.47	M	25.00	31.67	26.00	27.56	1.85	0.17	0.97
77	23.33	26.00	49.33	82.22	H	29.67	30.67	22.33	27.56	1.85	0.17	0.92
33	9.00	11.00	20.00	66.67	M	31.50	26.50	23.00	27.00	1.80	0.15	1.02
61	22.25	19.00	41.25	68.75	M	28.00	30.67	25.00	27.89	1.79	0.14	1.04
05	11.10	10.50	21.60	72.00	M	26.50	27.50	20.00	24.67	1.75	0.15	1.01
68	20.83	22.00	42.83	71.38	M	28.67	26.67	26.33	27.22	1.75	0.14	0.99
69	23.33	20.00	43.33	72.22	M	26.00	28.67	26.33	27.00	1.71	0.16	0.99
17	9.00	10.80	19.80	66.00	M	21.50	30.00	27.00	26.17	1.64	0.19	1.04
54	17.60	17.63	35.23	58.72	L	22.67	28.33	27.00	26.00	1.61	0.13	1.01
72	23.75	27.00	50.75	84.58	H	25.33	29.00	23.67	26.00	1.56	0.13	1.04
01	7.80	7.50	15.30	51.00	L	21.50	26.50	29.00	25.67	1.53	0.19	1.08
55	12.42	18.75	31.17	51.95	L	24.00	26.67	27.33	26.00	1.48	0.16	0.88
75	22.92	19.00	41.92	69.87	M	28.00	24.00	25.33	25.78	1.43	0.15	0.94

ID	Achievement exam scores					Diagnostic scores				Rasch estimates		
	Mid	Final	Total	(%)	Level	Task1	Task2	Task3	M	Logit	SE	Infit
80	23.33	21.00	44.33	73.88	M	28.00	29.67	22.00	26.56	1.42	0.13	0.93
40	18.08	19.88	37.96	63.27	L	23.67	28.67	24.67	25.67	1.41	0.15	0.84
41	16.00	19.50	35.50	59.17	L	22.67	27.67	26.33	25.56	1.38	0.15	0.97
22	8.00	8.00	16.00	53.33	L	26.00	22.00	28.00	25.33	1.33	0.14	1.08
45	19.67	19.88	39.55	65.92	M	26.33	26.33	21.33	24.67	1.31	0.13	0.93
11	9.30	10.80	20.10	67.00	M	23.50	19.50	28.00	23.67	1.25	0.14	1.13
76	25.00	19.00	44.00	73.33	M	16.33	24.67	26.00	22.33	1.24	0.13	1.02
47	18.42	21.00	39.42	65.70	M	19.67	24.33	26.00	23.33	1.22	0.15	0.88
23	9.00	12.00	21.00	70.00	M	28.67	27.33	27.67	27.89	1.17	0.18	0.97
43	21.83	19.13	40.96	68.27	M	20.33	26.00	27.00	24.44	1.16	0.15	0.86
53	16.58	18.75	35.33	58.88	L	22.67	28.33	27.00	26.00	1.16	0.15	0.98
59	19.58	17.00	36.58	60.97	L	24.00	26.67	27.33	26.00	1.11	0.12	1.07
15	8.70	10.20	18.90	63.00	L	23.67	25.33	16.33	21.78	1.10	0.14	0.96
42	17.58	20.25	37.83	63.05	L	22.67	21.67	22.67	22.33	1.09	0.14	0.89
66	21.67	19.00	40.67	67.78	M	25.67	26.00	22.00	24.56	1.09	0.14	0.97
44	19.42	22.88	42.30	70.50	M	29.00	20.00	24.67	24.56	1.08	0.12	0.92
49	16.67	18.75	35.42	59.03	L	25.00	23.00	21.00	23.00	1.07	0.14	0.90
67	24.58	25.00	49.58	82.63	H	28.00	30.67	25.00	27.89	1.03	0.14	0.96
58	19.08	22.50	41.58	69.30	M	30.67	31.00	29.33	30.33	1.02	0.12	1.03
29	11.00	12.00	23.00	76.67	H	28.33	30.33	25.67	28.11	1.00	0.14	0.97
39	14.50	20.25	34.75	57.92	L	27.67	29.00	27.67	28.11	0.96	0.12	0.99
34	7.50	10.00	17.50	58.33	L	28.33	28.00	28.33	28.22	0.95	0.17	0.91
37	10.00	10.00	20.00	66.67	M	21.00	29.00	22.00	24.00	0.95	0.17	0.96
10	8.70	10.10	18.80	62.67	L	24.33	27.33	19.67	23.78	0.93	0.14	1.16
32	8.00	10.00	18.00	60.00	L	28.67	26.67	26.33	27.22	0.89	0.14	1.07
56	19.17	20.63	39.80	66.33	M	26.00	28.67	26.33	27.00	0.89	0.12	1.04
60	20.00	21.00	41.00	68.33	M	30.33	29.33	26.67	28.78	0.89	0.14	1.06
51	16.33	19.88	36.21	60.35	L	25.00	31.67	26.00	27.56	0.82	0.12	0.97
57	20.58	20.25	40.83	68.05	M	25.33	29.00	23.67	26.00	0.77	0.14	1.04
38	17.92	18.00	35.92	59.87	L	29.00	28.33	27.67	28.33	0.76	0.12	0.91
16	9.30	9.60	18.90	63.00	L	30.33	31.33	30.33	30.67	0.67	0.13	0.98
13	9.15	10.20	19.35	64.50	L	28.00	24.00	25.33	25.78	0.65	0.13	0.96
50	17.83	18.38	36.21	60.35	L	24.00	26.00	23.67	24.56	0.54	0.12	0.92
36	9.00	10.00	19.00	63.33	L	29.67	30.67	22.33	27.56	0.51	0.16	0.91
28	7.00	8.00	15.00	50.00	L	26.33	28.33	27.67	27.44	0.21	0.13	0.96
46	13.42	20.25	33.67	56.12	L	30.33	28.33	26.00	28.22	-0.05	0.13	1.10
18	6.50	8.00	14.50	48.33	L	28.00	29.67	22.00	26.56	-0.11	0.13	1.16
M	15.34	16.15	31.49	68.03		25.09	26.32	25.53	25.65	1.55	0.16	1.00
SD	6.02	5.58	11.39	8.96		4.54	3.56	3.90	3.04	0.69	0.04	0.08

The table also includes the students' achievement exam scores and formative diagnostic scores. The maximum scores for the midterm and final exams for the CA course (Student ID: 01-37) are 15 each and for the CB course (Student ID: 38-80) are 30 each. Of the total exam, students receiving 75% and above, between 65% and 74%, and less than 65% were classified as high-, mid- and low-ability students respectively. The diagnostic scores include individual students' scores on each task and the average score of all tasks. The CA students' diagnostic scores were averaged from the two teachers' ratings whereas the CB student's diagnostic scores were averaged from the three teachers' ratings.

As shown in the table, the student mean logit was rather high (*M* = 1.55, *SD* = 0.69) with the mean standard error of estimate very close to 0 (*SE* = 0.16), indicating the precise estimation of the ability logits (Linacre, 2018). The student logits span the noticeable range of 3.55 logits with the 74th student having the highest ability (*Logit* = 3.44, *SE* = 0.23) and the 18th student showing the lowest ability (*Logit* = -0.11, *SE* = 0.13).

The Infit MS indices of all students fell within 0.80 - 1.20. This implies that individual students were consistently judged by virtue of no Infit MS values over 1.50, and the raters and the scale were able to differentiate the student writing ability with sufficient variability by reason of no Infit MS values below 0.50 (Engelhard, 2013). Since the rating design was not fully crossed (not all raters rated all students), the interpretation of the student PTM correlation is not trustworthy (J. M. Linacre, 2018, personal communication, December 4, 2018) and thus was not used for examining the student facet.

All in all, almost all students showed rather high logits, indicating their satisfactory mastery of the writing skills over the course. The student achievement levels, diagnostic scores, and ability logits were generally consistent with each other, meaning that the students receiving higher diagnostic scores and logits tended to gain higher end-of-course achievement scores.

4.3.3.3 Individual Descriptor Functioning

Table 4.12 shows the micro Rasch statistics of the descriptors arranged from the highest to the lowest logit values. Overall, the mean observed average (*M* = 0.75, *SD* = 0.17) and the mean fair average (*M* = 0.78, *SD* = 0.17) were relatively close, suggesting the six raters assigned somewhat similar ratings to the descriptors.

The difficulty logit mean was centred at 0 with the relatively high standard deviation of 1.27 logits. The most difficult descriptor (*Logit* = 3.84) was D28 (sentence accuracy) whereas the easiest descriptor (*Logit* = -4.16) was D25 (*use of simple sentences*). This means that most of the students produced more inaccurate, erroneous, or ungrammatical sentences than considered acceptable by the raters. D25 (*use of simple sentences*) was the most commonly-mastered descriptor and was assigned a score of 1 for virtually all students across the three tasks and six raters. This means that almost all students used simple sentences in their essays on the three tasks.

Table 4. 12 *Rasch Statistics of Individual Descriptors and Domains*

Descriptors (writing skills)	Ave Obs	Ave Fair	Rasch Logit	Logit SE	Infit MS	Outfit MS	PTM Obs	PTM Exp	D Index
28. Sentence accuracy	0.27	0.09	3.84	0.09	1.05	1.11	0.24	0.31	0.89
31. Punctuation	0.77	0.50	1.53	0.09	1.05	1.13	0.23	0.30	0.92
07. Main idea summarisation	0.54	0.53	1.43	0.08	0.98	0.97	0.37	0.34	1.11
11. Supporting idea logic	0.46	0.58	1.23	0.08	0.96	0.96	0.39	0.34	1.18
06. Thesis restatement	0.60	0.61	1.12	0.08	1.02	1.00	0.32	0.33	0.96
08. Essay ending	0.65	0.65	0.91	0.08	1.00	1.01	0.32	0.33	0.98
20. Part of speech	0.67	0.72	0.59	0.08	1.12	1.20	0.17	0.32	0.67
30. Word variety	0.36	0.72	0.58	0.08	1.08	1.17	0.21	0.33	0.71
14. Supporting idea connector	0.63	0.73	0.57	0.08	0.93	0.91	0.42	0.33	1.23
23. Article	0.68	0.73	0.54	0.08	1.07	1.14	0.22	0.32	0.81
04. Topic sentence specificity	0.75	0.76	0.37	0.09	0.88	0.80	0.46	0.30	1.21
26. Use of compound sentence	0.90	0.78	0.28	0.13	1.01	0.92	0.23	0.22	1.00
17. Content comprehension	0.75	0.79	0.20	0.09	0.97	0.91	0.35	0.30	1.07
16. Main idea connector	0.72	0.81	0.10	0.09	0.97	0.95	0.35	0.31	1.07
18. Content fulfilment	0.77	0.81	0.08	0.09	0.96	1.02	0.33	0.30	1.04
13. Supporting idea arrange	0.73	0.82	0.05	0.09	0.90	0.91	0.43	0.31	1.18
27. Use of complex sentence	0.92	0.82	0.04	0.14	1.12	1.60	-0.01	0.20	0.88
22. Tense and voice	0.79	0.83	-0.06	0.09	1.11	1.23	0.13	0.29	0.83
03. Topic sentence relevance	0.82	0.84	-0.08	0.10	0.93	0.81	0.39	0.28	1.10
21. Subject-verb agreement	0.81	0.85	-0.19	0.10	1.12	1.24	0.11	0.28	0.83
19. Content distribution	0.82	0.86	-0.28	0.10	0.97	0.95	0.31	0.27	1.04
32. Capitalisation	0.95	0.87	-0.32	0.17	1.01	0.93	0.16	0.17	1.00
05. Supporting idea relevance	0.85	0.87	-0.36	0.11	0.94	0.86	0.34	0.26	1.06
02. Thesis statement relevance	0.86	0.87	-0.39	0.11	1.01	0.99	0.25	0.25	0.99
12. Main idea unity	0.80	0.88	-0.45	0.10	0.90	0.81	0.42	0.29	1.14
29. Word choice	0.61	0.89	-0.58	0.08	1.04	1.05	0.28	0.33	0.84
15. Main idea arrangement	0.84	0.91	-0.71	0.11	0.99	1.00	0.27	0.26	1.00
09. Essay length	0.89	0.91	-0.74	0.12	0.89	0.81	0.37	0.23	1.09
10. Supporting idea unity	0.84	0.91	-0.78	0.11	0.89	0.85	0.40	0.26	1.12
24. Pronoun	0.89	0.92	-0.89	0.12	1.03	1.11	0.19	0.23	0.97
33. Spelling	0.98	0.94	-1.21	0.25	1.02	1.45	0.05	0.11	0.97
01. Essay topic introduction	0.97	0.98	-2.26	0.23	1.00	1.38	0.10	0.13	0.99
25. Use of simple sentence	1.00	1.00	-4.16	1.00	1.00	0.59	0.04	0.03	1.01
Vocabulary D29-D30	0.49	0.49	1.58	0.06	1.06	1.11	0.34	0.41	0.77
Coherence D10-D12	0.70	0.75	0.46	0.05	0.92	0.87	0.52	0.46	1.14
Cohesion D13-D16	0.73	0.77	0.37	0.04	0.94	0.94	0.40	0.35	1.10
Grammar D20-D24	0.77	0.81	0.12	0.04	1.10	1.18	0.24	0.34	0.85
Content D17-D19	0.78	0.81	0.10	0.05	0.97	0.96	0.34	0.30	1.05
Organisation D01-D09	0.77	0.83	-0.07	0.03	0.96	0.96	0.43	0.41	1.06
Sentence D25-28	0.77	0.94	-1.21	0.06	1.05	1.06	0.60	0.61	0.94
Mechanics D31-D33	0.90	0.95	-1.35	0.08	1.04	1.17	0.34	0.36	0.97
Overall scale	0.75	0.78	0.00	0.13	1.00	1.02	0.27	0.27	1.00

All descriptors had the Infit MS indices within the range of 0.80-1.2, which is very close to the expected value of 1, suggesting that they fit the expected Rasch model well (Linacre, 2018). The acceptable fit indices confirmed that the scale was internally

consistent in capturing the unidimensional underlying construct and each descriptor did not affect the scores assigned to other descriptors (Barkaoui, 2014).

However, D27 (use of complex sentences) exhibited a negative PTM correlation far below the expected positive value of 0.20. Thus, D27 was misfitting to the Rasch model and might potentially measure a secondary irrelevant dimension apart from the focal construct (Linacre, 2018). Nine descriptors (27%) in total showed the PTM correlations below 0.20 but almost all of these, except D27, were positive. The discrimination (D) indices of all descriptors were within the acceptable range of 0.50 -1.50 and close to the expected value of 1, suggesting that all descriptors discriminated the student essays equally well (Linacre, 2018).

Based on the Rasch logits, the writing skills showing the logits values over 0.50, 1.00, and 2.00 may be considered as slightly weak, relatively weak, and very weak skills respectively, whereas the skills displaying the negative logits below -0.50, -1.00, and -2.00 may be deemed as slightly strong, relatively strong, and very strong skills respectively. The skills showing the logit values between 0.50 and -0.50 may be regarded as the skills the students were developing as it is not clear if these skills are weak or strong. Throughout the course, although this group of Thai EFL students generally demonstrated relatively strong skills related to mechanics and sentence, the students appeared to have problem using punctuation (D31) and particularly sentences (D25) accurately. This means that they might have knowledge of sentence types (simple, compound, and complex sentences) and used various sentences but made more sentence errors than acceptably expected by the raters. Amongst the mechanical skills, punctuation seemed to be the weakest skill, indicating that the students did not use punctuation appropriately over the course. The students' weak skills and those showing the logit values from 0.50 to -0.50 should receive particular attention for pedagogical purposes as these skills can inform future instructional and remedial actions.

4.4 ANOVA, Correlation, and Regression Results

ANOVA, correlation, and regression statistics were used to investigate the relationship between (1) the formative diagnostic assessment and achievement assessment scores, (2) the rater agreement, descriptor difficulty, and essay quality, (3) the

student self-assessment and achievement assessment, and (4) the student self-assessment and teacher-led assessment. The data were based on the five teachers' ratings.

4.4.1 Formative Diagnostic Assessment and Learning Achievement

ANOVA and correlation statistics were used to examine the correspondence between the three-round formative diagnostic assessment and the summative achievement assessment, thus contributing information relevant to RQ1 and RQ4.

As shown in Tables 4.13, ANOVA results showed that there were significant differences in the diagnostic score means, $F(2, 77) = 12.101$, $p < 0.000$, and the diagnostic logit means, $F(2, 77) = 13.185$, $p < 0.000$ between the high, mid, and low achieving groups. Due to the homogeneous variance of the student groups as suggested by the non-significant Levene test, the Tukey HSD post-hoc test was used and it indicates that the high-achieving group's mean score and logit were significantly higher than those of the mid- and low-achieving groups. The mid-achieving group's mean score and logit were significantly greater than those of the low-achieving group.

The ANOVA results suggested that the scale and the teachers could diagnostically differentiate the student writing ability in line with the student achievement levels. In other words, the student's writing ability levels on the formative diagnostic assessment were consistent with their achievement levels on the exam tasks.

Table 4. 13 *Comparison of Score and Logit Differences Between Student Ability Groups*

Data	Groups	Descriptives			Homogeneity		ANOVA		Post-hoc test	
		N	M	SD	Levene	*p*	F	*p*	Paired	*p*
Score	High	15	28.13	1.95	2.755	0.070	12.101	0.000**	H > M	0.037*
	Mid	33	26.03	2.29					H > L	0.000**
	Low	32	24.09	3.28					M > L	0.013*
Logit	High	15	2.11	0.59	0.165	0.848	13.185	0.000**	H > M	0.047*
	Mid	33	1.66	0.61					H > L	0.000**
	Low	32	1.17	0.60					M > L	0.005**

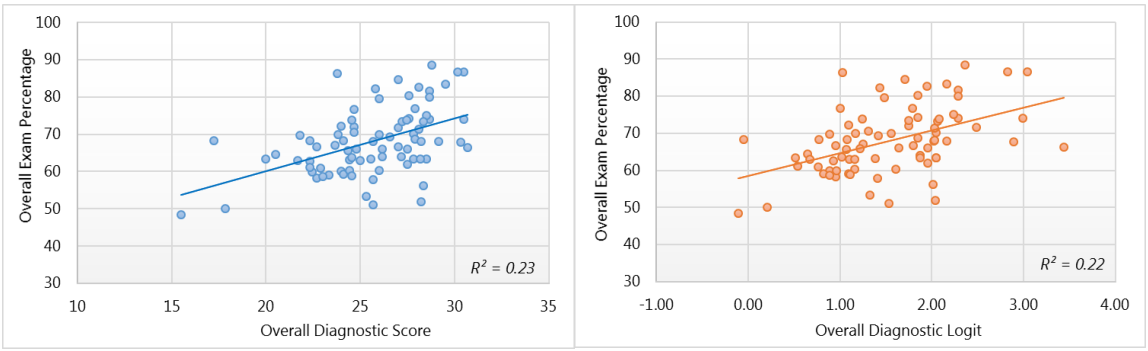
**p < 0.5, **p < 0.01*

Figures 4.2 displays the relationship between the formative diagnostic and summative achievement assessments. Correlation values of 0.10, 0.30, and 0.50 indicate low, medium, and high association respectively and coefficients of determination (R^2) of 0.01, 0.09, and 0.25 suggest small, medium, and large effect sizes of the predicting variable respectively (Cohen, 1988, 1992). Larson-Hall (2015) suggests that attention should also

be paid to understand the size of the correlation as the correlation tends be statistically insignificant if the sample size is quite small. As displayed in the figure, there were significantly positive and strong correlations between the diagnostic score and the exam percentage (N = 80, r = 0.48, p = 0.000), and between the diagnostic logits and the exam percentage (N = 80, r = 0.48, p = 0.000). This implies that the students diagnosed as having high writing ability tended to gain high learning achievement.

The regression was calculated to predict and explain the student learning achievement exam based on the formative assessment. Preliminary analyses were performed to ensure that there was no violation of the normality and linearity assumptions. The results revealed that the diagnostic score significantly predicted the achievement percentage, b = 0.48, $t(78)$ = 4.188, p = 0.000, and significantly explained 23% of the variance in the overall achievement percentage, R^2 = 0.23 ($F(1,78)$ = 23.484, p = 0.000. Furthermore, the diagnostic logit significantly predicted the achievement percentage, b = 0.48, $t(78)$ = 26.756, p = 0.000, and significantly accounted for 22% of the variance in the achievement percentage, R^2 = 0.22 ($F(1,78)$ = 22.283, p = 0.000.

Figure 4. 2 *Correlation Between Formative Diagnostic and Achievement Outcomes*



The correlation and regression results suggested that the current assessment, integrated into the ongoing teaching and learning, helped the teachers and students keep teaching and learning on track and move the students gradually towards achieving the learning goals.

4.4.2 Rater Agreement, Descriptor Difficulty, and Essay Quality

The correlation was used to explore the relationship of rater agreement with descriptor difficulty and essay quality in order to investigate rater consistency, thus

yielding information related to RQ2. As portrayed in Figure 4.3, there were significantly positive and strong relationships between the rater agreement percentages and the diagnostic scores (n = 80, r=0.72, p = 0.000) and between rater agreement percentages and the diagnostic logits (n = 80, r = 0.71, p = 0.000). Likewise, there were significantly positive and strong relationships between the rater agreement percentages and the CTT difficulty indices for the CA course (n = 33, r=0.85, p = 0.000) and the CB course (n = 33, r = 0.69, p = 0.000) as shown in Figure 4.4. Correlation results suggested that the teachers were more homogenous when judging easier descriptors and were less homogenous when judging harder descriptors.

Figure 4. 3 *Correlation Between Rater Agreements and Student Diagnostic Scores*

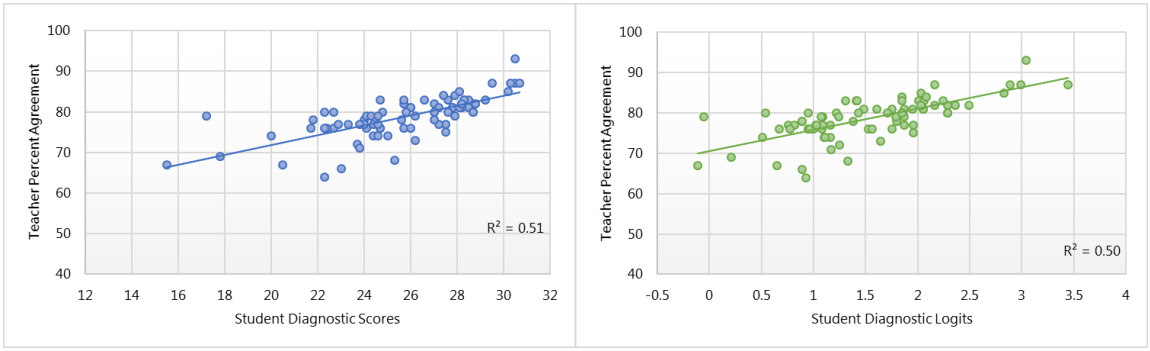
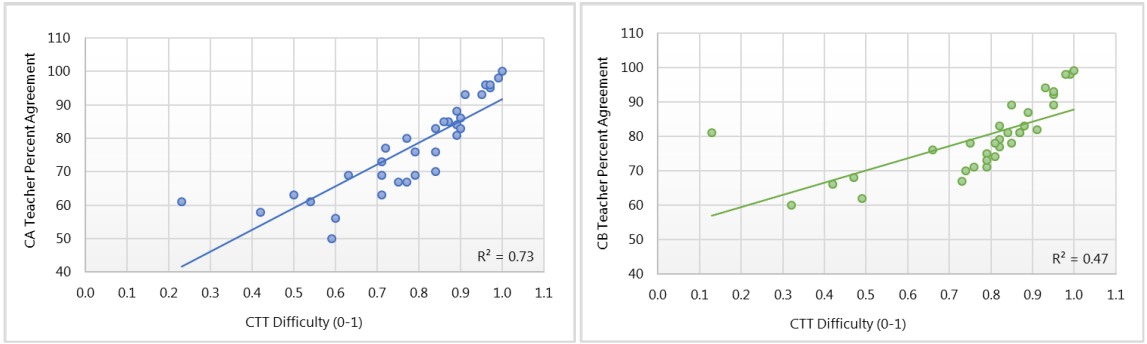


Figure 4. 4 *Correlation Between Rater Agreements and Descriptor Difficulty Indices*



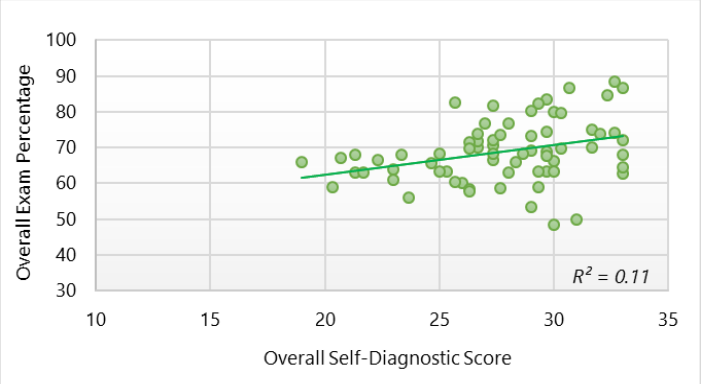
4.4.3 Student Self-Assessment Behaviour

The ANOVA, correlation and regression analyses of the self-assessment provide partial information on the formative impact of self-assessment on students' self-regulated learning and summative achievement, hence partly responding to RQ4. The results of the self-assessment behaviours were based on the scale scores obtained from 68 students' self-ratings and five teachers' ratings on these students. Twelve students' ratings were excluded as they did not score all descriptors on all tasks.

4.4.3.1 Student Self-Assessment and Learning Achievement

Figure 4.5 displays the correlation between the student self-assessment scores and the achievement exam scores. As can be seen, there was a significantly strong and positive correlation between the self-diagnostic score and the exam percentage (N = 68, r = 0.33, p = 0.006). The self-diagnostic score significantly predicted the achievement percentage, b = 0.33, t(66) = 5.683, p = 0.000., and significantly explained 11% of the variance in the overall examination percentage, R^2 = 0.11 (F(1,66) = 7.995, p = 0.006. The correlation and regression findings suggested that the students' formative self-assessment contributed to their learning achievement.

Figure 4. 5 *Correlation Between Self-Assessment and Achievement scores*



4.4.3.2 Student Self-Assessment and Teacher Assessment

The ANOVA and correlation were employed to examine the student's self-assessment accuracy and consistency vis-à-vis the teachers' ratings. As shown in Table 4.14, ANOVA results indicated that there were significant differences between the student group means on the overall task (F(1,134) = 15.120, p = 0.000), on Task 1 (F(1,134) = 24.957, p = 0.000) and on Task 3 (F(1,134) = 7.908, p = 0.006).

Regarding the high-ability students, there were significant differences between the group means on the overall task (F(1,26) = 4.382, p = 0.046), on Task 1 (F(1,26) = 4.393, p = 0.046) and Task 3 (F(1,26) = 5.219, p = 0.031). As for the mid-ability students, there were significant differences between the group means on the overall task (F(1,60) = 3.988, p = 0.050), on Task 1 (F(1,60) = 9.440, p = 0.003), and on Task 3 (F(1,60) = 4.237, p = 0.044). As regards the low-ability students, there were significant differences between the group means on the overall task (F(1,44) = 11.289, p = 0.002), on Task 1 (F(1,44) = 17.407, p =

0.000), and on Task 2 ($F(1,44) = 4.931$, $p = 0.032$). All this implies that the students tended to be lenient or overestimate their writing ability in comparison with their teachers.

Table 4. 14 *Rating Differences Between Students and Teachers*

Tasks	Raters	Descriptives			Homogeneity		ANOVA	
		N	M	SD	Levene	*p*	F	*p*
All	Student	68	27.75	3.47	0.663	0.417	15.120	0.000**
	Teacher	68	25.53	3.17				
1	Student	68	28.75	3.77	1.493	0.224	24.957	0.000**
	Teacher	68	25.16	4.57				
2	Student	68	27.16	4.37	2.106	0.149	1.809	0.181
	Teacher	68	26.22	3.80				
3	Student	68	27.34	4.74	0.773	0.381	7.908	0.006**
	Teacher	68	25.23	3.99				
All	High	14	29.76	2.23	0.345	0.562	4.382	0.046*
	Teacher	14	28.08	2.02				
1	High	14	30.71	2.49	0.043	0.837	4.393	0.046*
	Teacher	14	28.80	2.34				
2	High	14	29.07	2.95	3.047	0.093	0.166	0.687
	Teacher	14	28.69	1.88				
3	High	14	29.50	2.93	0.088	0.769	5.219	0.031*
	Teacher	14	26.75	3.42				
All	Mid	31	27.51	3.56	2.724	0.104	3.988	0.050*
	Teacher	31	25.99	2.30				
1	Mid	31	28.48	3.43	0.125	0.725	9.440	0.003*
	Teacher	31	25.88	3.23				
2	Mid	31	26.84	4.78	5.828	0.019	0.047	0.829
	Teacher	31	27.05	2.79				
3	Mid	31	27.19	4.56	0.369	0.546	4.237	0.044*
	Teacher	31	25.02	3.71				
All	Low	23	26.86	3.62	0.412	0.524	11.289	0.002**
	Teacher	23	23.37	3.41				
1	Low	23	27.91	4.51	0.074	0.786	17.407	0.000**
	Teacher	23	21.96	5.14				
2	Low	23	26.43	4.34	0.103	0.750	4.931	0.032*
	Teacher	23	23.59	4.36				
3	Low	23	26.22	5.54	1.148	0.290	1.208	0.278
	Teacher	23	24.57	4.56				

$*p < 0.05$, $**p < 0.01$

Table 4.15 shows ANOVA results comparing the students' self-rating leniency between ability groups. The differences between the student-ratings and teacher-ratings represent the size of the self-rating leniency and thus were used to compare if there were significant differences in the leniency between the student groups. Overall, the results showed that in spite of no significant differences in the leniency between the student ability groups, the low-achieving students generally showed higher degree of leniency.

Table 4. 15 *Self-Rating Leniency Differences Between Student Ability Groups*

Task	Group	Descriptive			Homogeneity		ANOVA	
		N	M	SD	Levene	p	Welch	*p*
All	High	14	1.68	2.00	4.770	0.012	1.160	0.324
	Mid	31	1.52	4.05				
	Low	23	3.48	5.46				
1	High	14	1.92	2.96	4.142	0.020	3.021	0.061
	Mid	31	2.60	4.06				
	Low	23	5.95	6.86				
2	High	14	0.38	2.00	8.075	0.001	1.823	0.175
	Mid	31	-0.22	3.93				
	Low	23	2.85	6.84				
3	High	14	2.75	4.46	1.240	0.296	0.151	0.861
	Mid	31	2.17	6.90				
	Low	23	1.64	8.03				

Figures 4.6 portrays the correlations between the students' self-ratings and the teachers' ratings across the tasks and groups. Overall, the students' self-ratings were consistent to a small extent with the teachers' ratings.

Figure 4. 6 *Correlation Between Students' Self-Ratings and Teachers' Ratings*

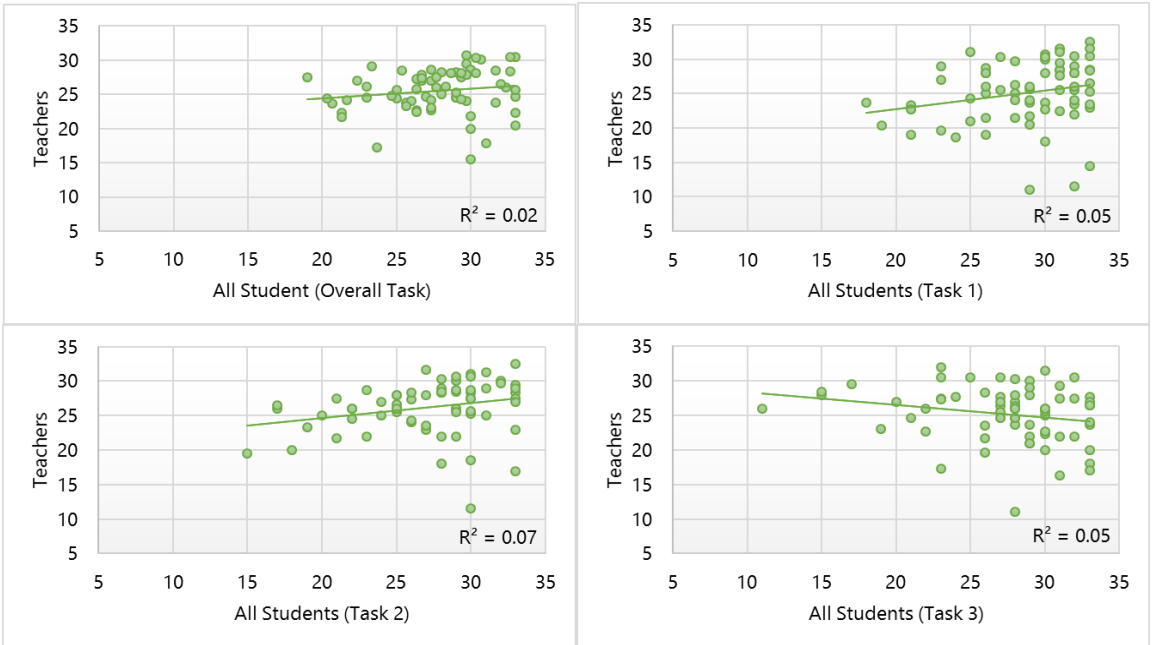
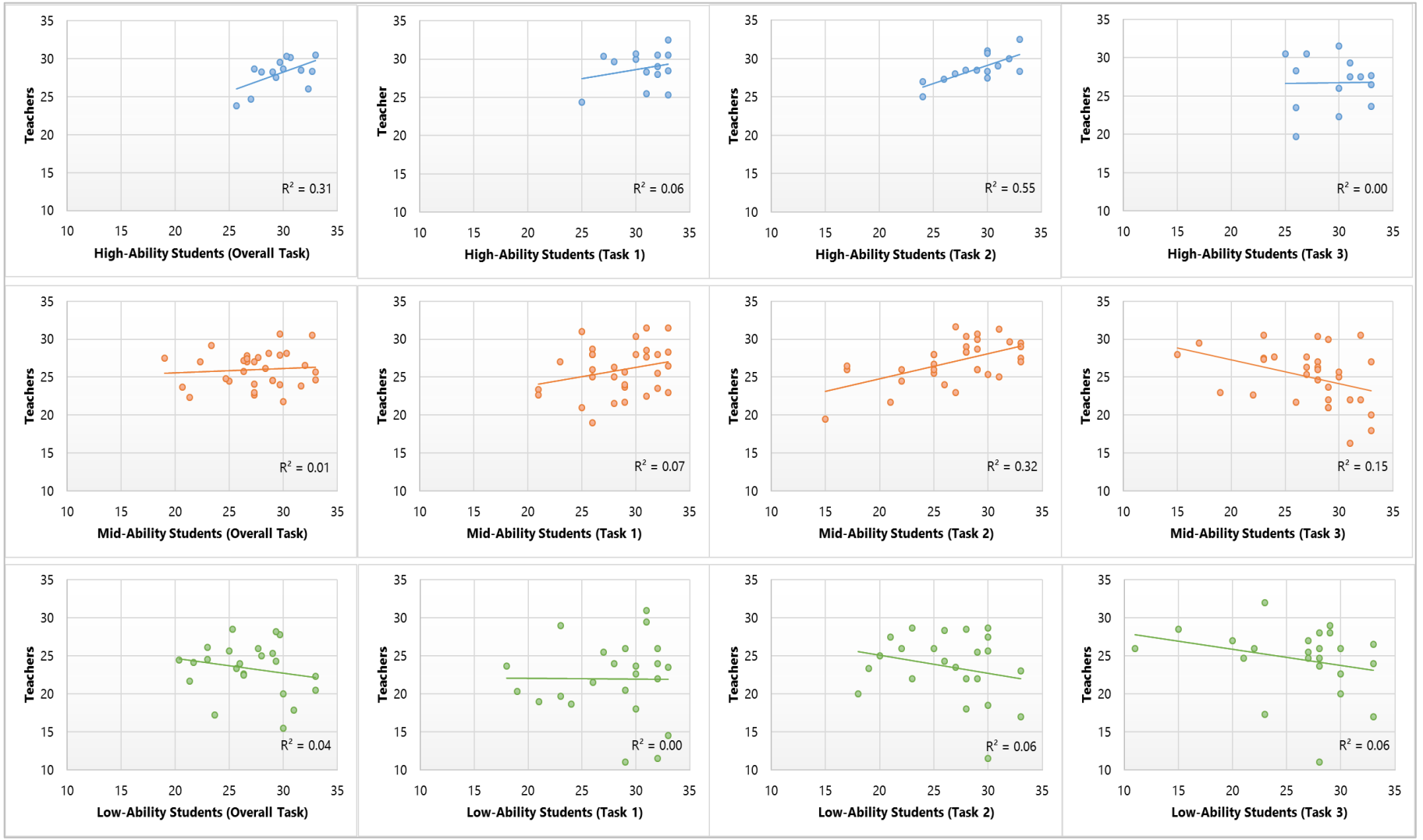


Figure 4.7 displays the correlations between students' self-ratings and teachers' ratings across student ability groups and tasks. Generally, the higher-achieving students' ratings were more accurate and consistent with the teachers' ratings than the lower-achieving students.

Figure 4. 7 *Correlation Between Students' Self-Ratings and Teachers' Ratings Over Tasks*



In addition, it was observed that the high-and mid-ability students' self-rating consistency over the tasks were similar to the levels of task performance logits (see also Figure 4.1). That is, they were most consistent when judging Task 2 showing the highest logit and were least consistent when rating Task 3 showing the lowest logit, which might imply that higher-ability students' self-rating consistency varies according to the levels of task difficulty. This is not true for the low-ability students showing decrease in self-rating consistency over the tasks, which might imply that whether the task is easy or difficult, they were still not able to self-judge their work.

4.5 Chapter Summary

This chapter has presented the quantitative results based on the descriptive, CTT, MFRM, ANOVA, correlation, and regression analyses. All in all, the quantitative results revealed that the scale functioned appropriately and consistently in the classroom assessment. The teachers also behaved appropriately, exhibiting no inconsistent and dependent scoring patterns. While the raters differed in the rating severity, each rater was self-consistent in interpreting the descriptors and diagnosing the student essays. It could also be implied from the findings that the formative diagnostic assessment supported the students' learning achievement. In addition, the students' self-assessment generally showed a certain degree of rating consistency with the teacher ratings with the higher-achieving students showing more consistent ratings than the lower-achieving students. Moreover, a certain degree of student self-assessment consistency suggests that the students were attentive to the self-assessment process, thus implying that they were developing self-regulated learning skills. In the next chapter, the qualitative results will be presented with regard to the teacher and student perceptions of the scale functioning, usefulness, and impact.

Chapter 5: Qualitative Results

This chapter presents the results from the qualitative content analyses of the five teachers' and 20 students' perceptions from the semi-structured interviews. The qualitative findings are concerned with the teacher and student perceptions of the scale functioning, usefulness, and impact as well as the students' self-assessment practices. The qualitative findings will later be integrated with the quantitative findings in response to RQ1, RQ3, and RQ4 and in justification of the validity argument in the discussion chapter. In this chapter, the teacher perceptions are presented first, followed by the student self-assessment strategies and perceptions.

5.1 Teachers' Perceptions of the Scale

Table 5.1 summarises the teacher perceptions, which were categorised in to three main themes: scale functioning, scale usefulness, and scale impact. Each of the main themes was further subcategorised into two-level hierarchical subthemes. Each subtheme is presented along with illustrative examples of the teachers' comments, including the researcher's prompts or questions in square brackets.

Table 5. 1 *Summary of Teachers' Perceptions of the Scale*

Main themes	Sub-themes	Perceptions
Functioning		
▪ Comprehensibility	• Criteria clarity	- Most descriptors are largely clear and understandable.
	• Criteria judgement	- Some descriptors are difficult to judge.
▪ Comprehensiveness	• Criteria specificity	- Criteria capture discrete and specific writing skills. - Error counting does not capture the quality of writing skills. - Binary rating does not capture the granularity of writing skills and should have more than two options.
	• Criteria coverability	- Criteria largely cover core writing skills and learning contents. - Other skills should be added to the criteria.
▪ Applicability	• Scale organisation	- The scale layout is largely arranged in a way that is easy to use. - The scale length should be a single page. - Micro-skill descriptors should come before macro-skill descriptors.

Main themes	Sub-themes	Perceptions
	• Rating format	- Binary rating is largely practical and easy to judge. - More rating options are easier to judge. - The 1-point option should come before 0-point option.
Usefulness		
▪ Teaching	• Diagnostic information	- The scale provides information about students' writing strengths and weaknesses.
	• Diagnostic feedback	- The scale provides information for detailed and digestible feedback.
	• Student improvement	- The scale provides information about students' writing improvement.
	• Diagnostic report	- The scale should include a concise report of individual students' diagnostic profiles.
	• Summative assessment	- The scale helps to better assess students' summative exam essays.
	• Teaching guideline	- The scale is useful as teaching resources and guidelines.
	• Scale practicality	- The scale has a lot of descriptors and thus is time-consuming to use in ongoing assessment.
▪ Learning	• Self-assessment and self-regulation	- The scale is largely useful for self-assessment and self-learning. - The scale is not effective for low-achieving students' self-assessment and self-learning. - Higher-ability peers should assist low-achieving students' self-assessment.
	• Writing development	- The scale-assisted self-assessment is largely useful for writing development - The scale is not useful for idea improvement.
Impact		
▪ Awareness raising	• Assessment fairness	- Assessment should be fair to students.
	• Self-assessment	- Self-assessment is necessary for students' writing development.
	• Feedback	- Feedback is important for students' writing development.
▪ Future plan	• Scale adaptation	- Teachers want to adopt and adapt the scale for future teaching and assessment.
	• Professional development	- Teachers have new ideas to improve future teaching and research. - Teachers are interested in doing research based on the diagnostic outcomes.

5.1.1 Functioning of the Scale

The teacher perceptions of the scale functioning were structured into three themes: scale comprehensibility, comprehensiveness, and applicability. Each of these themes was divided into associated subthemes, encompassing the variability in teachers'

viewpoints. The reader is reminded that only one teacher (Ivey) was interviewed in Thai and her interview protocol was translated into English.

5.1.1.1 Comprehensibility

The teacher perceptions of the scale comprehensibility were classified into two subthemes: criteria clarity and judgement. In general, all teachers perceived that the descriptors were clear and understandable as illustrated by the example quotes below:

Ken: *[Are the scale descriptors easy to understand in your opinion?] It is easy to understand yes.*

Ivey: *[Are the scale descriptors easy to understand? If not, specify descriptors that were ambiguous or not clear?] Yeah, it is easy to understand.*

However, individual teachers were uncertain how to judge some descriptors as these descriptors were not clear to interpret. To elaborate, Sara, Nana, and Cali were not certain how to rate D17 (*content comprehension*), and Nana and Ken were not certain how to judge D29 (*word choice*) and D30 (*word variety*). Furthermore, Nana was not confident in judging D03 (*topic sentence and thesis relevance*), D11 (*supporting idea convincing*) and D12 (*main idea and thesis relevance*). Sara was also not certain how to rate D28 (*sentence problem/accuracy*). Some quotes from the teachers' commentary are presented below:

Sara: *[Any descriptors you think are ambiguous or not clear or you find it difficult to judge?] Number 17 (content is understandable enough). [Ok why is that?] Em, as a Thai teacher, yes I understand what they try to tell the audience. If a foreigner, native-English speaker, have to tick ZERO or ONE on Number 17, it might be difficult for them [So, when you rated the essays, you kind of think of native-speaker readers right?] Yes.*

Ken: *What about these two as well I think Number 29 and 30 (words are used appropriately for the context) and also (various words). Mostly, I will give like ONE for words for Number 29 and ZERO for various words for these two descriptors. In my opinion [words are used appropriately for context] I think it should be something else instead of these two but I don't know what they are.*

5.1.1.2 Comprehensiveness

The teacher perceptions of the scale comprehensiveness were related to criteria specificity and criteria coverage. All teachers held the view that the scale criteria captured detailed, discrete, and specific writing skills as demonstrated by the selected quotes below:

Cali: *[Is the scale appropriate for identifying students' writing strengths and weaknesses in an ongoing classroom instruction?] Em yes, I think it's appropriate. I think all the points are*

like very very in detail and we can measure about the knowledge and skills of the students via their writing.

Nana: *[Do you think the descriptors are specific enough to capture I mean detailed skills or several writing skills?] I think yes.*

Although the criteria largely captured discrete and specific writing skills, some of the teachers voiced concerns over their specificity. Ken and Cali found that the frequency of errors did not necessarily indicate the overall writing quality of an essay. Example quotes from the teachers' comments are offered below:

Ken: *So, sometimes for example if one student wrote just only simple sentences throughout from the very beginning to the last sentence, they do not have any errors at all something like this.*

Cali: *If you ask me about any recommendation or suggestion, it is that sometimes when the students write an essay and em even though they lack of the skills in all of these but it's just like little little mistake about punctuation, little mistake about capitalisation, spelling, and everything. It means like they still have got ONE for all of these. But if you see their writing, it's still not very good. [On the whole, right?] Yes, on the whole thing. [On the macro level] Yes.*

Apart from the error counting problem, Ivey and Cali were concerned that the binary rating was crude and did not capture the detailed granularity of writing quality. They thus suggested that more rating options should be added in the rating format. Examples of their comments are quoted below:

Ivey: *[Is the judgement or scoring of the scale descriptors appropriate? If not, explain why?] Yeah but I think it's a little bit crude. [What would you want it to be?] Maybe we can make it like more scale or degree. [What about two option, what do you think it's it difficult to rate?] As I said, it's not detailed. Sometimes, students' writing meets the descriptor requirement but it has only one feature that appear in the descriptor. [What about in an ongoing classroom do you think you can finish rating in time using several rating options?] Yeah, I think we can finish rating in time.*

Cali: *So, if you would like it to be in very detail, maybe you would have some kind of like the rating ONE, TWO, THREE, like this. [So, you prefer more rating options?] Yeah. [More than two levels?] Yeah because this one is "Right" or "Wrong." You should have some kind like right wrong or something in between. But it depends on like you know what level you'd like to measure the students anyway.*

With regard to criteria coverage, all teachers perceived that the scale criteria covered essential writing skills and teaching contents in the classrooms as explicated in the example quotes below:

Sara: *[What about the number of the descriptors do you think it's too many descriptors or it's kind of ok for you for detailed information?] It's ok ok. These are the things we're looking in their essays. It's em common. [You mean it kind of contains necessary skills that students have to write in the essays] Yeah to be able to write.*

Ivey: *[Do you think that the rating scale cover all productive skills and textual features of the expository writing in classroom? If no, specify descriptors that were irrelevant to expository writing] I think it all covers essential skills according to the course objectives, for example five paragraph essay, what is a paragraph.*

In addition, individual teachers proposed that other writing features were useful and should be included for writing assessment. These features included: *(a) academic language* by Cali, *(b) consistent use of English style* by Sara, *(c) genre-specific features* by Nana, *(d) additional grammar features* by Ivey, *(e) overall impression* by Ken, and *(f) standard or natural English* by Sara and Cali. Example quotes are given below:

Ivey: *…but it is not detailed enough for grammatical features. [How do you want the scale to be specific?] The scale focuses more on content, the skills of expository writing structure. The scale should include more specific details about grammar points. There is a small number of grammar descriptors.*

Ken: *[If you want to improve the scale to make it more like practical] and applicable to my classroom [and focus on like more specific skills, what would you want to improve if you want to improve?] I think these criteria itself overall I mean it's already good in my opinion, but there is something missing I mean if you can add more something like content overall impression you can add more. Sometimes, you know when we take a look at all the sentences or a piece of writing by students there is no error at all, but they lack content. I mean the writing are not appropriate. Yes, if we weigh between this and also overall impression of the content, so sometimes we take a look if they don't use them.*

5.1.1.3 Applicability

The teacher perceptions of the scale applicability were pertinent to the scale organisation and rating format. Overall, all teachers perceived that the scale layout was well organised and easy to follow as illustrated by the following comments:

Ivey: *[How it is easy to use in your opinion?] The scale shows the picture of the writing structure. [What do you mean by the picture?] For example, the scale helps to know the structure of a paragraph and show how the paragraph is composed of and what elements should be included in the paragraph.*

Sara: *[Do you think the scale is user-friendly for classroom assessment? I mean is it easy or complicated or practical in your opinion?] Yes, for me.*

For a more user-friendly application of the scale, Nana and Cali preferred that the scale length should be a single page, making it easier for them to use and rate as their quotes illustrate:

Nana: *It would be better if you put it like only one page, maybe 20 descriptors. [You want to make all the descriptors I mean appear in one page?] Yeah, that would be very easy for teachers and students coz we don't have to like flip.*

Cali *For me, I think it's better if everything will be on the same page because sometimes you need to [So that it's easier for you because you don't have to flip] yeah.*

To make it easier to rate, Cali preferred that grammar and micro-skill descriptors come before organisation and macro-skill descriptors as she typically checked grammar skills and errors first as her comment illustrates:

Cali: *I think the organisation should be the latest one for me. [You mean put the grammar first] Yeah because you know when I do it normally I will check this and I'll check this and I go back to this yes. So, I think like why it's supposed to be like this every time [But you don't have to follow the order when you rate the essays] Yeah but it's ok you know. It doesn't matter a lot maybe other instructors may have some other techniques to do it as well.*

In addition to scale organisation, all teachers perceived that the binary rating is practical and easy to judge as explicated by some of the quotes below:

Nana: *[Do you think the scale is user-friendly for classroom assessment?] Yes, the reason is the same reason I provided [how?] the students or teachers only need to judge right only ZERO or ONE. So, it's easy for them for me as well.*

Ken: *[Is the judgement or scoring of the scale descriptors appropriate? I mean only two options ZERO and ONE or strong and weak] Yes, it is I think you have or we have more it's gonna be complicated. Two is enough.*

However, Ivey and Cali perceived that multiple rating options would not put more burden in their scoring decision and would be easier to judge as illustrated by their comments:

Ivey: *[So, when we actually rate, which one do you think is easier or more difficult, between two options and several options, which one requite more time] Em, I think several options may be ok and maybe easier to judge whether the skill exist and then how much it exist. [You mean if there more several options you think it's easy to judge] Yeah. [All descriptors?] Yeah, we can add more options for all descriptors.*

Cali: *[So, if the scale has more than two rating options do you think it's practical in classroom do you think you will have to spend more time?] No not at all because even though when you check right, you need just like count the mistakes of the student already and they have like one or two mistakes and you would like to say like oh your skill in this point is not very good but you still need to tick ONE. It means like good but it's not like not good yet.*

Despite the practicality of the binary rating, Nana expressed that view that the ordering of the 0-1 rating options was not familiar with her cognition as she was more familiar with the yes-no format of the checklist. She thus suggested that the 1-point

option should precede the 0-point option to make it consistent with her cognition. Nana's view is illustrated below:

Nana: *[Were there any, for example, errors or mistakes that happened because of the, you know, any problems that happened because of the scale format or the scale characteristics?] Ah at first, I often tick ONE instead of ZERO when I want to give ZERO. So, I tick in the Number ONE instead of ZERO because we normally think that ZERO. [Some students mentioned they are familiar with "Yes" first and then "No". So, "Yes" is supposed to be ONE] Yeah. [So, ONE should come first, right?] Right right, that's the point. [In the first option] But later after I've been using this I'm getting used to it.*

5.1.2 Usefulness of the Scale

The teacher perceptions of the scale usefulness were categorised into two main themes: teaching and learning each with a number of associated subthemes.

5.1.2.1 Teaching

The teacher perceptions of the scale usefulness for ongoing teaching were sub-categorised into several aspects: diagnostic information, diagnostic feedback, student improvement, diagnostic score report, summative assessment, teaching guideline, and scale practicality. To begin with, all teachers concurred that the scale provided diagnostic information which helped them to identify students' writing strengths and weaknesses as demonstrated by the following sample quotes:

Nana: *[Did the scale provide useful information for you to improve how you teach?] Yes, coz I when I rated the students then I found out that, for example, they lose their scores in part of organisation. So, most of them got ZERO here, so I know that I should tell them to be more carefully on this part. So, it would be useful you can track your students' problems right and then you can improve your students. Normally I do that.*

Ivey: *[Anything else?] Students often use spelling and punctuation incorrectly, but I understand that and when I gave feedback, I talked about this but not that much. Yeah, it's interesting. If we don't have the scale, we will not see these mistakes.*

In addition, Sara, Nana, and Cali mentioned that the scale helped them to provide specific and targeted feedback to students as illustrated by their comments:

Nana: *When I give students' feedback, I always show I mean I show this with the written comments in the students' essays together like this. So, I show them and let them see what point they loose and what point they gain so that the students can see clearly which parts or which skills they should put more focus on or improve more on that. So, I think it's useful when I use this with students' writing and when I gave students' feedback, so I can like group and clearly show students the quality of their writing ability, writing skill in each essay.*

Sara: *[Do you think that the diagnostic rating scale provides useful information for improving the way you assess students' expository writing? If so, why and how?] Sure sure, I think I*

have already mentioned two points that are very essential from these descriptors. I think it's because of this detailed information about sentence and mechanics. So, after I rated all students' writing, I usually maybe I didn't do it individually but I give the overall feedback to my students for this kind of like useful information about how to write more compound sentences because they use fewer compound sentences from what I graded their writing.

As well as diagnostic information and feedback, Ivey and Cali reported that the scale gave them insight into know students' improvement as illustrated by their quotes below:

Ivey: *When they write the first and second drafts, they did not do well on the first draft but when I rated the second draft I can see that they improved though it is not as good as I expected. This shows that students listened to feedback and then go back to revise their writing. So many students did better on the second draft better than the first draft though there are some mistakes.*

Cali: *For example, the best student in my class, you already met her. She is very outstanding in class. Whenever she speaks or writes, she is more advanced than others. Without this scale, she thinks she is already good and have nothing to improve, but when she do the self-assessment, she know her weaknesses. Sometimes, she try to make simple sentences to become compound or complex sentences but she write run-on complicated sentences. When she know this, she try to improve for the next essays and make shorter and simple sentences and use various words as well.*

To enhance the interpretation of the diagnostic outcomes, Nana suggested that there should be a brief and more digestible report of students' diagnostic profiles:

Nana: *[Anything else you want to say in terms of the scale help you to improve your teaching?] Yeah, I mentioned it already. So, I looked at the scale I gave to the students each time and I am not a kind of statistic person but if I can do it I can like put it on the programme so that we know what are the weakness of the students. [You mean there should be something like a report] Yeah so that we know right so we can analyse right the students' strengths weakness whatever so it would be very useful for. [Yeah I'll will produce that report I mean diagnostic profile reports] Yeah, ah profile report right it will be useful for every teacher.*

As a result of applying the diagnostic criteria, all teachers reported that the scale helped them to better understand and more carefully interpret the assessment criteria used to evaluate the students' essays in the summative midterm and final examinations as illustrated in the following comments:

Sara: *[Even though you used another scale for the midterm and final exams, do you think at some point the criteria in the diagnostic scale helped you to kind of better judge midterm exams even though you used another rating scale] Yeah it reminded me actually most of the items on your scale yeah, we include them in another criteria we use yes. [But the descriptions] is different a bit different.*

Ken: *[Ok what about assessment?] Assessment. [Any changes in classroom assessment do the scale affect your decision on the midterm and final exam criteria?] Yes yes sure sure of course sure. Em, I don't normally take a look at everything in details. I usually gave ah overall impression about their writing rate all and then grade it but now. [You then to holistic-evaluate student performance holistically] Yeah yeah not this. [And how the scale has any influence on your assessment] Just like you said, I am more careful about diagnostic. [You focus more on specific skills] Yes, I try to balance between the two holistically and specific details.*

Furthermore, Sara, Ivey, and Cali reported that they used the scale as the teaching resource and guideline:

Cali: *You know when Teacher Ken and I received this scale, we teach exactly like the scale because we want students to learn exactly like the scale [So, the scale is part of your teaching materials] Yes, it's part of our teaching materials. We talked to each other if we want students to get ONE for all of these, what kind of things that we should teach them. We train them on each point according to this scale.*

Sara: *[You mean the scale served as the guideline for you right as a framework for your teaching and for assessment as well] Yes, it's practical. It's very easy to use yeah for my teaching and for the student as well.*

In respect of scale practicality, Nana was concerned that the scale contained a lot of descriptors, which made it time-consuming to continually rate multiple essays in an ongoing classroom:

Nana: *But I found out that maybe it's too many descriptors here but another but again because I do understand that you have to cover every writing skill right you provide a lot of descriptors and information here but in terms of I mean it's good but in terms of practical [The number of descriptors is too many?] Yes, too many yes.*

5.1.2.2 Learning

The teacher perceptions of the scale usefulness for ongoing learning are associated with two subthemes: self-assessment and self-regulated learning, and writing development. In respect to self-assessment and self-regulated learning, all teachers perceived that the scale was useful for promoting student self-assessment and self-regulated learning because the students used the scale to guide and revise their essay writing. Examples of the teacher comments are offered below:

Nana: *[How it is important, self-assessment, in writing classroom and assessment?] Ok this is from my student view, they said they look at this before they handed in the essays to me so the self-assessment helped them to prepare their task. So, they know that the good essay should include these domain or descriptors right so they kind of have an outline already before they write so they can follow by trying to cover all these descriptors. [So, should it be included?] Yeah for sure.*

Ken: *Em, I think descriptors in my opinion because students have this with them and then they have to follow all the descriptors. So, this might have something in their mind yes, for example, the introduction paragraph, they should have a clear topic they should have a clear thesis statement for their writing or for their introduction.*

However, Sara and Ivey held the view that the scale may not be very useful and effective for the low-achieving students' self-learning and self-assessment because they were not able to identify whether their skills are weak or strong or pay attention to the teacher feedback as explicated in the quotes below:

Sara: *Some students they don't even know what the mistakes in their essays. So, it would be difficult for some of them right. They don't even know they created like incorrect sentences [You mean they don't have knowledge in terms of writing skills in their head] Yeah. [And it's difficult for them to judge whether their writings] is good or not [are good or fit into the descriptors or not] Yeah.*

Ivey: *For some students they show little improvement as they may listen to my feedback but were not attentive to my feedback. [Which group is larger between attentive and non-attentive students] In this course, there were four groups or classrooms as I talked with other teachers, we feel that most of the students in my class are relatively poor while most of better students were enrolled in other groups.*

To help the low-achieving students more effectively use the scale, Sara suggested that low-achieving students need further support from higher-ability peers and teachers as her comment illustrates:

Sara: *Here is my idea maybe if possible but it's unlikely to happen. They just need someone who is better in English than them work together with them. [Their higher-proficiency peers] Yeah while they are doing the self-assessment once or twice so that they would get some basic idea on how to grade themselves.*

Apart from student self-assessment and self-learning, all teachers perceived that the scale-assisted self-assessment was useful for writing development as the following example quotes indicate:

Cali: *[Do you think that the self-diagnostic assessment helped and or hindered the students' writing improvement? If so, why and how?] Sure, I think it does not hinder it helped the students a lot because you know when they know their goal, it's easy for them to reach the goal.*

Ken: *[Do you think that the self-diagnostic assessment helped and or hindered the students' writing improvement? If so, why and how?] I think it's both both. [Go with improvement first] Improvement, they are aware of what should be in their writing such as organisation coherence cohesion.*

Nevertheless, Ken went on to say that the scale was not useful for promoting the students' idea development:

Ken: *…but sometimes this will affect their ideas or content because they try to focus more on something like this to have good introduction how to have good coherence how to write good grammar so sometimes these will hinder their flow of ideas.*

5.1.3 Impact of the Scale

The teacher perceptions related to scale impact were associated with two main themes: awareness raising and interest. Each of the main themes was made up of its associated subthemes, each of which included the teachers' perceptions indicative of the subthemes. The teacher perceptions of the scale impact were as follows.

5.1.3.1 Awareness Raising

The teachers were aware of the importance of fair assessment, self-assessment, and feedback in learning and teaching following the implementation of the scale. In terms of the teacher awareness, Sara and Ivey realised that assessment should be fair and transparent to students as their quotes illustrate:

Sara: *[It can be in your future course how it will benefit your writing teaching, self-assessment in class you said that it should be included so it's useful right] It can be used as an evidence when my students have some questions on how I graded them. So, I can use this criteria and point out to them how their scores come. So, it's like an evidence to defend myself [To justify yourself, right?] Yeah to justify myself when they have questions like why you gave me B why not A. So, I can use it as a proof yeah. I think it if we have specific criteria, we can justify give a justification so that the students would how can I say. I don't want to leave it unexplained you know, so it can be used as evidence when they have questions.*

Ivey: *Ah I think it is good guideline to develop a scale, but it should be adjusted according to the teaching course and subject so that it is useful for teachers and learners and the teacher team. So, when we use the same scale, there will be no question about bias judgement or score assignment. [Anything else?] no.*

In addition to fair assessment, all teachers were aware that self-assessment should be included in writing classrooms as it is valuable for students' writing development as revealed by the example comment below:

Cali: *[Should self-assessment of writing be used in classroom teaching and assessment?] Actually, I think it' useful and I don't think Thai students get used to this kind of technique or even they finish their paper or exam they just sent it straightaway. [Most of the students who participated in my interview never used this kind of self-assessment before] I think Thai students don't get used to this kind of technique even me as well because Thai style of writing is that when you finish it means you finish.*

Ken: *[Should self-assessment of writing be used in classroom teaching and assessment?] You mean [The way students rate their own essays] Yes yes. [Why can you explain a little bit more about this] If we want to do something like try to ah ah ride a motorbike and then you don't know how but if there is a manual for you to follow. I think it's the same thing*

with that I think if they want to have good writing, they should have this criteria they should have these descriptors as a guideline for them to follow.

Additionally, Nana realised, as a result of participating in the research project, that both teachers' feedback and students' self-feedback are important for students' writing development as the following quote illustrates:

Nana: *I think that feedback is very important after I've been through this project, I think feedback is very important, both teacher feedback and student feedback and I look forward to seeing the report coz the result would be juicy right. [You mean the diagnostic outcomes score report] Right yeah, I really want to see coz that would be very useful for my future class [teaching preparation] right yeah right.*

5.1.3.2 Future Planning

Teachers' positive reaction to the scale was evident in comments about its future use in their classroom context. Four teachers mentioned that they were interested in adapting the scale for future use as illustrated by the following example comments:

Sara: *My writing course I think in the future, if I have a chance to teach expository or argumentative writing again, I would use the scale and give it to my students and explain some major points they need to acquire.*

Ken: *I think it is applicable in my opinion because something you use in this research can be applicable to ah my classroom and other people's classrooms as well. This criteria is useful and if possible we can adapt some of them and us it classroom.*

Some teachers saw the scale as tool for improving teaching and for conducting research. As a result of implementing the scale, Nana came up with new ideas to improve her future writing teaching as this quote illustrates:

Nana: *Even this class, I know that the student they lack the models, the examples, coz I learned that from using this rating scale. So, I know that the students they lack they know how to write individually they know how to write but they don't know how to em student were focused predominantly on the element of the essay but lack good essay examples or they did not much analyse the whole input essay or analyse the elements in good essay models [So do you think it's important for students to analyse good essays models or learn from the characteristics of good essays] Yes coz normally we always teaches separate elements of essays like each part of the essay [you mean you focus on teaching the skills necessary to write] right receptive skills.*

Inspired by the diagnostic assessment outcomes, Nana was interested in doing research related to student writing as indicated in the following comment:

Nana: *Em, can I say something like it strikes me up a little bit I want to do another research maybe in the future I want to compare student writing nativeness coz after using this rating scale I found out that some students they completely write a perfect sentence no grammatical mistake at all but the sense of language you understand what I mean the*

> *sense of language fell Thai you remember the first time that we discussed a bout Thainess that influences students' writing. So, I really want to know if this information or this scale [You mean you learned about what you want to do in the future because this project sparks you something that you want to further study] Like I mentioned coz some students write perfect sentences, but it doesn't feel native-like at all. Sometimes, I am still confused about that they put part of speech together very perfectly, but it doesn't make sense you know.*

5.2 Students' Self-Assessment Practices and Perceptions

The student interview findings, obtained from 20 volunteer students, are concerned with their self-assessment practices and scale perceptions. The self-assessment practices are concerned with how students generally applied the scale for self-assessment. The student perception themes are similar to but not as detailed as the teacher perception themes. The student self-assessment practices are presented first, followed by the student perception themes, each of which is illustrated by excerpted quotes with the researcher's prompts or interjections in square brackets.

5.2.1 Students' Self-Assessment Practices

Table 5.2 gives an overview of students' self-assessment practices. The word "*yes*" indicates that the students demonstrated the behaviour whereas "*no*" indicates that students did not. If the researcher did not ask the students about a particular behaviour, the letter "*n/a*" is used. Overall, most students reported that they had never done self-assessment before and that they spent about 5 - 30 minutes for each self-rating session. In addition to rereading and trying to understand essay topics, all students reported that they better understood descriptors after participating in the rating training session, and practically all of them understood and reread the task instructions.

In addition, most of the students read descriptors before rating their essays and reread some descriptors while self-rating their essays in order to make sure they clearly understood the descriptors. Most students also thought of the descriptors while rating their essay whereas most of them did not think of the model essays used as a guideline in the rating training. Virtually all students tended to change their rating decisions on some descriptors to make sure they made the right decision and all students compared their self-rating results with teachers' rating results.

Table 5. 2 *Students' Self-Assessment Practices*

Student background						Self-assessment strategies											
Student ID	Achievement level	Diagnostic score Task 1	Diagnostic score Task 2	Diagnostic score Task 3	Self-assessment experience	*Approximate time (min) spent on essay rating*	*Better understand descriptors after rating training*	*Understand assignment task instruction*	*Reread assignment task instruction*	*Try to understand essay topics*	*Reread essay topics*	*Read descriptors before rating an essay*	*Reread descriptors while rating an essay*	*Think of descriptors while rating an essay*	*Think of model essays in rating training when rating an essay*	*Change rating decision on some descriptors*	*Compare self-rating results with teacher-rating results*
17	L	21.5	30.0	27.0	no	*20-30*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
32	L	28.6	26.6	26.3	no	*30*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*n/a*	*n/a*	*no*	*yes*	*yes*
39	L	27.6	29.0	27.6	no	*10-15*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*	*yes*	*yes*
46	L	30.3	28.3	26.0	no	*15-20*	*yes*	*yes*	*no*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
54	L	22.6	28.3	27.0	no	*10-15*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*
03	M	26.5	26.5	33.0	no	*15-20*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
04	M	31.5	29.5	30.5	no	*30*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
09	M	28.0	25.5	30.0	no	*5*	*yes*	*yes-no*	*yes*	*yes*	*yes*	*yes*	*yes*	*n/a*	*no*	*yes*	*yes*
12	M	29.0	24.5	29.0	no	*20-30*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
23	M	28.6	27.3	27.6	no	*15*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
56	M	26.0	28.6	26.3	no	*5-15*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
57	M	25.3	29.0	23.6	no	*5*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*no*	*yes*	*yes*
61	M	28.0	30.6	25.0	no	*5-10*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*
63	M	28.3	30.3	25.6	no	*5*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*n/a*	*yes*	*no*	*yes*	*yes*
64	M	27.6	29.0	27.6	yes	*10-20*	*yes*	*yes*	*n/a*	*yes*	*yes*	*n/a*	*yes*	*yes*	*yes*	*yes*	*yes*
21	H	32.5	32.5	26.5	yes	*30*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
26	H	28.5	27.0	30.5	yes	*20*	*yes*	*yes-no*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
31	H	30.5	27.5	30.5	no	*5-10*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*yes*	*yes*
62	H	28.5	27.0	30.5	no	*5-10*	*yes*	*yes*	*yes*	*yes*	*n/a*	*yes*	*n/a*	*n/a*	*yes*	*yes*	*yes*
65	H	28.3	28.0	28.3	no	*10-15*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*

5.2.2 Students' Perceptions of the Scale

Table 5.3 shows the main themes and subthemes of the students' perceptions of the scale. As was the case with the teachers, student perceptions could be classified into three main themes: scale functioning, scale usefulness, and scale impact, each with associated subthemes. The student perceptions are presented in the sections that follow.

Table 5. 3 *Summary of Students' Perceptions of the Scale*

Main themes	Sub-themes	Perceptions
Functioning		
▪ Comprehensibility	• Criteria judgement	- Some descriptors are difficult to judge.
▪ Comprehensiveness	• Criteria specificity	- Binary rating is not detailed and more rating options should be added.
	• Criteria coverability	- Other skills should be added to the criteria.
▪ Applicability	• Scale criteria	- The scale is well organised and easy to use. - Micro-skill descriptors should come before macro-skill descriptors.
	• Rating format	- The rating option should be verbal labelling instead of numeric labelling.
Usefulness		
▪ Learning	• Diagnostic information	- Self-assessment help to know and realise writing strengths and weaknesses.
	• Self-assessment	- Self-assessment help to guide and revise writing essays - Self-assessment help to become attentive to or engaged in learning and writing - Self-assessment help and does not help to become motivated in learning and writing. - Peer-assessment should complement self-assessment for more reliable and unbiased assessment.
	• Diagnostic feedback	- Teacher's scale-assisted feedback is useful for writing revision and improvement.
	• Learning improvement	- Self-assessment helps to improve writing.
	• Diagnostic report	- The scale should have the overall evaluative description of the diagnostic results.
Impact		
▪ Awareness raising	• Self-assessment	- Self-assessment is useful and should be included in writing classroom.

5.2.2.1 Functioning of the Scale

The student perceptions of the scale functioning covered three areas: scale comprehensibility, comprehensiveness, and applicability. Scale comprehensibility is concerned with criteria clarity and ease of judgement. In general, 19 students found it difficult to judge certain descriptors. Of these descriptors, those frequently mentioned included (*a) coherence (n = 4), (b) cohesion (n = 4), (c) content distribution (n = 4*), *(d) supporting idea arrangement (n = 3),* and *(e) sentence problem (n = 3)*. Sample comments are shown below:

12L: *It's not that difficult but I feel like I am not confident if I should give ZERO or ONE. [Can you give example?] It's pretty much about errors and appropriate length and paragraph balancing. [Is it content in the paragraph?] I mean I am not sure if five paragraphs are well balanced. Sometimes, the introduction is shorter than others and sometimes the conclusion is longer than others.*

61M: *[Were there any particular rating scale descriptors you found difficult to understand and judge? If yes, please identify] Coherence and cohesion. [Why?] It's like I have to rethink, for example, main ideas and supporting ideas, sometimes I am not sure if the teacher will think it is appropriate. I think it's appropriate but the teacher may not think it's appropriate. [Anything else?] No.*

The student perceptions of scale comprehensiveness were related to criteria specificity and coverage. Regarding the criteria specificity, nine students were concerned that the binary rating was not sufficiently detailed and should include more rating options as indicated below:

09M: *[Do you think ONE and ZERO is easy to rate?] Yeah but I think it doesn't have to be only ONE and ZERO. (So, do you think there should be more options?) Yeah, this one is like "have" and "don't have". [For all descriptors?] No, for example, most descriptors in the front page of the scale, but for grammar I think it is more objective and only ONE and ZERO is ok for this.*

57M: *I think there should be five rating options because it's more detailed but it's time-consuming.*

In relation to criteria coverage, two students suggested that additional criteria should be assessed, including: (a) *citation of information sources* and (b) *standard or native-like English*, as the following quotes illustrate:

26H: *If possible, you may have a descriptor about how to cite references or the sources of information for writing.*

39L: *Em, I think there is no descriptors talking about the use of natural or native-like sentences. The teacher often told me to write naturally so that it is easier to understand by international people.*

The student perceptions of the scale applicability are concerned with the scale criteria and rating format. In general, all students perceived that the scale was well-organised and easy to use as illustrated below:

03M: *It is very easy to assess my essay. [Do you agree with categorisation of the domains?] Yes, it makes me identify the problem more easily.*

04M: *It is easy to understand because it has clear explanation of the three parts of the essay as well as other minor elements. It makes my writing read more smoothly.*

46L: *I think it's easy to use. There are only ZERO and ONE. If there is the middle point, we tend to tick the middle point.*

However, one student perceived that micro-skill descriptors such as grammar, sentence and vocabulary should precede macro-skill descriptors such as organisation, coherence, and cohesion as illustrated by the following quote:

64M: *[What about the order the arrangement of the descriptors] I think coherence and cohesion should be put at the end of the scale or after grammar. [Why?] I can check grammar skills first which are small details in sentences and then check coherence and cohesion which are overall writing picture to see if all paragraphs are related.*

Apart from descriptor ordering, two students preferred that the rating options be verbally labelled as "Yes" and "No" instead of "0" and "1" as illustrated by their comments:

32L: *[What about ZERO and ONE?] Em, it's ok but it should be letter like "Yes" and "No" instead of number like ZERO and ONE.*

39L: *[What about rating options?] I think it should be "Yes" or "No" instead of ZERO and ONE. The rating option should be "Yes" and "No".*

5.2.2.2 Usefulness of the Scale

The student perceptions of the scale usefulness were concerned primarily with learning promotion. To begin with, all students perceived that the scale-assisted self-assessment helped them to identify their writing strengths and weaknesses as the following comments illustrate:

21H: *[Did the rating scale descriptors help you to know your strengths and weaknesses in writing?] Yes, em often I am not sure about how to use articles such as when to use "the" and I am not sure if the ideas are related or not. When I check the descriptor, I have to think more and revise my essay again.*

23M: *I think this project is useful to me because it helped me to focus on small detailed I previously overlooked, and it helped me to know my strengths and weaknesses. For example, punctuations, I normally think it is not difficult to use punctuations but after using the scale, it is not like what I thought.*

Regarding self-regulated learning, all students perceived that the scale-assisted self-assessment helped them to write and revise their writing as the following examples of the student quotes explicate:

17M: *I helped me a lot to realise and spot my weak points in my writing. Before submitting an essay, I have to be more careful and pay more attention to an essay. Before doing self-assessment, I didn't check an essay much before submitting but when having self-assessment, I feel like I have to reread, recheck an essay again before submitting an essay.*

62H: *Em first, I never knew before that when writing we can self-rate our work first before the teacher rate our work. After doing self-assessment in this project, I feel like before submitting an essay, we have to self-rate our essay first to see how many scores we should get before submitting an essay. If we get low scores, we have to correct and revise an essay before submitting.*

Moreover, all students reported that they became more attentive to and engaged in learning and writing as revealed in the quotes below:

21H: *Yes, I enjoyed the writing course more and paid more attention to my writing. For example, I compared the first and second drafts, I can see that the second draft was better than the first draft. [How were you engaged?] I think I was more careful when I write an essay and put more effort because I didn't want to get ZERO.*

63M: *Yes, when I didn't have the rating scale, I didn't feel like I want to learn but when I have the scale, I have the guideline and pattern and I follow the scale step by step and this make it easier for me to write and I feel like I want to learn.*

Apart from learning engagement, almost half of the students reported that they felt motivated to write and learn during the course as a result of implementing the scale for self-assessment. It was observed that some students started to perceive the benefits of the scale for learning after overcoming their initial resistance to using the scale while others were more motivated to write after realising the value of the scale for learning as the following comments illustrate:

56M: *[Do you feel motivated after doing the self-assessment?] At first, it is not because I think that there are so many things to do, but after using the scale, my writing improves, so I see how important it is, I notice my improvement, so I feel motivated.*

65H: *Em writing is not my favourite subject [What is your favourite skill?] Speaking. It's not that I don't like writing but I feel like the scale helps me to write better and when I can write better I want to write.*

However, 11 students reported that using the scale for self-assessment did not, in itself, help them to become more motivated in learning and writing. Of the 11 students, four perceived that it is rather the writing topic that motivates them to learn and write (as

commented by 12L below) and two perceived that it is the teacher's teaching style that motivates them to learn and write (as commented by 23M below):

12L: *Yes, I pay attention to learning [Did you have motivation in learning?] Motivation for me I think it's about the topic of writing that motivates me to write. If I am interested in the topic, I am happy to write and want to learn but if the topic is not interesting, I feel like I am not happy to write and for some topics I have to search more information about the topic and when I don't find much information on the topic I don't really want to write anymore.*

23M: *Em, I think I was more attentive to writing and focused more on details in writing. [Did you feel you want to study writing more?] Em, for me I think the teacher is the main reason that makes me want to learn in the course.*

For a more effective self-assessment, three students suggested that self-assessment should be accompanied by high-ability peers' or more-experienced writers' assessment for reliable and unbiased results as illustrated by the following comments:

04M: *The scale is good. I think that it should have peer assessment because self-assessment is not reliable because it can be biased.*

57M: *It should and should not be included. I think it's not necessary because when student self-rate, we may not see mistakes or problems. More experienced writers or raters may be able to better rate writing and provide more useful comments. But self-assessment may help me to review my writing skills and what I write.*

In addition to self-regulated learning, all students perceived that the teacher's scale-driven feedback was useful for their writing revision and improvement as demonstrated in the selected quotes below:

31H: *Yes, like I said I think I did well on some descriptors and I gave ONE but the teacher gave ZERO. So, I went back and revised that skill [Did you remember the descriptor?] No.*

65H: *[Did you find your teacher's feedback from the teacher rating scale helpful? How and give examples?] Yes, it is useful. [How?] During the first task, the teacher gave very detailed feedback and after that Teacher Cali told me that my ideas are better but there are some skills such as fragment and tense that I still made mistakes.*

In terms of learning, all students perceived that the scale-assisted self-assessment helped them to improve writing as the following examples illustrate:

23M: *[Do you think that the use of the diagnostic rating scale for self-assessment helped you improve your writing ability?] Yes. [How?] Like I said, if I forget to write some skills or elements such as compound and complex sentences, I would change from simple to compound or complex sentences.*

31H: *[Do you think that the use of the diagnostic rating scale for self-assessment helped you improve your writing ability?] Yes. [How? can you explain more about it?] Em, it helps me to learn how to paraphrase information from other resources and I think my essay is more united and has various types of simple, compound and complex sentences and various words as well.*

62H: *[Do you think you make progress in learning writing in the course?] Yes. [In what way, better or worse?] It's better. Like I said, I used the criteria as a guideline when I write my essay. [In your opinion, how much do you think your writing improve in percentage?] Em, it less than 50%. [How much on average?] Em, it's about 30%. [How much did the self-assessment help your improvement apart from other things like teacher' instruction or feedback? how much percentage for self-assessment?] Em, it's around 15-20%. [Ok anything else you want to say more about self-assessment?] The scale helped me to know the criteria that are used to rate writing and give scores and what skills or elements in an essay I should use when scoring writing.*

To make it easier and faster to interpret the diagnostic results, one student (see 62H below) suggested that the scale should have an overall evaluative description of the diagnostic results, which would make it easier to know her overall writing strength, weakness and progress:

62H: *I want to add more information. For example, if my essay gets more than half out of 33 descriptors, then I pass or if my essay gets more than 25 descriptors, my essay is good or if my essay gets low scores, what skills should be further improved. I also used the rating scale in my assessment and evaluation course, the teacher commented that the scale is very good, but it doesn't have the evaluative description of the diagnostic results.*

5.2.2.3 Impact of the Scale

With respect to the scale impact, after using the scale for self-assessment, all students reported that the scale-assisted self-assessment was useful for improving their particular essay writing skills and it should, therefore, be included in writing classrooms as the following quotes demonstrate:

61M: *[Do you think self-assessment should be included in writing classroom?] I think it should because it will help my writing to look good. If I submit an essay without self-assessment, I will know my results only from the teacher's feedback and then I know what to revise. But when I self-rate my essay, I know what to correct before submitting an essay so it also helps the teacher as well.*

63M: *[Do you think self-assessment should be included in writing classroom?] I think it should be included because when we learn in class we forget what we learn and the teacher just give lecture in class. But in the scale, we have pattern and skills that we can follow when we write.*

5.3 Chapter Summary

This chapter has presented the qualitative findings pertaining to teacher and student perceptions of the scale. Notwithstanding some negative reactions to the scale, the teachers and students largely have positive perceptions of the scale functioning, usefulness, and impact. In addition, the students employed important self-assessment strategies, which implied that they were attentive to the self-assessment process and thus

were developing self-regulated learning skills. This chapter emphasises that the scale users' perceptions are necessary for validation research. Despite the acceptable psychometric properties of the scale, this does not guarantee that the users are satisfied and comfortable with the scale characteristics, functionality, and applicability. In the next chapter, all the quantitative and qualitative findings will be synthesised and discussed in response to the research questions and in justification of the validity argument.

Chapter 6: Discussion

This chapter synthesises and discusses all the research findings, presented in Chapters 4 and 5, in response to each of the research questions and in justification of each of the validity inferences. In this chapter, the research findings are first discussed in light of existing research and literature and then the overarching validity argument for the binary diagnostic rating scale is evaluated in terms of the soundness and coherence of the evidential sources provided to justify the validity inferences.

6.1 Results Related to Research Question 1

The first research question – *To what extent does the diagnostic rating scale function appropriately for the formative diagnostic assessment in the EFL writing classroom?* – investigated the appropriateness of the scale functioning from perspectives of psychometrics and user perceptions. The results for Research Question (RQ)1 provide empirical evidence to justify the evaluation, explanation, and extrapolation inferences. Overall, the psychometric properties of the scale functioning and rater behaviour as well as the teacher and student perceptions support the appropriate functioning of the scale, with some possibilities for further revisions of the scale in the Thai EFL university writing classroom context.

6.1.1 Psychometric Indicators of the Scale Functioning

Based on the psychometric indicators of the scale functioning, most of the MFRM, CTT, and descriptive indices indicated that the scale functions appropriately. First, the scale generated the observed scores acceptably matching the expected scores generated by the Rasch model as informed by: *(a) the acceptable data-model fit indicators and (b) no overfitting and underfitting (misfitting) descriptors*. Second, the descriptors were independent of one another in yielding the observed scores as shown by: *(a) the acceptable local independence indicators* and *(b) no overfitting descriptors*.

Third, the descriptors were consistent with one another in capturing the prime dimension of the defined construct as the following indicators illustrated: *(a) the acceptable data-model fit indicators, (b) the acceptable unidimensionality indicators, (c) the*

acceptable fit indices of all descriptors, (d) acceptable Point-Measure (PTM) correlations of 32 descriptors, (e) the high alpha internal consistency reliability, and (f) the acceptable corrected item-total (CIT) correlations of all descriptors.

Fourth, the scale succeeded in targeting the wide and varied range of the defined construct of student writing ability as demonstrated by: *(a) the wide visual dispersion of student logits, (b) the significant heterogeneity index of student logits, (c) the high separation indices of student logits, (d) no overfitting descriptors, (e) the reasonable range (3.55 logits) and SD of student logits and scores.* The student variability differentiated by the current scale is relatively consistent with that of Kim's (2010) EDD scale (4.6 logits), and Knoch's (2007, 2009b) analytic scale (around 3.0 logits) given the fact that the number of examinees in this research is much lower than that in Kim's and Knoch' studies.

Fifth, the descriptors spanned the wide and varied range of difficulty levels as evidenced by: *(a) the wide visual distribution of descriptor logits, (b) the significant heterogeneity index of descriptor logits, (c) the high separation indices of descriptor logits, and (d) the reasonable range of descriptor logits and CTT difficulty indices.* Even if the extremely easy and difficult descriptors were to be excluded, the difficulty would range nearly 3.0 logits, relatively close to the difficulty range of Kim's (2010) EDD scale (3.23 logits).

Lastly, the scale was found to differentiate student assignment performances in a way consistent with their achievement exam results as indicated by: *(a) the student locations on the visual variable map, (b) significant differences in the diagnostic results between student ability groups on the achievement exam, and (c) significant, positive, and strong correlations between the diagnostic and achievement exam results*. It should be noted, however, that although the formative diagnostic and summative achievement assessments were based on different writing tasks and rating criteria, the teachers' ratings of the students' essays on the achievement exam tasks could be, at least to some extent, influenced by their familiarisation of the diagnostic criteria. This is evidenced by some teachers reporting that the use of the scale helped them to better understand the rating criteria on the achievement exams.

6.1.2 Psychometric Indicators of Rater Behaviour

In relation to the psychometric indicators of rater behaviour, most of the MFRM and descriptive results demonstrated that the raters applied the scale appropriately. To start with, the raters' observed ratings acceptably matched the expected ratings generated by the Rasch model. In other words, they did not exhibit a lack of rating variability and abnormal rating as indicated by no overfitting and underfitting raters.

Additionally, the raters did not exhibit unrealistically too similar agreement as their observed agreement slightly exceeded the expected agreement generated by the Rasch model. This was suggested by the Rasch-Kappa slightly over 0 and the percent observed exact agreement slightly over the expected agreement. Moreover, the raters behaved independently of one another when rating the essays as indicated by no overfitting raters and students.

Finally, the raters were able to target the wide and varied range of the defined construct as suggested by: *(a) the significant heterogeneity index of student logits, (b) the high separation indices of student logits, (c) the wide visual distribution of student logits, (d) the reasonable SD and range of student logits and scores, and (e) no overfitting raters and students*.

6.1.3 Rater Perceptions of the Scale Functioning

In terms of the raters' perceptions of the scale functioning, the teachers and student were largely positive about the appropriateness of the scale functioning. Firstly, the teachers and students perceived that the descriptors were mostly comprehensible, suggesting their confidence in interpreting and judging the descriptors and the clarity and interpretability of the descriptors. However, some descriptors, including *D03 (Topic sentence relevance), D11 (Supporting idea logic), D12 (Main idea unity), D17 (Content comprehension), D28 (Sentences accuracy), D29 (Word choice), D30 (Word variety),* were perceived by individual teachers as difficult to judge and most of these descriptors contain subjective terms, for instance, "*convincing*", "*enough*", "*appropriate*", and "*appropriately*", which were also found in previous research to be unclear and subjective to raters (Kim, 2010). Such subjective terms represent a degree or continuum of writing quality, which could be variably interpreted by individual raters and be difficult for them to interpret and

judge dichotomously. It is, however, impossible to eliminate entirely the subjective nature of performance assessment criteria.

Secondly, while the teachers generally viewed the descriptors targeting discrete and specific writing skills in keeping with the scale designer's intention to elicit fine-grained diagnostic information as suggested by several scholars (e.g., Alderson, 2005; Alderson et al., 2015; Lee, 2015), the teacher perceptions revealed some issues related to error counting and the binary rating of specific skills.

While the intention of using error counts to measure the quality of grammar, sentence, vocabulary, and mechanics skills was to make the rating of such skills more objective, some teachers realised that the number of errors in certain essays did not give an accurate picture of students' writing ability. They observed that lower-quality essays tended to have fewer errors as was likewise discerned by raters in Kim's (2010) study. This is probably because low-proficiency writers are more likely to play safe, whereas those of high-proficiency writers tend to take risks using advanced and complicated grammar and language. Moreover, some of the counted errors might not genuinely reflect students' deficient writing knowledge as these might be unsystematic and random errors caused by memory lapses, physical tiredness, emotion, and/or a slip of tongue (Corder, 1981). Furthermore, the counting of errors is by its very nature a meticulous and error-prone process (Xie, 2019). In light of this, the composite score based on error-counted descriptors could misrepresent the overall quality of an essay or each student' writing ability, as was also pointed out by Kim (2010). The current and prior findings suggest the tension between specific and overall estimates of student's ability.

Apart from the problem of error counting, binary rating, while practical and easy-to-judge, does not provide detailed diagnostic information on individual skills. In general, the teachers and students found the binary rating easy to judge and practical, which is consistent with Park and Yan's (2019) finding showing that some raters found binary rating as easier and more objective to judge than a holistic scoring. However, some studies revealed that raters found binary rating as cognitive-loaded and difficult to judge (Kim, 2010) and more difficult to judge than holistic scoring (Park & Yan, 2019). Despite its ease of use, some teachers and students in this study found binary rating a rather crude means of capturing the detailed quality of specific skills, which corroborates the finding of previous research (Kim, 2010). The present and previous findings indicate that raters have

mixed opinions about binary rating both with respect to its ease of application and the specificity of diagnostic information provided. While detailed rating options may have the value of providing fine-grained feedback about writing quality, too many rating options could increase raters' cognitive load unduly and create difficulty in discriminating between rating options or points, hence contributing to measurement errors (DeVellis, 2017; Johnson & Morgan, 2016).

Thirdly, the teachers perceived that the scale largely represented the essential writing skills taught in the classroom although individuals later mentioned additional features to be assessed, including *(a) academic language (b) consistent use of English style (c) genre-specific features, (d) more grammar features, (e) overall essay impression, and (f) standard or natural English*. This is probably because the teachers were able to pilot-rate only a small number of student essays due to time constraint in the scale trialling stage. Thus, they did not have enough exposure to the various writing problems, features, and skills produced in a wide variety of student writing performances. While increasing the number of scripts used at the piloting stage would clearly be desirable if time permitted, there are practical limitations on the amount of detail that can be covered in a scale designed for repeated application in the classroom context. Precisely for this reason, it was decided at the outset that including genre-specific features in the scale, as recommended by a small number of respondents, was not feasible and would be too demanding for the teachers.

Finally, the teachers and students generally perceived that the scale characteristics, including scale layout, criteria structure, and rating format, were applicable although some of them expressed certain negative perceptions about scale length, descriptor ordering, and labelling and ordering of rating options. This indicates that while the teachers, following rater training, appeared to provide appropriate ratings, they were not wholly comfortable or satisfied with the scale functionality. For example, some teacher suggested that the micro-level descriptors (grammar, sentence, vocabulary, and mechanics) should come before the macro-level descriptors (organisation, content, coherence, and cohesion). It can be argued, however, that the ordering of the descriptors in the current scale may force teachers to focus on the macro skills first, which could have a positive washback effect as it might force them to first read a writing sample for more global aspects, which is thought to represent more natural reading, rather than focusing on

linguistic features, in particular errors, first. Besides, some students recommended that the rating options should be verbally labelled. Previous research revealed that variations in scale labelling (verbal labels and numerical values) on rating scales can affect respondents' response style, measurement quality, and cognitive process (Menold, 2020; Moors et al., 2014) and variations in the order of verbal labels and numerical values can impact respondents' responses (Betts & Hartley, 2012). This suggests that variability in criteria structure and rating format on a rating scale can differently impact individual raters. It is not clear to what extent the negatively perceived features could affect the raters' decision-making process and the score variability in this study. However, these features deserve particular attention for future scale revision to reduce raters' psychological and cognitive load, to facilitate raters' rating strategies and decision-making processes, and to avoid construct-irrelevant and erroneous ratings. That being said, it is impossible to develop a rating scale that perfectly suits all needs, given the wide variety of raters' rating styles, strategies and preferences which have been revealed in previous studies (Cumming et al., 2002; Han, 2017; Zhang, 2016).

In summary, notwithstanding acceptable psychometric results showing the appropriate functioning of the diagnostic scale, the teacher and student users were not entirely positive about the scale functionality. This emphasises that psychometric evidence favouring scale validity does not guarantee the usability of an assessment instrument from the raters' perspective. Positively-perceived descriptors and scale characteristics are more conducive to facilitating raters' scoring decisions, leading to accurate ratings while poorly perceived ones are prone to negatively influence raters' cognition and decision-making, potentially giving rise to measurement errors and construct-irrelevant ratings (Lane, 2019). It should be noted that the teachers' negative perceptions are arguably of greater concern in this research as their scoring decisions impact the interpretations and uses of the scale-based information in the classroom, whereas those from the students are of less concern as they are typically associated with limitations in their writing knowledge. The negatively-perceived descriptors may be considered for exclusion to increase the quality and practicality of the scale. However, by deleting poor items without considering content coverage, the scale might fail to capture important aspects of the construct, resulting in construct underrepresentation (Schumacker, 2004). Therefore, the difficult-to-judge descriptors, together with the negatively perceived scale properties, should be revised

rather than excluded to ensure that important skills students need to learn are covered and that the quality of ratings and scores is maintained.

6.2 Results Related to Research Question 2

The second research question – *To what extent does the diagnostic rating scale function consistently for the formative diagnostic assessment in the EFL writing classroom?* – examined the consistency of the scale functioning by looking at the psychometric indices of the scale functioning and rater behaviours. The RQ2 findings yield evidence pertinent to the generalisation and explanation inferences. Considering the varied, low-stakes, and non-standardised nature of the current classroom assessment, the psychometric indices of the scale and raters indicate acceptable consistency of the scale functioning.

6.2.1 Psychometric Indicators of the Scale Functioning

As regards the statistical indices of the scale, the MFRM and CTT results generally confirm that the scale is consistently applied across the raters to an acceptable extent. To elaborate, individual descriptors functioned consistently across the raters and the student essays as suggested by: *(a) no underfitting descriptors, (b) no underfitting students, (c) no underfitting raters, (d) acceptable PTM correlations of 32 descriptors, (e) high alpha internal consistency reliability, and (f) acceptable CTT percent rater agreement of most descriptors.* The relatively high consistency of the current binary scale is in line with previous findings indicating that a binary rating tends to be judged highly consistently (Kim, 2010; Wagner, 2015) and more consistently than a multiple-point or polytomous scale (Park & Yan, 2019). Yet, six descriptors (D11, D14, D20, D23, D28, D29) showed unacceptably low percentages of interrater agreement. Amongst these, D11, D14, D20, D23, and D28 exhibited relatively or very high difficulty logits. To put it another way, most students did not do well on the skills associated with these descriptors. D11, D28, and D29 were perceived as difficult to judge by some teachers and most involved the counting of errors and subjective terms. The unacceptably low percent agreement of the descriptors is not nevertheless of grave concern since the MFRM indices suggested their consistent functioning.

6.2.2 Psychometric Indicators of Rater Behaviour

The psychometric indicators of the raters' behaviour, considering the raters' varying background, largely support an acceptable level of consistency in judging the descriptors and diagnosing the student writing performances over the three sequential tasks despite their differing interpretation of some descriptors.

To start with, the raters were acceptably congruent in judging the majority of the descriptors and diagnosing the student performances as evidenced by: *(a) the acceptable percentages of interrater agreement on most descriptors and students and (b) the high Rasch percent observed exact agreement*. This finding corroborates those of previous studies revealing that well-trained teachers exhibited a Rasch observed exact agreement of about 64.7% (Kim, 2010) and average interrater agreements of about 77.3% (Kim, 2010), about 85.2% (Wagner, 2015), and between 80.53% and 70.18% (Park & Yan, 2019) on binary rating. However, previous research showed that raters demonstrated lower agreements for holistic rating with an average agreement of 34.37% (Park & Yan, 2019), and for analytic judgement, with Rasch observed exact agreements of about 37.9% and 51.2% (Knoch, 2007, 2009b), and between 15.2% and 43.8% (Park & Yan, 2019). Based on the present and prior findings, it can be inferred that raters tend to judge dichotomous rating choices more consistently than multiple rating options.

In addition, the raters differed substantially in their levels of severity, suggesting that they interpreted the descriptors in different ways. This is indicated by: *(a) significant heterogeneity index of rater logits, (b) high separation indices of rater logits, (c) wide visual dispersion of rater logits, and (d) noticeable SD and range of rater logits.* In spite of the raters' differing severity, each rater exhibited consistent level of severity across the essays and descriptors, providing partial evidence of fairness, as illustrated by: *(a) no underfitting raters, (b) no underfitting students, and (c) no underfitting descriptors.* Given the raters' limited experience in teaching essay writing and limited formal training on essay rating, their severity differences are to be expected and not significantly higher than those exhibited by well-trained raters in Kim's (2010) study (showing a severity range of 1.08 logits) and by professional raters in Knoch's (2009b) study (showing a severity range of almost 0.5 logits and nearly 1.0 logits). The raters' heterogenous severity could potentially be caused by such factors as the characteristics of the scale, the rater training and rating conditions, and individual raters' background and personality traits, low experience in

essay teaching and rating, and workload during the ongoing teaching and assessment. These sources of rater variability are flagged in the literature (e.g., Eckes, 2015; Engelhard, 2013; Knoch & Chapelle, 2018; Weigle, 2002). It is well-established in a body of research that even well-trained and experienced raters tend to differ in their rating severity (e.g., Elder et al., 2005; Elder et al., 2007; Knoch, 2011; Knoch et al., 2007).

Finally, there was evidence of relationships between raters' decision-making behaviours, descriptor difficulty indicators, and student essay scores. The teachers were inclined to judge easier descriptors more homogenously than judging harder descriptors as demonstrated by the significant and strong correlations between the rater agreement percentages and the descriptor difficulty indices. One possible explanation may be that difficult descriptors represent challenging or advanced writing skills which are highly abstract and complex in nature and thus necessitate a more complicated decision-making process, resulting in high rater variability on these skills. Another explanation may be the limited proficiency of the raters themselves which might have influenced the raters' confidence in judging. This finding is partly supported by another quantitative finding that most of the descriptors showing unacceptably low agreement also demonstrated high difficulty logits. In addition, the teachers were inclined to judge higher-score essays more homogenously than judging lower-score essays as indicated by the significant and strong correlations between the rater agreement percentages and the student essay scores. This result corroborates previous findings that variability of essay characteristics differentially influences raters' decision-making behaviours (e.g., Barkaoui, 2007a; Han, 2017; Huang et al., 2014; Şahan, 2018; Şahan & Razı, 2020). The decreased rater agreement on lower-score essays may be owing to the fact that these essays may include textual features or writing skills that the student were developing and acquiring and thus it is not clear whether such features or skills are weak or strong, making it difficult for the raters to make binary choices.

Ideally, the teachers should have interpreted all descriptors and judged all essays homogeneously irrespective of descriptor difficulty or essay quality as failure to do so may result in systematic rater error or construct-irrelevant ratings. As with any rater-mediated assessments, it is impossible to eliminate all the factors that negatively impact raters' decision-making process. Given that the rater consistency statistics in this study remained within acceptable bounds, it can be concluded that the descriptor difficulty and essay

quality did not affect the raters' rating consistency to a worrying degree. For future training of raters, the difficult descriptors or skills should receive more attention and time than easier skills so as to norm raters' interpretation of the descriptors. It is also important to train raters particularly on judging poorly-written essays by providing them with benchmarked models of low-quality essays with various characteristics and with more practice on rating poor essays. A think-aloud method may be used to uncover why raters tend to differ when binarily judging difficult descriptors and lower-score essays.

To sum up, despite the raters' varying severity levels, statistical indicators confirm the acceptable consistency of the scale functioning. Under the MFRM approach, as long as each rater applies the scale consistently either too harshly or too leniently in relation to other raters, it is not necessary for all raters to reach consensus because differences in rating severity can be estimated and corrected for in the estimation of each student's ability logit and fair average (Linacre, 1989, 2018; McNamara et al., 2019) Moreover, high rater consistency and consensus are deemed as less critical in formative classroom assessment which is in nature less-formal and non-standardised (Andrade & Heritage, 2018).

6.3 Results Related to Research Question 3

The third research question – *To what extent does the diagnostic rating scale support formative decisions about teaching and learning in the EFL university writing classroom?* – examined whether the diagnostic rating scale supports teachers' and students' formative decisions about teaching and learning in the classroom. The results for this question are obtained primarily from the qualitative content analyses of the teacher and student perceptions of the scale and are used to justify the decision inference. In general, the perception findings revealed that the scale is useful to support teaching and learning by virtue of its practical usefulness and meaningful diagnostic information provided in the classroom.

6.3.1 Practical Usefulness of the Scale

In general, the teachers and students perceived that the scale is easy to use even though one teacher found it somewhat time-consuming for multiple assessments due to the number of descriptors. As discussed in the RQ1 results, scale practicality is strongly

associated with criteria comprehensiveness. While diagnostic criteria should target discrete and specific language skills and cover essential skills and learning contents (e.g., Alderson, 2005, Lee, 2015), overly comprehensive criteria could reduce the practicality of the scale, which is deemed important for classroom assessment (Alderson et al., 2015). Just like determining an appropriate level of diagnosis specificity, determining an appropriate level of criteria coverage, while ensuring practicality, is challenging and may depend on assessment purposes, learning and teaching practices, and stakeholders' needs in a given context.

6.3.2 Usefulness of Diagnostic Information

Apart from the scale practicality, the teachers and students expressed that the scale provides diagnostic information interpreted as writing strengths and weaknesses although some of them preferred a brief report which would help better understand not only writing strengths and weaknesses but writing progress as well. This suggests that the use of the binary rating options (0 and 1) helped the teachers to differentiate between the strength and the weakness on specific skill as indicated by the teacher comments. There was also evidence from the perception findings that the inclusion of several discrete-point descriptors in the checklist helped the teachers to know students' strength and weakness on specific skills. This may be due to the fact that in the Thai EFL university classroom context under study, the teachers normally use analytic rating criteria to evaluate student essays and thus tend to focus more on a few global traits or domains (e.g., content, grammar, organisation) rather than more specific skills. This suggests that the use of the binary checklist could provide specific and useful diagnostic information in supporting the teachers' and students' diagnostic decision about writing strengths and weakness and about teaching and learning adjustment.

Although the current binary diagnostic rating scale on its own could, to a reasonable extent, help the teachers and students to understand strengths and weaknesses on specific skills, it is, however, simply a score-based reporting tool and lacks the detailed verbal and/or graphical descriptions of the diagnosed skills which might make it easier, clearer, and faster for teachers to give formative feedback and for both teachers and students to identify writing strengths, weaknesses, and improvement, as commented by one teacher (Nana) and student (62H). As pointed out by Jang (2012), Jang

and Wagner (2014), and Kunnan and Jang (2009), a well-designed, detailed, and individualised diagnostic profile report is recognised as a necessary component of diagnostic assessment as it can engage students in and also attract students' attention to the diagnostic information and feedback. However, generating a quality diagnostic profile report takes time and need assistance from technology to optimise the provision and quality of diagnostic feedback in an ongoing classroom (Alderson, 2005; Kunnan & Jang, 2009). The lack of a more user-friendly diagnostic profile report somewhat undermines the meaningful interpretation and practical usefulness of the diagnostic information in this study.

Although qualitative findings revealed that the scale and diagnostic results were generally perceived to support formative decisions about teaching and learning improvement, it was revealed that the scale may not effectively support low-ability students' learning as they are not able to self-diagnose whether their skills are weak or strong. One teacher and some students suggested that low-ability students may need more support from teachers or higher-ability peers so as to effectively apply the scale and prevent unbiased assessment. The concern about low-ability students' self-assessment seems to be supported by the quantitative results showing significant rating difference and very small correlation between the self-assessment and teacher-led assessment, which corroborates previous findings (Brown & Harris, 2013; Ünaldı, 2016). The teacher's and students' suggestions are in line with those proposed in previous studies that self-assessment should be complemented with peer-assessment which was found in previous research to be more accurate and consistent with teacher ratings (Esfandiari & Myford, 2013; Hung et al., 2016; Matsuno, 2009; Salehi & Masoule, 2017) and less biased and independent of their own writing performance (Matsuno, 2009).

To sum up, the qualitative findings generally confirmed that the diagnostic rating scale was practical and useful to support the formative decisions to improve teaching and learning in the Thai EFL university writing classroom. However, some findings uncovered that the lengthy number of descriptors and a lack of a diagnostic profile report might reduce the scale practicality and meaningful interpretation of the diagnostic information respectively. In particular, low-ability students might not make appropriate decisions due to their deficient knowledge of language and learning contents, and thus need closer monitoring and more support from teachers or higher-ability peers.

6.4 Results Related to Research Question 4

The fourth research question – *To what extent does the formative diagnostic assessment have beneficial consequences for teaching and learning in the EFL university writing classroom?* – investigated whether the scale-driven formative diagnostic assessment led to beneficial consequences on teacher instruction, student self-regulated learning, student learning progression, student learning achievement, and assessment impact. The RQ4 was examined through descriptive, ANOVA, correlation, regression, and MFRM of the diagnostic scores and the qualitative content analyses of the teacher and student perceptions of the scale. The results provided evidence to justify the consequence inference.

6.4.1 Teacher Instructional Practice

In general, qualitative findings support the notion that the assessment helped improve the teachers' instructional practice. To begin with, the scale-based diagnostic information helped the teachers to give detailed and targeted feedback to their students, which is also confirmed by the students' comments that the teachers' scale-supported feedback was useful for them to revise and improve their essays. As discussed earlier, the inclusion of several and specific descriptors on the scale helped the teachers to not only identify strengths and weaknesses on specific skills but also provide detailed and targeted diagnostic feedback to the students. It was also observed from the interview data that there were some variations in the way individual teachers provided diagnostic feedback, generated by the scale, to their classroom students. Some teachers focused on providing feedback on specific skills for individual students as reflected in one teacher' s (Nana) comment and some focused on giving overall patterns of strengths and weaknesses to the student group as mentioned in one teacher' s (Sara) feedback. This suggests that the formative value of the diagnostic scale and associated feedback on the students' learning could vary depending on the teachers' modes and methods of feedback provision, as also acknowledged by Jang and Wagner (2014).

Furthermore, the scale was perceived by some teachers as a useful teaching resource and guideline. This result indicates that the diagnostic scale positively influenced ongoing teaching, providing partial evidence that the current assessment was really integrated into ongoing teaching and learning as was intended. In addition, the teachers

reported that the scale criteria helped them to better understand and more carefully interpret achievement assessment criteria used in the midterm and final exams, suggesting another positive consequence of the scale on the part of teachers. One explanation may be that the teachers never before interpreted detailed assessment criteria on several and specific skills until they were trained to judge the specific descriptors on the current scale. This helped them to understand what to look for when applying the broader analytic assessment criteria used in the achievement exams.

6.4.2 Student Self-Regulated Learning

In terms of student self-regulated learning, some qualitative and quantitative findings suggested that the scale-assisted self-assessment helped promote the students' self-regulated learning. As revealed from qualitative findings, the students employed a number of self-assessment strategies to engage with the self-writing and self-assessment process, with almost half of them feeling motivated to write and learn as a result of the self-assessment. This finding is in line with previous research showing that rubric-assisted self-assessment has positive effects on EFL students' writing quality, learning strategies, and attitudes (Kim, 2019).

Another effect of the self-assessment on the students' self-regulated learning could be inferred from the finding that the students compared their self-rating results with teachers' diagnostic results or feedback as one student (31H) reported. The process of comparing self-assessment with external sources or teacher feedback is considered as necessary in promoting self-regulated learning (Andrade, 2019; Andrade & Heritage, 2018) and enhancing the effectiveness of diagnostic feedback on learning (Alderson et al., 2015; Kunnan & Jang, 2009; Lee, 2015). Without self-assessment, such a process may not have taken place. This finding strengthens the significance and necessity of including self-assessment as one key component of diagnostic language assessment (Alderson et al., 2015; Lee, 2015) and other forms of classroom assessment (Brown et al., 2015).

In addition, evidence of the students' self-regulated learning could be drawn from the descriptive and ANOVA results showing that the students were able to diagnose their own ability to a certain extent although they, irrespective of ability levels, are inclined to overestimate their own ability vis-à-vis the teacher-led assessment. This tendency is in line with a body of previous research revealing that student self-assessors tended to self-rate

their language ability more leniently than their peers or teachers (Esfandiari & Myford, 2013; Ünaldı, 2016). However, Matsuno (2009) found that EFL Japanese students tended to underestimate their own writing performances in comparison to teachers and higher achieving students were not more severe than lower-achieving students.

Additionally, correlation results indicate that the student self-assessment was generally somewhat consistent with the teacher-assessment with correlations ranging from weak to moderate. This finding corroborates those of previous systematic review studies (Andrade, 2019; Brown & Harris, 2013) reporting that small-to-moderate correlations between student self-ratings and teacher ratings were the norm. On closer investigation, the high-achieving students' ratings showed better correlations with the teachers' assessments than the mid- and low-achieving students' ratings. In fact, the low-achieving students' self-assessments demonstrated negative correlations with the teachers' ratings. This implies that the higher-ability students were better able to self-assess their essays than the lower-ability students, and supports the conclusions of existing systematic review studies (Andrade, 2019; Baleghizadeh & Hajizadeh, 2014; Brown & Harris, 2013) and empirical studies (Brown & Harris, 2013; Ünaldı, 2016).

It was also observed that the high-and mid-ability students generally showed the highest self-rating consistency on the task for which they received the highest scores and exhibited the lowest rating consistency on the task for which they gained the lowest score. This was not true for the low-ability students, who exhibited a decrease in self-rating consistency over the tasks. It may be that the students' self-rating consistency is influenced by the levels of task difficulty as flagged in meta-analysis research (Brown & Harris, 2013). However, it is not clear in this study whether the students' scores over the tasks are predominantly reflective of task difficulty or of student progress, let alone changes in the raters' severity over the tasks.

Inaccuracy, either overestimation or underestimation, and low consistency undermine the reliability and validity of self-assessment scores (Brown et al., 2015). The variability between students' self-assessment and teachers' judgement could be caused by, for instance, student inability to apply assessment criteria, self-bias, and even the teachers' unreliable assessments, which are factors also noted by Ross (2006). A lack of accuracy in self-assessment is not a matter of great concern in the current study context where the main goal of the self-assessment was to enhance motivation, engagement, and

self-regulation. In fact, focusing too much on the quality of student self-assessment could undermine its effectiveness for formative purposes, where the internal process of self-assessment, rather than its accuracy, is of critical importance to cultivate students' self-regulated learning (Brown et al., 2015; Harris & Brown, 2018). Indeed, Ross (2006) suggested that differences between self-assessment and teacher-led assessment can lead to productive impact in terms of teacher-student conversations about student learning needs.

All in all, in spite of showing poorer self-assessment quality than the teachers' ratings, the students generally demonstrated as reasonable a degree of rating accuracy and consistency as could be expected from students given trends reported in previous studies. This suggests that they were engaged and attentive in the self-assessment process and hence self-regulating learning.

6.4.3 Student Learning Progression

It could be concluded from some findings that the formative diagnostic assessment helped improve the students' learning in the classroom. Qualitative findings suggested that the teachers and students discerned writing improvement over the course. It is also worth noting that the students' comments about their writing progress were not always consistent with the patterns of their overall diagnostic scores on the first-draft essays over the tasks. However, it appears from some teacher commentary (Ivey) that this progress was observed on the second drafts which were not rated using the diagnostic scale.

In addition, descriptive and MFRM results, based on the diagnosis of the first-draft essays over the three sequential assignment tasks, could not be used to draw sound conclusions about the student learning progression as it is not clear in the current assessment circumstance whether changes in the student scores over the tasks resulted from writing improvement and/or the task variability (see Appendix K). Although the students were assigned the same genre on the same task, some teachers allowed students to choose their own topic, some allowed students to choose one topic from the provided topics, still others assigned the same topic for all students. Clearly then, the topics, and prompts on the same genre were quite different. Moreover, the order of the assigned tasks might affect the patterns of learning improvement if the genre and task

characteristics are not equivalent in difficulty. That is, if the genre and task characteristics on the third assignment are more difficult than the first and second ones, then the student improvement curve could be suppressed due to the effects of the task difficulty and ordering. In fact, the third tasks included compare-contrast and argumentative essays. In particular, argumentative writing is considered as complex and challenging in academic writing (Ahmad, 2019). It is thus possible that the third tasks were more demanding for the students than the first and second ones. Variability and complexity in genres, topics, and task characteristics have different effects on test-takers' language performances as revealed in previous studies (e.g., Cho, 2008; Huang, 2009; Jeong, 2017; Jiuliang, 2014). However, it is unclear in this study whether the difficulty levels of the tasks are equivalent or different and to what extent the variability and difficulty of the genres and tasks have different effects on individual students' writing performances. Since task variability is typical in a formative classroom assessment, it is difficult to standardise assessment tasks and procedures to minimise sources of measurement errors and therefore individual students' learning and assessment results are influenced by a variety of factors (Andrade & Heritage, 2018; Kane & Wools, 2020; Moss, 2016). As this study was not appropriately designed to trace student progress and aimed at designing the writing assignment tasks, it was impossible to manipulate and control the effects of task and genre characteristics and to track student writing improvement.

Also, it is very important to note that the non-standardised and rather varied nature of the current formative assessment limits the meaningful interpretation of the students' learning progress based on the quantitative scores obtained from the first-draft essays. The students' scores might or might not change on the second-draft essays, might vary due to the high variability of the assignment tasks, and might be influenced by varying writing conditions outside of the classroom. What is more, there might be some degree of plagiarism involved in some essays that the teachers could not detect and the students' engagement and effort on individual tasks might change over time. Clearly, there is a variety of variables and factors underlying the students' assignment scores, which limits the power of the present measurement-driven diagnostic assessment in detecting the students' learning progression. This does not mean, however, that the students did not improve on their writing over the course. The students' learning progression could be inferred from other methods of assessment and various sources of

evidence (e.g., second-draft essay diagnosis and teacher conversation with students) during ongoing teaching and learning as noted by Andrade and Heritage (2018), Kane and Wools (2020), and Moss (2016). However, within the scope of this study, these routine and real-time sources of learning evidence could not be collected.

6.4.4 Student Learning Achievement

In respect of student learning improvement, it can be inferred partly from the quantitative results that the scale-driven formative diagnostic assessment contributed to the students' learning achievement, which is one of the focal purposes of formative assessment (Andrade & Heritage, 2018). This is suggested by the regression results showing that both self- and teacher-assessment results were significantly correlated with student achievement outcomes. The present findings appear to corroborate previous research showing a positive effect of formative assessment on EFL students' learning achievement (Asadifard & Afghari, 2019) and a positive relationship between self-assessment and learning achievement (Andrade, 2019). However, it is important to bear in mind that the contribution of the formative diagnostic assessment to the student achievement might not have resulted primarily from the use of the scale per se but could be influenced by a variety of factors (e.g., teacher feedback and teaching methods) in the current Thai EFL classroom context.

6.4.5 Assessment Impact

With respect to formative diagnostic assessment impact, qualitative findings revealed that the diagnostic rating scale and the assessment had some positive consequences on the teachers and students. That is, the teachers and students acknowledged that the student self-assessment was useful for improving their essay writing skills and it should be included in future writing classrooms. Following the implementation of the scale and formative diagnostic assessment, all teachers were interested in adopting or adapting the scale for future teaching and assessment. Some were aware of the importance of teachers' feedback and students' self-feedback in writing development. In addition, using the diagnostic scale made some teachers more conscious of the importance of transparency in assigning scores so that assessment results appeared fair to the students. Some also came up with new ideas to improve future teaching and

conduct future research. All this suggests the positive washback of the scale and assessment in the current Thai EFL classroom context.

In summary, the quantitative and qualitative findings revealed that the formative diagnostic assessment process generally served its intended beneficial consequences on teaching and learning to the extent that was possible to gauge from the current study. This corroborates previous studies discovering that incorporating a formative assessment in EFL classrooms has positive impacts on writing teaching and learning (Lee, 2011), students' self-regulated learning (Jing, 2017; Xiao & Yang, 2019), students' writing ability (Mohamadi, 2018; Naghdipour, 2017), and students' academic achievement (Asadifard & Afghari, 2019). While the quantitative results do not provide a reliable source of evidence for student learning due to the lack of standardisation across tasks and rather varied nature of the assessment, evidence from qualitative findings shows some promising trends.

6.5 Development of the Validity Argument

Now that all of the quantitative and qualitative findings have been synthesised and discussed, this section integrates all the evidential sources across the three study stages, including research procedures and empirical quantitative and qualitative findings. Different sources of evidence may more or less support (✓), threaten (X), or question (?) assumptions underlying each of the IUA inferences. The evidentiary sources are aligned with the assumptions for the warrant of each inference and are evaluated to determine the plausibility of the warrant. The evidence across the inferences is then evaluated to determine the coherence and completeness of the IUA in order to establish the overarching validity argument for the scale-driven formative diagnostic assessment. It is very important to keep in mind that as validity of the proposed interpretation and use of the scale was contextualised in the Thai EFL university writing classroom of interest. Therefore, it should not be generalised to other contexts.

6.5.1 Evidence Justifying the Domain Description Inference

Table 6.1 summarises all evidentiary sources for the domain description inference resting on the warrant that the scale criteria represent academic writing ability and skills in student writing performances and learning contents in the target language use (TLU)

domain of EFL university classroom. This warrant depends on two assumptions: *(1) the expected academic writing quality features, writing skills, and learning contents in the classroom can be identified, and (2) the characteristics of writing assignment tasks in the classroom can be identified.* Both assumptions are justified by evidence related to research procedures and classroom practices delineated in Chapter 3.

Table 6. 1 *Evidence for Backing of the Assumptions for the Domain Description Inference*

Evidence	Study stage	Assumptions	
		1	2
• Review of L2 writing ability theories and existing scales	1	✓	-
• Review of classroom-related materials	1	✓	✓
• Expert review and preliminary trialling of the scale	1, 2	✓	-
• Teacher review and trialling of the scale	2	-	-
• Teacher discussion about classroom practices	2	✓	✓
• Variability of classroom materials and assignments	3	-	X

As presented in the table, the first assumption is reasonably supported due to the evidence that the scale criteria were developed and revised based on a multisource review of theoretical, intuitive, and empirical sources during the scale construction and trialling stages. In so doing, a representative sample of writing quality features, writing skills, and learning contents were reductively extracted from L2 writing ability theories and existing scales, empirical classroom materials and student writing performances, and expert and teacher intuitive feedback and comments.

The second assumption is partly substantiated owing to the evidence that the characteristics of writing assignment tasks were identified through a review of prior classroom materials and student writing performances in the scale construction stage and teacher discussion about the prospective characteristic of classroom assignment tasks in the scale trialling stage. However, the second assumption is partly undermined by the evidence that the teacher-made assignment tasks in the scale implementation stage were understandably rather varied and different from those identified in the scale construction and trialling stages. In the context, the writing courses may be conducted by the same teachers or different teachers and teachers may more or less change or alter learning materials and assessment tasks over time as they see fit to accommodate their students'

needs and characteristics at the time. Accordingly, writing assignment instructions and prompts, albeit similar genres, may vary more or less each year.

Overall, despite some unsupported evidence stemming from the variability of the teacher-made assignment tasks in the classroom context, there is sufficiently sound evidence to support the assumption that the diagnostic criteria represent the academic writing ability and skills, and the learning contents in the Thai EFL university writing classroom. As sufficient supportive evidence has been provided to substantiate the warrant of the domain description inference, it is possible to move on to the evaluation inference.

6.5.2 Evidence Justifying the Evaluation Inference

Table 6.2 presents all evidence for justifying the evaluation inference which claims that the scale provides observed scores reflective of the academic writing ability and skills in student writing performances in the classroom. This warrant depends on five assumptions: *(1) the rating format is appropriate to assess the strengths and weaknesses of the student writing ability, (2) the scale shows acceptable psychometric properties to ensure accurate functioning, (3) the raters are positive about the scale functioning, (4) the raters go through appropriate rater training and rating procedures, and (5) the raters show acceptable psychometric properties to ensure appropriate rating behaviours*.

As shown in the table, the first assumption is reasonably supported in that the rating format was informed by a conceptual review of scale development, the expert and teacher review and trialling of the scale in the scale construction and trialling stages. Nonetheless, the feasibility of this assumption is called into some question a few negative perceptions of the capacity of binary rating to capture the granularity of strengths and weaknesses on writing skills as discussed in the RQ1 results.

Despite the negative PTM correlation of one descriptor, the second assumption is mostly supported by the MFRM, CTT, and descriptive indicators discussed in the RQ1 results, confirming the appropriate psychometric functions of the scale at the scale implementation stage. The plausibility of third assumption, though slightly undermined by a few negative perceptions of the scale functioning, is reasonably supported by the generally positive perceptions of the scale functioning as discussed in the RQ1 results.

The fourth assumption, considering the non-standardised nature of formative classroom assessment, is largely substantiated in that the raters received appropriate training, practice, and opportunity to pilot-rate the scale with student essays during the scale trialling and implementation stages. The non-standardised nature of the assessment could negatively affect the teachers' rating consistency. The fifth assumption that is largely substantiated by the MFRM, CTT, and descriptive indicators at the scale implementation as discussed in the RQ1 results, confirming the raters' appropriate behaviours in applying the scale.

Table 6. 2 *Evidence for Backing of the Assumptions for the Evaluation Inference*

Evidence	Study stage	Assumptions				
		1	2	3	4	5
• Conceptual review of scale development	1	✓	-	-	-	-
• Expert intuitive review and pre-trialling of the scale	1, 2	✓	-	-	-	-
• Teacher intuitive review and trialling of the scale	2	✓	-	-	-	-
• Rater training and practice with support documents	2	-	-	-	✓	-
• Scale trialling on student essay samples	2	-	-	-	✓	-
• Low standardisation of the classroom assessment	3	-	-	-	X	-
• Acceptable data-model fit	3	-	✓	-	-	-
• Acceptable unidimensionality of the scale	3	-	✓	-	-	-
• Acceptable local independence of the scale	3	-	✓	-	-	-
• Wide visual dispersion of descriptor logits	3	-	✓	-	-	-
• Wide visual dispersion of student logits	3	-	✓	-	-	✓
• Significant heterogeneity index of descriptor logits	3	-	✓	-	-	-
• Significant heterogeneity index of student logits	3	-	✓	-	-	✓
• High separation indices of descriptor logits	3	-	✓	-	-	-
• High separation indices of student logits	3	-	✓	-	-	✓
• Noticeable range and SD of descriptor logits	3	-	✓	-	-	-
• Noticeable range and SD of student logits	3	-	✓	-	-	✓
• Acceptable fit indices of all descriptors	3	-	✓	-	-	-
• Acceptable fit indices of all students	3	-	✓	-	-	✓
• Acceptable fit indices of all raters	3	-	-	-	-	✓
• Acceptable independence indices of raters	3	-	-	-	-	✓
• Acceptable PTM correlations of descriptors	3	-	✓	-	-	-
• Unacceptable PTM correlation of one descriptor	3	-	?	-	-	-
• High alpha internal consistency reliability of the scale	3	-	✓	-	-	-
• Acceptable CIT correlations of all descriptors	3	-	✓	-	-	-
• Noticeable range and SD of student observed scores	3	-	✓	-	-	-
• Positive perceptions of scale functioning	3	-	-	✓	-	-
• Negative perceptions of scale functioning	3	-	-	?	-	-
• Negative perceptions of binary rating	3	?	-	-	-	-

Overall, although the variability of assessment tasks, the unacceptable PTM correlation of one descriptor, and a few negative perceptions of the scale functioning could detract from the quality of the observed scores, there is sufficient supportive evidence to justify the conclusion that the scale provides accurate observed scores. It is thus reasonable to proceed to the generalisation inference.

6.5.3 Evidence Justifying the Generalisation Inference

Table 6.3 summarises all evidentiary sources pertaining to the generalisation inference claiming that the scale provides observed scores as estimates of the expected scores across raters and student writing performances in the classroom. This warrant depends on two assumptions: *(1) the scale shows acceptable psychometric properties to ensure consistent functioning, and (2) the raters show acceptable psychometric properties to ensure consistent rating behaviour*. Both assumptions are justified by quantitative findings obtained at the scale implementation stage as discussed in the RQ2 results.

Table 6. 3 *Evidence for Backing of the Assumptions for the Generalisation Inference*

Evidence	Study stage	Assumptions	
		1	2
• Wide visual dispersion of rater logits	3	-	X
• Significant heterogeneity index of rater logits	3	-	X
• High separation reliability of descriptor logits	3	✓	-
• High separation reliability of student logits	3	✓	-
• High separation reliability of rater logits	3	-	X
• Acceptable fit indices of all descriptors	3	✓	✓
• Acceptable fit indices of all students	3	✓	✓
• Acceptable fit indices of all raters	3	✓	✓
• Noticeable range and SD of rater logits	3	-	X
• High percent observed exact agreement of raters	3	-	✓
• Acceptable percent interrater agreement of most descriptors	3	✓	✓
• Unacceptable percent interrater agreement of six descriptors	3	X	X
• Acceptable percent interrater agreement of most students	3	-	✓
• Unacceptable percent interrater agreement of six students	3	-	X
• High alpha internal consistency reliability of the scale	3	✓	-
• Noticeable range and SD of rater observed ratings	3	-	X

As outlined in the table, the first assumption, albeit threatened by six descriptors with unacceptably low percent interrater agreement, is reasonably supported by most

MFRM and CTT indicators confirming the scale consistency across the raters and student writing performances. While partly threatened by some MFRM, CTT, and descriptive indicators showing the substantial variability of the rater severity, the second assumption is substantiated by some MFRM and CTT indices revealing the self-consistency of individual raters' behaviours.

Overall, considering that the raters were using the scale for the first time and the formal training they received may not have been sufficient to homogenise their interpretation of the descriptors. In any case, rater heterogeneity is not unusual in classroom assessment circumstances. There is arguably, therefore, sufficient evidence to support the generalisation inference in a classroom context such as this where the stakes are relatively low and there are multiple opportunities for teachers to refine their judgements. Accordingly, it is sensible to proceed from the generalisation inference to the explanation inference.

6.5.4 Evidence Justifying the Explanation Inference

Table 6.4 shows all evidentiary sources in justification of the explanation inference stating that the scale provides observed scores as estimates of the expected scores attributed to the defined academic writing construct required in the classroom. This warrant relies on following two assumptions: *(1) the diagnostic scores are internally consistent with the defined writing construct, and (2) the diagnostic scores reflect the academic writing skills learned and assessed in the classroom*. Both assumptions are justified by the domain description inference evidence and quantitative and qualitative findings achieved at the scale implementation stage as discussed in the RQ1 results.

In spite of one descriptor with negative PTM correlation, the first assumption is largely supported by MFRM and CTT indicators confirming the unidimensionality and internal consistency of the descriptors in capturing the defined writing construct. The second assumption is partly substantiated by supportive evidence for the domain description inference, and overall positive perceptions of the criteria comprehensiveness. However, this assumption is slightly weakened by the unsupportive evidence of the domain description inference and a few negative perceptions of the criteria coverage of the writing construct.

Overall, notwithstanding some unsupportive evidence, most of the evidential sources provide sufficiently sound evidence to back the warrant of the explanation inference. Accordingly, it is feasible to move on to justify the extrapolation inference.

Table 6. 4 *Evidence for Backing of the Assumptions for the Explanation Inference*

Evidence	Study stage	Assumptions	
		1	2
• Acceptable data-model fit	3	✓	-
• Acceptable unidimensionality of the scale	3	✓	-
• Wide visual dispersion of student logits	3	✓	-
• Significant heterogeneity index of student logits	3	✓	-
• High separation indices of student logits	3	✓	-
• Acceptable fit indices of all descriptors	3	✓	-
• Acceptable PTM correlations of almost all descriptors	3	✓	-
• Unacceptable PTM correlation of one descriptor	3	?	-
• Acceptable CIT correlation of all descriptors	3	✓	-
• High alpha internal consistency reliability of the scale	3	✓	-
• Supportive evidence for domain description inference	1, 2	-	✓
• Unsupportive evidence for domain description inference	2, 3	-	X
• Positive perceptions of scale comprehensiveness	3	-	✓
• Negative perceptions of scale comprehensiveness	3	-	X

6.5.5 Evidence Justifying the Extrapolation Inference

Table 6.5 presents all the evidentiary sources for justifying the extrapolation inference claiming that the scale provides diagnostic scores accounting for the quality of student academic writing ability on other tasks in the classroom. This warrant rests on two assumptions: *(1) the diagnostic results distinguish between low-, mid-, and high achieving students, and (2) the diagnostic results have a positive relationship with student learning achievement*. The two assumptions are justified by quantitative findings obtained at the scale implementation stage as discussed in the RQ1 results.

As outlined in the table, the first assumption is reasonably substantiated by (a) the visual variable map showing the students' locations and (b) the ANOVA results showing significant differences in diagnostic results between high-, mid-, and low-achieving students, grouped according to their total achievement exam results. The second assumption is upheld by correlation results indicating significant, positive, and strong correlations between the student diagnostic results and the student achievement results. However, there is partial evidence from the teachers' perceptions that the use of the scale

influenced their assessment of the student exam essays. This casts some doubt on the interdependence between the formative and summative assessments and in turn the assumptions for the extrapolation inference. However, given that the writing tasks, rating criteria, and writing conditions between the two assessments were different, the influence of the scale utilisation on the teachers' exam judgement should not be deemed as a significant threat to this inference.

Table 6. 5 *Evidence for Backing of the Assumptions for the Extrapolation Inference*

Evidence	Study Stage	Assumptions	
		1	2
• Student locations on the visual variable map	3	✓	-
• ANOVA significant differences in diagnostic results between high-, mid-, and low-achieving students	3	✓	-
• Significant, positive, and strong correlation between diagnostic results and exam percentages	3	-	✓
• Dependency between formative and summative assessments	3	?	?

Overall, ANOVA and correlation results provide partial but adequate evidence to claim that the scale provides diagnostic scores which can be extrapolated to the quality of the students' academic writing ability on other writing tasks in the classroom. Even so, more investigations are required to gain stronger evidence for strengthening the extrapolation inference. As the warrant for the extrapolation inference is apparently upheld, it is reasonable to proceed to the decision inference.

6.5.6 Evidence Justifying the Decision Inference

Table 6.6 summarises all evidentiary sources relevant to the decision inference resting on the warrant that the scale is useful to support formative decisions about teaching and learning in the classroom. This warrant relies on three assumptions: *(1) the scale is practical for teachers and students in the ongoing classroom, (2) the scale provides diagnostic information meaningfully interpretable by teachers and students, and (3) the scale provides useful diagnostic information to inform teachers' formative decisions about instruction and learning.* All assumptions are justified by qualitative findings that emerged at the scale implementation stage as discussed in the RQ3 results.

Though slightly weakened by the teacher concern about the scale length, the first assumption is largely supported by the overall positive perceptions of the scale practicality. The second assumption is generally substantiated by the overall positive perceptions of the diagnostic score interpretation. However, the concern about the lack of a diagnostic profile report to accompany each student's scores could be seen as partly threatening the meaningfulness of the diagnostic information yielded by the scale. The third assumption is reasonably well substantiated by the overall positive perceptions of the scale in informing the teacher formative decisions. In spite of the teacher concern about the value of the scale in supporting low-ability students' formative decisions, the fourth assumption is reasonably supported by the overall positive responses.

Table 6. 6 *Evidence for Backing of the Assumptions for the Decision Inference*

Evidence	Study stage	Assumptions		
		1	2	3
• Positive perceptions of scale practicality	3	✓	-	-
• Negative perceptions of scale practicality	3	X	-	-
• Positive perceptions of diagnostic information	3	-	✓	-
• Negative perceptions of diagnostic information	3	-	X	-
• Positive perceptions of the scale usefulness	3	-	-	✓
• Negative perceptions of the scale usefulness for low-ability students	3	-	-	X

Overall, leaving aside some negative perceptions about the scale length, the scale interpretation, and the lack of diagnostic profile report, most of the feedback from the teachers and students support the claim that the scale provides useful diagnostic scores to inform formative decisions about teaching and learning. Accordingly, it is reasonable to move to the consequence inference which is central to the validity of the formative diagnostic assessment.

6.5.7 Evidence Justifying the Consequence Inference

Table 6.7 lays out all evidentiary sources for justifying the consequence inference claiming that the scale-driven assessment has beneficial consequences for teaching and learning in the classroom. This warrant relies on five assumptions: (*1) the scale provides diagnostic information to improve teacher instruction and feedback, (2) the scale supports self-assessment in promoting student self-regulated learning, (3) the assessment system*

promotes student learning progression, (4) the assessment system contributes to student learning achievement, and (5) the assessment system has potential positive impacts on teachers' and students' academic development. All assumptions are justified by the empirical findings at the scale implementation stage as discussed in the RQ4 results.

Table 6. 7 *Evidence for Backing of the Assumptions for the Consequence Inference*

Evidence	Study stage	Assumptions				
		1	2	3	4	5
• Positive perceptions of teaching and feedback improvement	3	✓	-	-	-	-
• Negative perceptions of binary rating	3	X	-	-	-	-
• Overall significant differences in self-assessment scores between high-, mid-, and low-achieving students	3	-	✓	-	-	-
• Significant score differences in diagnostic scores between self-assessment and teacher-assessment	3	-	✓	-	-	-
• Some positive correlations between self-assessment and teacher-assessment	3	-	✓	-	-	-
• Student use of relevant self-assessment strategies	3	-	✓	-	-	-
• Positive perceptions of learning engagement/motivation	3	-	✓	-	-	-
• Negative perceptions of learning motivation	3	-	X		-	-
• Positive perceptions of writing improvement	3	-	-	✓	-	-
• No quantitative results provide meaningful indication of learning progression	3	-	-	?	-	-
• Significant prediction of formative diagnostic and self-assessment results on student achievement.	3	-	-	-	✓	-
• Positive impact on teacher professional development	3	-	-	-	-	✓
• Positive impacts on student academic development	3	-	-	-	-	✓

The first assumption is reasonably supported in that the scale was largely perceived as enhancing the teacher instruction and feedback even though the binary rating was deemed as providing crude diagnostic information on the quality of writing skills. Although some students did not feel the self-assessment motivated them to learn and write, the second assumption is supported to a certain extent by the following evidence. That is, the students generally (a) showed a reasonable degree of self-rating consistency and accuracy with the teachers' ratings based on ANOVA and correlation results, (b) used strategies in writing, self-rating, and revising their essays during self-assessment process, and (c) expressed their engagement and motivation in learning and self-assessment.

The third assumption is supported to a certain degree by the positive feedback with respect to the student writing improvement. Yet as a consequence of the rather varied assignment tasks, it is not logical to draw conclusions about student learning progression based on quantitative results. Having said that, this lack of evidence should not be deemed a failure of the assessment in promoting the student learning progression or a threat to this inference.

The fourth assumption is supported to a reasonable extent by regression results suggesting that the formative assessment and self-assessment significantly predicted and accounted for the student learning achievement as assessed by the midterm and final exams. In other words, the assessment successfully contributed, at least in part, to the students' overall learning achievement on their writing course.

The fifth assumption is supported by the teachers and students' reflections. To elaborate, the utilisation of the assessment system raised the teacher awareness of the importance of fair assessment and the usefulness of feedback and self-assessment, ignited the teacher ideas for future teaching and research, and kindled the teacher interest in using the scale for future teaching and assessment. The students also realised the usefulness of self-assessment in learning to write.

Overall, there are sound, albeit partial sources of evidence, to support the intended beneficial consequences of the scale-driven assessment on teaching and learning. As noted above it was the qualitative feedback from teachers and students rather than the quantitative data which lent more convincing support in relation to the consequence inference, but even the qualitative evidence was somewhat limited in scope and could usefully be built on in future research. More evidence is necessarily called for to reasonably ensure that the scale-driven assessment satisfies its intended purposes in improving teaching and learning in general and learning progression in particular.

6.6 Challenges in the Current Classroom Assessment Validation

As with several studies on L2 classroom assessment validation (Chapelle, Chung, et al., 2010; Chapelle et al., 2015; Ranalli et al., 2017), Kane's argument-based approach, albeit framed initially for test-based, standardised, and high-stakes assessment (Kane & Wools, 2020; Moss, 2003, 2013, 2016), was found to be generally viable for the current formative diagnostic assessment in the classroom. There are, nevertheless, certain

challenges worthy of note in the validation of the current Thai EFL university writing classroom assessment due to its non-standardised, multifaceted, dynamic, and varied nature.

To begin with, as the scale-driven diagnostic assessment was integrated into ongoing classrooms for promoting teaching and learning and focused on diagnosing writing assignment tasks, the assessment was thus rather non-standardised. Consequently, the diagnostic scores were inevitably influenced by desirable and undesirable sources of measurement error in the context. In this circumstance, as already noted above, individual students' writing performances and individual teachers' rating behaviours could be influenced by varying uncontrollable factors. That limited the utilisation of the diagnostic scores over the tasks to inform learning progression. It is well acknowledged that formative classroom assessment is non-standardised and varied in nature (Andrade & Heritage, 2018; Kane & Wools, 2020; Moss, 2003, 2013, 2016) and thus imposing standardisation during ongoing learning and assessment is difficult and not pedagogically practical and sound for individuals' learning (Moss, 2003). Accordingly, the diagnostic scores generated in the current assessment are not sufficient to reflect learning progress, which is another source of validity evidence for the consequence inference.

In addition to the score-based diagnostic information, there was a variety of sources that could reveal the students' strengths and weaknesses and shape the students' learning progression during regular teaching and learning activities. As noted by Andrade and Heritage (2018), Kane and Wools (2020), and Moss (2016), evidence of learning, the focus of validity in classroom assessment, is not only inferred psychometrically from observed changes in individual students' abstract knowledge or assignment performances over time but also inferred from various sources, including *(a) students' engagement with learning and the tasks through successive drafts, (b) students' ongoing interactions about their work with other students and teachers, (c) teachers' ongoing conversations with students and colleague, (d) teachers' observations of students' interactions with others, and (e) reading materials students are assigned and locate on their own, and formal feedback from teachers*. Any change in one of these sources may affect the nature of assessment for a particular student, a particular group of students, or the entire classroom (Moss, 2003). In this regard, the use of varying methods, in addition to psychometric analyses, are needed to investigate and collect evidence of learning emerging during regular

learning (Andrade & Heritage, 2018; Moss, 2016). For example, Can Daşkın and Hatipoğlu (2019) drew on a conversation analysis of teacher-student interactions in an informal EFL formative classroom assessment to examine teachers' informal assessment of student knowledge and understanding. Their findings showed that teachers can seek evidence of student learning through informal assessment during interaction with students. This way of investigating the validity of classroom assessment is consistent with a functional (Cronbach, 1988; Kane & Wools, 2020) or conceptual (Moss, 2016; Murnane et al., 2009) perspective to validation. An investigation of the validity of classroom assessment that focuses on the interpretations and uses of score-based information is in line with a measurement (Cronbach, 1988; Kane & Wools, 2020) or instrumental (Moss, 2016; Murnane et al., 2009) perspective to validation. It is recommended that both perspectives should be complementarily used in validation of classroom assessments (Kane & Wools, 2020; Moss, 2016). In particular, attention should be centred on a functional perspective which can more effectively illuminate students' learning while a measurement perspective, given its limitations, should play a supporting role (Kane & Wools, 2020). That being said, it was impossible for this study to document various sources of learning evidence emerging from regular classrooms since the researcher did not directly teach, observe, and interact with the students in the classroom.

Furthermore, the current assessment was administered on multiple sequential occasions to provide ongoing information to inform instructional decisions and interpretations about the students' strengths and weaknesses as well as cumulative learning over time. In this manner, the interpretations and uses of the diagnostic results and the decisions and actions informed by the diagnostic information were regularly made and perhaps altered by the local teachers in order to adjust upcoming teaching and learning activities. Although the current proposed interpretations and uses were modified over time, they were broadly predetermined from the perspective of the scale developer or researcher rather than the teacher and student users who directly interpreted and used the diagnostic results on a regular basis. For meaningful interpretations and uses of an ongoing classroom assessment, it is suggested that relevant stakeholders be involved in the specification and modification of the interpretations and uses of assessment results (Kane & Wools, 2020).

Finally, learning sources and assignment tasks in the classroom under study were rather varied, thereby posing a threat to the assessment validity particularly in respect of the domain description inference. Even though an analysis of the TLU domain was carried out to inform the scale development, the learning materials and assignment tasks used in the actual classrooms were rather different across the writing classrooms and from those specified in the TLU domain analysis and used in the previous classrooms. This was probably due partly to the fact that the intact classrooms were conducted by different teachers with varying background and expertise and thus they might have different perceptions of how assignment tasks should be designed to suit students' needs. Chapelle and Voss (2014) acknowledged this challenge of defining the relevant domain in classroom assessments. They showed the case of a relatively low-stakes achievement test for a reading classroom, where learning materials were changed on a regular basis. In this scenario, test developers analysed coursebooks on reading and vocabulary development to identify a pool of words to be tested but an online learning source subsequently replaced the coursebooks, thereby making it difficult to define the domain of actual words that students studied. Precisely for this reason, they pointed out that it is necessary to redesign a test when a TLU domain changes to ensure that test scores represent the learning objectives across time. Im et al. (2019) also suggested that during a domain analysis, researchers should work with language users and relevant stakeholders in a local assessment context to identify language knowledge, skills, and abilities and relevant tasks in the TLU domain. All this implies that a TLU domain analysis and test development in a classroom assessment tends to be a mutually-informed and continual process in the sense that when there are changes in, for example, learning materials, tasks, and teaching activities, in a TLU domain, a test should be redesigned accordingly.

6.7 Chapter Summary

In this chapter, the research findings, together with other sources of evidence have been synthesised to support the seven inferences as the basis for the establishment of the overarching validity argument for the newly-developed diagnostic scale. Overall, the findings confirm that the scale functions appropriately and consistently and supports teaching and learning in the current assessment context. In addition, theoretical, procedural, and empirical evidence has reasonably substantiated the assumptions

underlying the domain description, evaluation, generalisation, explanation, extrapolation, and decision inferences. Yet, it has been argued that more evidence, apart from a measurement-driven assessment approach, is needed to support the decision and particularly consequence inferences. The current IUA is driven primarily by the scale-based information which could not capture learning progression that can be inferred from various sources of information. Moreover, some aspects of the ongoing learning and teaching, for instance, high variability in learning tasks and resources and dynamic and multifaceted learning development, pose a challenge to the applicability of the argument-based approach in the current formative classroom assessment. In the next chapter, the conclusion of the current scale development and validation research will be presented, along with the study implications, limitations, and suggestions for future research.

Chapter 7: Conclusion

This research set out with two primary objectives: (1) to develop a diagnostic rating scale for a formative diagnostic assessment for the purpose of diagnosing students' strengths and weaknesses in academic English essays in an ongoing EFL university classroom in Thailand, and (2) to examine the validity of the scale situated within a formative classroom assessment following Kane's argument-based approach to validation. To achieve these objectives, four research questions were formulated to investigate the appropriateness and consistency of the scale functioning as well as the usefulness and consequence of the scale utilisation. A three-stage exploratory sequential mixed-methods research design was employed as the research methodology to address the research questions and objectives mentioned above. The scale was designed and constructed based on multiple sources of information at the scale construction stage and was subsequently trialled and modified at the scale trialling stage. Finally, the scale was operationalised at the scale implementation stage, where empirical data were accumulated from a three-round diagnosis of students' first-draft assignment essays and retrospective semi-structured interviews of teachers' and students' perceptions of the scale with a view to answering the research questions. The diagnostic scores were analysed using Classical Test Theory (CTT) and Many-Facets Rasch model (MFRM) psychometrics and descriptive, ANOVA, correlation, and regression statistics, whereas the perception protocols were analysed following a qualitative content analytic approach.

7.1 Summary of Research Findings and Validity Argument

Despite certain limitations and challenges that will later be highlighted, this research has satisfactorily accomplished its two primary objectives. Overall, it can be argued that the quantitative and qualitative findings ascertain the appropriateness, consistency, usefulness, and positive consequence of the scale and offer reasonable backing for the overarching validity argument for the scale.

To elaborate, the scale development procedures were informed by multiple sources and systematically conducted, hence supporting the domain description inference. The quantitative and qualitative findings related to the first research question

revealed that the scale: (1) provided accurate diagnostic scores, (2) homogeneously captured the prime dimension and substantial variability of the writing construct, (3) yielded diagnostic results well aligned with the summative achievement outcomes, and (4) was deemed by the raters or users (teachers and students) to be largely comprehensible, comprehensive, and applicable. While the psychometric indicators confirm the scale appropriacy, several issues emerged from the raters' perceptions, including (a) scale comprehension and comprehensiveness, (b) aspects of scale structure, (c) ordering of descriptors, and (d) the requirement to count errors and make binary judgements. These problems may potentially threaten the validity of the scale and assessment. Since the findings suggested the acceptable accuracy of the diagnostic scores, the substantial variability of the defined writing construct captured by the scale, and the correspondence between the scale functioning and the achievement exams, they offer reasonable support for the warrants of the evaluation, explanation, and extrapolation inferences respectively.

As for the second research question, the quantitative results generally indicated that the scale descriptors provided consistent diagnostic scores. Although the raters significantly differed in levels of severity and their diagnostic judgements appeared to vary according to the levels of descriptor difficulty and essay quality, they were acceptably self-consistent and congruent in judging the descriptors and student essays over the three sequential tasks. It is, therefore, reasonable to say that the warrants of the generalisation and explanation inferences are sufficiently supported. As regards the third research question, the qualitative findings showed that the teachers and students generally perceived the scale to be practical and useful in identifying writing strengths and weaknesses. They also viewed the diagnostic information to be useful for their decisions about teaching and learning. These positive perceptions thereby reasonably substantiate the warrant of the decision inference.

The qualitative findings related to the fourth research question partly showed that the scale-driven assessment was deemed to improve the students' writing and learning as well as the teachers' instruction in terms of their provision of feedback, assessment of learning achievement, and teaching preparation. The scale and assessment were also reported to have some positive impacts on the students' and teachers' awareness as well as the teachers' future teaching and professional development. Even though the variable

tasks, formative diagnostic assessment design, and assessment conditions made it difficult for the assessment to gauge learning progress over the course, feedback from the students and teachers suggested that using the instrument made a positive and significant contribution to learning. With respect to the self- assessment, the quantitative analyses revealed that despite variations in student self-assessment between the student ability groups and over the tasks, the student self-assessment showed a reasonable degree of alignment between students' self-ratings and those of their teachers and this was particularly true for high achieving students. In addition, the students reported the use of self-assessment strategies and learning motivation. In combination, these findings offer suggestive evidence for the positive effects of the scale-driven self-assessment on the students' self-regulated learning. It can thus be argued that, the empirical findings, to the extent possible within the constraints of this study, suggested the beneficial consequences of the scale-driven assessment on teaching and learning. It is important to note that, the consequence inference typically requires more time and research to examine and gather evidence to fully back its warrant.

By and large, it can be concluded that the research findings and validity argument generally ascertain the usefulness of the multisource-driven approach for the diagnostic scale development, the generalisation of the scale to the Thai EFL student population and classroom context of interest, and the applicability of the argument-based approach, albeit with certain limitations, for the formative diagnostic assessment in the classroom.

7.2 Implications

The findings of the present study have a number of important implications. The implications could inform the development of theory, pedagogy, and methodology in relation to diagnostic and formative assessments as well as to the teaching and learning of writing in the EFL classroom context.

7.2.1 Theoretical Implications

Theoretical implications are concerned with diagnostic language assessment, rating scale development, and classroom assessment validation.

7.2.1.1 Diagnostic Language Assessment

This study contributes to the theory of diagnostic language assessment (Alderson, 2005; Alderson et al., 2015; Jang, 2012; Knoch, 2007, 2009a, 2009b, 2011; Kunnan & Jang, 2009; Lee, 2015) by providing a fuller picture of the formative process of diagnostic assessment than has been offered in previous research. This has been achieved by exploring the insights generated from repeated applications of a diagnostic scale and considering the value of student self-assessment in the process. The findings of this research showed the positive impact of the inference between the diagnostic scale, self-assessment, and repeated assessment in enhancing teaching and learning in the EFL classroom. A number of key insights can be drawn from the current findings regarding the specificity and practicality of a diagnostic scale, the interpretation and utilisation of diagnostic scores, and the rater diagnostic decision-making behaviour.

One of the insights is that while diagnostic instruments should be designed to be user-friendly, discrete and efficient for teachers to make diagnostic decisions, should be suitable for administration in the classroom, and should generate rich and detailed feedback for students (Alderson et al., 2015; Knoch, 2011; Lee, 2015), there is a potential trade-off between the quality of diagnostic information that can be provided by a diagnostic rating scale and the practicality constraints of conducting ongoing formative assessment in the classroom situation. While multiple rating options and more descriptors can enhance the specificity and comprehensiveness of diagnostic criteria and thus provide fined-grained diagnostic information, this may reduce the practicality of a repeated diagnostic tool.

Another insight is that interpretation of individual students' diagnostic scores in the rather varied formative assessment context needs to be undertaken with care and should be linked to or triangulated with other sources of assessment information. In such a context, individual students' observed scores can be attributed to various uncontrollable sources of measurement error and this affects meaningful interpretation of diagnostic scores regarding learning strength, weakness, and progression. Therefore, scores obtained solely from a single diagnostic tool alone may be insufficient and teachers may need to draw on additional methods of diagnosis, alongside other forms of assessment, to arrive at optimal diagnostic outcomes. As pointed out by Alderson et al. (2015, pp. 258), *"diagnostic assessment itself needs to be situated within the range of other assessment*

practices that might routinely take place in and outside the classroom, and we would emphasise that diagnostic assessment is just one type of assessment that provides useful information for students and teachers."

A final insight is related to teachers' diagnostic rating behaviours. As suggested in the findings of this study, the teachers' diagnostic binary judgements seemed to be influenced by essay quality and descriptor difficulty, showing higher rating variability on low-quality essays and difficult writing skills than on high-quality essays and easy skills. This suggests the need for further attention to developing teachers' diagnostic assessment literacy and diagnostic rating training. As Alderson et al. (2015, pp. 318) highlighted, "*it is not the test that diagnoses, it is the user of the test*." Involving teachers in developing and revising a rating scale from beginning to end may help them to better interpret diagnostic criteria as they gradually become familiar with the criteria during the scale development process. At the scale trialling stage, teachers need to receive appropriate scale trialling and rater training under assessment conditions that represent, to the extent possible, the target actual classroom context. Depending on the time allowed, a representative and varying sample of student writing products should be used for scale trialling and rater training so that teachers have exposure to various essay characteristics. In particular, the trialling and training should be focused on diagnosing difficult skills and poorly-written essays. All this is, however, impractical in the real world, where ongoing classroom contexts are dynamic and variable by nature, making it challenging to manipulate such trialling and training conditions and to obtain a representative sample of student writing performances.

It is very important to note that a diagnostic rating scale, ongoing assessment, and self-assessment are just three key recommended elements of effective diagnostic language assessment. There are still other aspects to be explored and examined in order to shape the theory and practice of effective diagnostic language assessment in a classroom context.

7.2.1.2 Rating Scale Development

In relation to rating scale development, the present study contributes to the recent line of a multisource (also known as hybrid) approach to scale development (e.g., Banerjee et al., 2015; Kim, 2010; Knoch, 2007, 2009b; Montee & Malone, 2014; Wagner, 2015) which

draw on multiple sources of information to inform scale development. Previous scale development approaches tend to rely on one or a few information sources typically from theory, rater, performance, and statistics, which are not sufficient to account for the multifaceted nature of L2 writing and learning construct in the classroom context (Bachman & Palmer, 2010; Cumming, 2016; Hirvela et al., 2016; Knoch, 2011; McNamara, 1996; Weigle, 2002). The current multisource approach not only drew on writing theories, existing scales, and the classroom curriculum but also involved context-external specialists and local classroom teachers in evaluating the diagnostic criteria and trialling them with student performance samples. By doing so, various sources of information input were triangulated to arrive at optimal outcome during the iterative process of scale development. Therefore, it could be argued that the combination of external and context-sensitive sources makes a rating scale both theoretically defensible and richly representative of the hitherto unexplored Thai tertiary classroom context. More information from further applications of the scale in other contexts could be used for curriculum development purposes so that teaching is more sensitive to learner needs. It should also be informed that, the selection of multiple sources of information for rating scale development needs to be driven by assessment purposes, relevant policies, and score uses in a particular context and the impact of the various sources on assessment quality (e.g., score generalisability and rater reliability) also varies depending on scale developers' design choices (Knoch et al., in press).

7.2.1.3 Validation of Classroom Assessments

With regard to validation of classroom assessment, this study adopted Kane's argument-based approach to validating the interpretation and use of the scale scores. The primary sources of empirical evidence collected were based on the three-round teacher-and self-diagnostic assessments of first-draft assignment essays, and on the retrospective interviews of the teachers' and students' perceptions of the scale. Overall, the current findings indicated that the approach could capture important interpretations and uses of the scale-driven diagnostic information and reasonable sources of evidence to justify the interpretations and uses. However, there are some challenges that could minimise the usefulness of the argument-based approach to the validation of the current formative classroom assessment, including (a) the non-standardised nature of writing

assessment tasks, (b) the complex and multi-componential construct of learning, (c) the dynamic interpretations and uses of assessment results, and (d) the variability of learning resources in the ongoing classroom context.

Clearly, the current classroom assessment is more multifaceted, nuanced, and dynamic than high-stakes standardised testing contexts, at which the argument-based approach is particularly aimed. In light of this, the measurement-driven diagnostic assessment was not sufficient to capture the nature of learning problems and development in the classroom. The current validation framework, building on the argument-based approach, was primarily measurement-oriented, too broad and somewhat static, and still underrepresented other interpretations and uses of assessment results that the teachers and students made on a regular basis as well as other sources of learning evidence emerging during ongoing learning. All these aspects need to be considered in order to make a validation framework fit into the nature of ongoing classroom for more effective implementation and validation of formative classroom assessment. The insight discussed above supports Kane and Wools' s (2020) and Moss' s (2003, 2013, 2016) perspectives on classroom assessment and validation by revealing some limitations of both a measurement-oriented approach and Kane's argument-based approach to classroom assessment and validation as already mentioned. Both approaches pay insufficient attention to the dynamic and multifaceted nature of the classroom assessment process and to the kinds of qualitative evidence needed to understand and document the outcomes of classroom language learning and development. This does not mean, however, that the earlier conceptualisations of Kane's argument-based approach (e.g., Kane 1992, 2006, 2011, 2012, 2013, 2016a, 2016b) is not effective, but owing to its focus on the psychometric aspect of validity, it is not sufficient in its current formulation to thoroughly validate classroom assessment in a local context. Very recently, Kane and Wools (2020) have revisited the approach in order to accommodate the complex nature of classroom assessments.

7.2.2 Pedagogical Implications

The current findings revealed the formative impact of integrating diagnostic assessment in an ongoing classroom in promoting teaching and learning in a tertiary EFL writing classroom. Clear, specific, and discrete diagnostic descriptors can help teachers

and students to identify strengths and weaknesses in the language skills or learning contents to meet curriculum goals. However, diagnostic feedback yielded by a diagnostic tool alone is not sufficient and effective unless teachers know how to meaningfully interpret diagnostic results and employ appropriate types and modes of feedback provision (e.g., oral, written, direct, and/or indirect feedback) to further describe and explain diagnostic results to a learner as pointed out by Jang and Wagner (2014) and Kunnan and Jang (2009). All this could enhance the impact of a diagnostic rating scale by helping learners better digest diagnostic feedback, know his or her status of mastery towards expected criteria or learning goals, and realise the areas for further improvement.

In addition, given that the process of self-assessment could enhance effectiveness of diagnostic feedback and self-regulated learning skills, the findings of this study point to the limited usefulness of a diagnostic rating scale for low-ability students as they may not be able to appropriately comprehend diagnostic criteria and identify their own strengths and weaknesses in specific skills. Teachers, therefore, need to provide more assistance and support to help low-proficiency students gain maximum benefit from self-assessment procedures and diagnostic results. For example, depending on the learners' preferences, teachers may need to provide direct corrective feedback pointing the learners to their mistakes and the correct forms they need to follow to correct their mistakes. Previous research showed positive effect of corrective feedback on low-proficiency ESL learners' writing (Mekala & Ponmani, 2017). Another potential approach, as recommended by some teachers and students in this study, is to have low-ability students work and discuss their work with their higher-ability peers through self- and peer-assessment processes, which was found in previous research to promote ESL students' writing improvement (Yu & Lee, 2016; Yu & Hu, 2017).

In addition, the present findings indicated that this group of Thai EFL higher-education learners have the serious problem regarding sentence accuracy, punctuation use, main idea summarisation, supporting idea logic, and thesis restatement. Over the three sequential tasks, they made too many mistakes of these features and skills, particularly ungrammatical sentences such as fragment and run-on sentences. In particular, sentence problem, the most serious problem of the students, is one of the common and persistent problems for L2 learners, which could be caused by several factors such as the interaction between developing linguistic competence and basic principles of

information ordering (Yates & Kenkel, 2002), first language interference, and overgeneralisation of English language rules (Reid, 1998). Although composition writing courses may be focused on the discoursal or organisational level, teachers need to pay attention to sentential-level and language use problems as well, for too many of these errors could negatively affect the meaning and flow of ideas or information in an essay. If, at all, possible, as Lee (2015) suggested, teachers should go beyond identifying students' strengths and weaknesses to investigate the root causes underlying the problems, in particular sentence accuracy, by for example, probing students to explain their reasons for producing ungrammatical sentences and other problematic skills as they did. This could help teachers to provide more targeted feedback and remedial intervention and work out specific strategies, related to explicit corrective feedback and error corrections, to enable students to become aware of their weaknesses and take relevant remedial actions effectively. Apart from the information about the writing problems, there may be other tendencies that emerged from repeated applications of the diagnostic scale that could inform the writing curriculum.

7.2.3 Methodological Implications

The current findings provide some insights regarding the contribution of the multi-stage exploratory sequential mixed-methods research design (Creswell & Plano Clark, 2018) to scale development and validation research. The multi-stage mixed-methods research allows researchers to choose and combine qualitative and qualitative methods of data collection and analysis to collect and analyse multisource data concurrently and/or sequentially over time in order to develop and revise an assessment instrument, while at the same time accumulating multisource evidence to justify assessment validity. In this regard, the multi-stage exploratory sequential mixed-methods research design is particularly well-suited for a longitudinal nature of scale development and validation research.

In this study, the quantitative methods included CTT and MFRM analyses which provided a variety of psychometric indices indicating the quality of the scale, the rater behaviours, and the student writing performances. The current study's findings, however, draw attention to the limited usefulness of the MFRM method for diagnostic language assessment, which is focused more on individuals' diagnosis. The MFRM analysis is

particularly useful for diagnosing the scale functioning, rater behaviours, and group-level student ability. Yet, it does not seem to yield user-friendly diagnostic results pointing to individual learners' strengths and weaknesses on domains and specific descriptors. Such diagnostic information is useful for individual learners' feedback and should be obtained from a diagnostic language assessment. Although the Rasch analysis can generate the KIDMAP (see Jin et al., 1999) pointing to each student's strengths and weaknesses, the KIDMAP is not easy to interpret and is not user-friendly to students and even teachers. It needs to be modified and complemented with additional elements, for instance verbal description, to make it more digestible and user-friendly. An alternative analytic approach is to use a Diagnostic Classification Model (DCM), aka Cognitive Diagnostic Model, technique (e.g., Chiu et al., 2018) which can generate more interpretable and digestible diagnostic information showing students' strengths and weaknesses on both specific skills and ability domains at the group and individual levels. However, it is very important to ensure that the number of examinees is sufficient for a particular DCM method so as to generate reliable estimates.

As well as the quantitative analyses, this research employed a qualitative content analytic approach to examine the teachers' and students' perceptions of the scale, elicited by a semi-structured interview method. The qualitative findings complemented the findings from the quantitative analyses by revealing additional information about the functioning, usefulness, and impact of the diagnostic rating scale and formative diagnostic assessment, and particularly the problems and pitfalls of the scale functioning and characteristics which could not be detected by the quantitative methods. This indicates that sound validity evidence based merely on psychometric approaches does not suffice to build a sound validity argument for an assessment tool. Qualitative approaches are thus necessary to triangulate and complement psychometric evidence with, for instance, raters' decision-making behaviour and users' perceptions of an assessment instrument.

7.3 Limitations of the Study

Developing a diagnostic rating scale and validating a formative assessment in the classroom context involves multiple information sources, research activities, and close collaboration from teachers and learners. Notwithstanding the best possible attempts to

ensure reliable and valid research results, the current PhD study has not been done without limitations.

To begin with, this study focuses on designing and developing a measurement-driven diagnostic rating scale for diagnostic language assessment situated within a formative assessment to provide useful information to promote teaching and learning in a language classroom. However, the extent to which the formative diagnostic assessment improve student learning depends not only on the quality of diagnostic feedback yielded by the scale but also on the types and modes of feedback delivery individual teachers used to provide diagnostic feedback to students. As pointed out by Jang and Wagner (2014), the formative potential of diagnostic assessment to advance student learning is realised when diagnostic feedback is meaningfully interpreted and used by teachers and learners. The full potential of the diagnostic assessment on student learning improvement also depends on teachers' diagnosis experience and expertise (Alderson et al., 2015; Kunnan & Jang, 2009) and their follow-up remedial teaching in response to students' learning problems (Alderson et al., 2015; Lee, 2015). Therefore, the diagnostic rating scale alone is not sufficient to effectively promote student learning as the teachers' feedback provision and remedial teaching methods could mediate the consequence of the formative diagnostic assessment and hence the results of this study. Since this research did not probe into how teachers gave diagnostic feedback and remedial teaching to learners, it is thus impossible to gauge the extent to which such factors influence the impact of the formative diagnostic assessment.

Although the formative diagnostic assessment was designed to be longitudinal to examine its formative contribution to teaching and learning over the course, the scale was implemented only on students' first-draft essays on three assignment tasks, which did not cover the final writing task assigned in the classrooms. Knowing the diagnostic results on students' second-draft essays and on the final assignment tasks could have provided a fuller picture of the formative consequences of the assessment on student learning progression. A more longitudinal diagnostic assessment might have shed more light on students' writing improvement as certain writing skills may be challenging and need more time for them to improve on. Accordingly, improvement on these skills might not have been explicitly observed over a period of one semester. An experimental design using, such as the one-group pretest-posttest design and the experimental- control group,

pretest-posttest design with delayed posttest, could help to ascertain language development. Moreover, a focus group or in-depth interviewing could shed more light on individual students' score-based learning profiles showing, for instance, small improvements, large improvements, or no improvements. All this, nevertheless, was not a focus of the present study and beyond the scope of what could be implemented.

Another factor that may have impacted the current MFRM results is the rating process used for this research. In order to link the score data assigned by teachers in different courses for the MFRM analysis, I, as the scale developer and teacher in the context, randomly rated about half of the students' essays in each intact classroom. Yet, I and the teachers rated student essays under different conditions. During the ongoing classrooms, the teachers' decision-making processes were influenced by a variety of factors related to teaching and professional workload and this was not true for me. Without my ratings, the research results might have been different.

The methods for eliciting qualitative data may also have impacted the findings of the current research. Due to the issue of research ethic, it was impossible for this study to start collecting data at the scale construction stage since the research ethic application was not approved yet at this stage. If the teachers' voices had been gathered to inform the scale construction from the beginning of the scale development process, the findings might have been different. At the scale trialling stage, a semi-structured interview was conducted to ask each teacher about students' learning and writing problems and in the group discussion, the teachers reviewed and trialled the scale with two samples of student essays before providing feedback. Although all the teacher feedback from the interview and group discussion was put into the scale criteria revision at the scale trialling stage, the teacher interview and group discussion may not have tapped effectively into the teachers' rating behaviours and perceived rating criteria. In addition, the characteristics of the student trialling essays with which the teachers interacted, while qualitatively trialling the scale, may not have represented those produced by the students in the actual classroom context and thus were limited to effectively trigger the teachers' perceived features of writing quality. If more effective methods, such as a concurrent think-aloud protocol, and a more representative number of student writing performances had been used, the teachers may have provided more comprehensive feedback about the writing skills and features to be included in the diagnostic criteria. Moreover, at the scale

implementation stage, four teachers responded to the perception interview in English, which is their second language. This could limit the effectiveness of their expressions about the scale perceptions, which might have impacted the qualitative findings.

In addition, the student self-assessment behaviours were investigated through statistical analyses of group-level data. As each student self-rated only his or her own essays, the self-assessment data were not appropriate for the MFRM analysis. Besides, the student self-regulated learning development was inferred from the student perception interview and the self-assessment performance based on statistical analyses. Therefore, other aspects of the self-assessment and self-regulated learning behaviours and their causal relationship remain underexplored. Using other quantitative methods, such as the MFRM analysis, to investigate individual students' self-assessment behaviour and other qualitative methods, such as in-depth interview or think-aloud protocol, to examine individual students' self-regulated learning behaviours would have yielded more insightful results about the self-assessment and its impact on student self-regulated learning. This would in turn provide stronger evidence to support the consequence of the formative diagnostic assessment.

Another limitation is related to instrument validation. It should be noted that validation is a lengthy process which typically requires multiple studies over an extended period of time to accumulate evidence for the inferences, particularly the consequence inference. It is therefore beyond the scope of this research to provide thorough evidence for all inferences.

7.4 Recommendations for Future Research

A number of areas and issues related to diagnostic language assessment remain understudied in this research. Future research should extend the current study to diagnose students' writing process, which could help teachers to uncover the underlying causes of students' weaknesses on writing products. Diagnosing both writing process and product could provide more insightful diagnostic information that could help teachers to effectively solve students' writing problems and improve their learning and writing ability.

The scope of this study was limited to the diagnosis stage, focusing on development and administration of a diagnostic rating scale. To bring diagnostic language assessment to its full fruition in a language classroom, a more longitudinal

research design is needed and should incorporate the design of feedback delivery and remedial instruction in diagnostic language assessment procedures since feedback provision methods and remedial intervention shape the effectiveness of diagnostic language assessment on student learning (Alderson et al., 2015; Lee, 2015). Examining the effects of different types and modes of feedback delivery on students with different proficiency levels could also provide further insights into how students with different proficiency levels process diagnostic feedback and which types and modes of diagnostic feedback are most effective for different groups of learners.

In this study, the descriptors were judged binarily to ensure the scale practicality but the binary scoring was perceived as providing crude information on the quality of writing skills. Future research could consider varying rating formats (e.g., dichotomous, polytomous, partial credit, or mixed formats) in a diagnostic scale in order to appropriately capture the information or granularity of different writing skills. However, as already noted, researchers should keep in mind that the number of rating options, points, or bands on individual descriptors affects rater judgement, scale practicality and diagnostic information. Fewer rating options tend to enhance scale practicality and rating consistency but can result in less detailed diagnostic information. Several rating options tend to provide more detailed diagnostic information but can reduce scale practicality and rating consistency. Future research should also examine the effects of different types of scoring format on quality of diagnostic assessment results, raters' diagnostic decision-making behaviours, and learners' interpretation and processing of diagnostic results. Another interesting area for future research is to investigate the effects of the characteristics of descriptors with different difficulty levels and essays of different quality on raters' diagnostic judgements on binary or polytomous descriptors. The results would be useful to inform descriptor wording and rater training.

Despite the perception findings suggesting the usefulness of the diagnostic scale in supporting teaching and learning, the diagnostic scores could not be used as reliable evidence of learning progression due to the non-standardised nature of assessment and task variability. Future research should consider both functional and measurement perspectives with more emphasis on the functional perspective when developing and evaluating a classroom assessment as proposed by Kane and Wools (2020) and Moss (2003, 2013, 2016). As the functional perspective emphasises how well an overall

assessment support the attainment of assessment purposes and consequences (Kane & Wools, 2020), future research should employ qualitative methods (e.g., interview, focus group, eye-tracking, stimulus recall, think-aloud protocol, classroom discourse or corpus analysis, and conversational analysis) which are particularly suitable for investigating and collecting real-time evidence of learning, particularly learning progression and self-regulated learning. For example, a classroom conversational analysis could reveal students learning progression emerging from teacher-student interactions, thus providing evidence for the consequence inference. A discourse or error analysis of student writing performances could shed more light on student progress on, for instance, writing accuracy, complexity and fluency. A think-aloud protocol is effective to gain insight into teachers' rating behaviours and perceived features of writing quality as well as students' self-assessment and self-regulated learning behaviours. A think-aloud protocol is also useful to examine if raters' cognitive processes while making judgement are aligned with relevant theoretical models, thus providing evidence for the explanation inference (Knoch & Chapelle, 2018). More research is called for to investigate diagnostic language assessment from the qualitative and functional perspectives.

The teacher and student perceptions revealed that the diagnostic scores on the scale alone may not fully provide meaningful and descriptive diagnostic information. Future research should design and incorporate a separate diagnostic profile report which includes numeric, visual and verbal descriptions of individual students' strengths and areas for further improvement on both specific skills and domains. Such a diagnostic profile report can ease teachers' and students' interpretation of diagnostic results and writing improvement over drafts or tasks, thereby enriching the scale usefulness. To optimise and accelerate a formative diagnostic assessment and immediate feedback delivery, further research should harness data science and analysis technology, such as R or RStudio, and psychometric methods, such as nonparametric diagnostic classification models (e.g., Chiu et al., 2018; Li et al., 2020) to develop a computer-based or online formative diagnostic assessment system through which teachers can assess students' writing performances and immediately provide diagnostic results in the form of diagnostic profile reports. With the interface and assistance of an advanced psychometric DCM and open-source R data science programme, the techno-enhanced formative diagnostic assessment system with built-in psychometric analysis could yield reliable diagnostic

results and generate immediate diagnostic profile reports digestible and user-attractive to teachers and students in a small-scale rater-mediated classroom assessment. With such assessment system, future research may need to train teachers on how to utilise the assessment system to promote individualised feedback, differentiated instruction, and personalised learning. An investigation of teachers' and students' perceptions of a diagnostic profile report will also indicate how well the system-generated diagnostic profile report promotes learning. Developing such technology-enhanced formative assessment system is undeniably very challenging, particularly for local classroom teachers, but once well established, it could significantly optimise the power and impact of diagnostic language assessment on teaching and learning. As pointed out by Alderson (2005) and Kunnan and Jang (2009), without the assistance of technology, it would be challenging to provide immediate and quality diagnostic feedback in an ongoing classroom.

Although the present IUA framework was judiciously altered to fit the research scope and assessment practice, it remained rather broad and still underrepresented the interpretations and uses of the diagnostic information made during the ongoing classroom. Further scale development initiatives should allow teachers and/or key stakeholders to get involved in specifying and adapting the IUA, which may initially be broadly framed by researchers, in order to fit how they actually make interpretations and uses of assessment results so that newly-emerging interpretations and uses can be progressively added. In this way, the actual interpretations and uses could be thoroughly documented, thereby enriching the IUA framework. The well-specified IUA structure in turn directs the types of backing evidence to be examined and collected to justify the IUA. To accomplish this, researchers, classroom teachers, and/or other key stakeholders need to work closely, collaboratively, and continually to develop the IUA and work out ways to accumulate relevant evidence to the IUA, as advocated by Kane and Wools (2020).

7.5 Concluding Remarks

Developing an effective diagnostic scale and validating a formative assessment in the classroom involve multiple information sources, multiple methods of data collection and analysis, continual collaboration between researchers and classroom stakeholders, and thus multiple stages of research activities. While research has been dedicated to

large-scale, high-stakes, and standardised diagnostic language assessment, this study aims to advance the fields of diagnostic language assessment and formative assessment in the L2 classroom context, where EFL learners probably have the most opportunity to learn and develop a second language. Nevertheless, this study does not claim that the newly-developed diagnostic scale can fully serve diagnostic language assessment purposes, which may vary from context to context. There were some problems in terms of the functionality, characteristics, and application of the scale in the real-world classroom situation and other sources of validity evidence remained under-explored within the current validation framework. It should also be borne in mind that a diagnostic test, repeated assessment, and student self-assessment are just some elements of diagnostic language assessment and its full formative potential to advance teaching and learning also depends on other elements and variables situated within a particular assessment context, which need to be further investigated. It is very much hoped that this research contributes to the field by offering a number of insights into how a diagnostic tool should be designed to serve its intended purposes and how a fuller formative diagnostic assessment should be designed and validated in a specific classroom context.

References

Ahmad, Z. (2019). Analyzing argumentative essay as an academic genre on assessment framework of IELTS and TOEFL. In S. Hidri (Ed.), *English language teaching research in the Middle East and North Africa* (pp. 279-299). Palgrave Macmillan.

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

Alderson, J. C. (2005). *Diagnosing foreign language proficiency: The interface between learning and assessment*. Continuum.

Alderson, J. C. (2010). Cognitive diagnosis and Q-matrices in language assessment: A commentary. *Language Assessment Quarterly, 7*(1), 96–103. https://doi.org/10.1080/15434300903426748

Alderson, J. C., Brunfaut, T., & Harding, L. (2015). Towards a theory of diagnosis in second and foreign language assessment: Insights from professional practice across diverse fields. *Applied Linguistics, 36*(2), 236–260. https://doi.org/10.1093/applin/amt046

Alderson, J. C., Brunfaut, T., & Harding, L. (2017). Bridging assessment and learning: a view from second and foreign language assessment. *Assessment in Education: Principles, Policy & Practice, 24*(3), 379–387. https://doi.org/10.1080/0969594X.2017.1331201

Alderson, J. C., Haapakangas, E.-L., Huhta, A., Nieminen, L., & Ullakonoja, R. (2015). *The diagnosis of reading in a second or foreign language*. Routledge.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. American Educational Research Association.

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

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andrade, H. L. (2019). A critical review of research on student self-assessment. *Frontiers in Education, 4*(87), 1–13. https://doi.org/10.3389/feduc.2019.00087

Andrade, H. L., & Brown, G. T. L. (2016). Student self-assessment in the classroom. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of human and social conditions in assessment* (pp. 319–334). Routledge.

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge.

Andrich, D. (1988). *Rasch models for measurement*. Sage.

Arter, J., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Corwin Press.

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

Asadifard, A., & Afghari, A. (2019). The Effect of Systematic Implementation of Formative Assessment on Male and Female EFL Learners' Academic Achievement. *Research in English Language Pedagogy, 7*(1), 71–90. doi: 10.30486/relp.2019.663423

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

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford University Press.

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

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

Baker, B. A. (2012). Individual differences in rater decision-making style: An exploratory mixed-methods study. *Language Assessment Quarterly, 9*(3), 225–248. https://doi.org/10.1080/15434303.2011.637262

Baleghizadeh, S., & Hajizadeh, T. (2014). Self- and teacher-assessment in an EFL writing class. *GIST Education and Learning Research Journal, 8*, 99–117. https://doi.org/10.2687/16925777.116

Banerjee, J., Yan, X., Chapman, M., & Elliott, H. (2015). Keeping up with the times: Revising and refreshing a rating scale. *Assessing Writing, 26*, 5–19. https://doi.org/10.1016/j.asw.2015.07.001

Barkaoui, K. (2007a). Participants, texts, and processes in second language writing assessment: A narrative review of the literature. *The Canadian Modern Language Review, 64* (1), 97–132. http://dx.doi.org/10.3138/cmlr.64.1.099

Barkaoui, K. (2007b). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12*(2), 86–107. https://doi.org/10.1016/j.asw.2007.07.001

Barkaoui, K. (2008). *Effects of scoring method and rater experience on ESL essay rating processes and outcomes* (Publication No. NR44703) [Doctoral dissertation, University of Toronto]. ProQuest Dissertations and Theses Global.

Barkaoui, K. (2010). Variability in ESL essay rating processes: The role of the rating scale and rater experience. *Language Assessment Quarterly, 7*(1), 54–74. https://doi.org/10.1080/15434300903464418

Barkaoui, K. (2014). Multifaceted Rasch analysis for test evaluation. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–22). John Wiley & Sons, Inc. https://doi: 10.1002/9781118411360.wbcla070

Bereiter, C., & Scardamalia, M. (1987). *The psychology of written composition*. Lawrence Erlbaum.

Betts, L., & Hartley, J. (2012). The effects of changes in the order of verbal labels and numerical values on children's scores on attitude and rating scales. *British Educational Research Journal, 38*(2), 319–331. http://dx.doi.org/10.1080/01411926.2010.544712

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Sage.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading.* Association for Supervision & Curriculum Development (ASCD).

Brown, A., & Lumley, T. (1991). *The University of Melbourne ESL Test. Final report*. Language Testing Research Centre, University of Melbourne.

Brown, G. T. L., & Harris, L. R. (2013). Student self-assessment. In J. H. McMillan (Ed.), *The Sage handbook of research on classroom assessment* (pp. 367–393). Sage.

Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: directions and cautions for research. *Assessment in Education: Principles, Policy & Practice 22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brown, J. D. (2012). *Developing, using, and analyzing rubrics in language assessment with case studies in Asian and Pacific languages*. National Foreign Language Resource Center.

Can Daşkın, N., & Hatipoğlu, Ç. (2019). Reference to a past learning event as a practice of informal formative assessment in L2 classroom interaction. *Language Testing, 36(4)*, 527–551. https://doi.org/10.1177/0265532219857066

Canale, M. (1983). On some dimensions of language proficiency. In J. Oller, (Ed.), *Issues in language testing research* (pp. 333–342). Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*(1), 1–47. https://doi.org/10.1093/applin/I.1.1

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3–22. https://doi.org/10.1177/026553229701400102

Chalhoub-Deville, M. (2003). Second language interaction: current perspectives and future trends. *Language Testing, 20*(4), 369–383. https://doi.org/10.1191/0265532203lt264oa

Chan, S., Inoue, C., & Taylor, L. (2015). Developing rubrics to assess the reading-into writing skills: A case study. *Assessing Writing, 26*, 20–37. http://dx.doi.org/10.1016/j.asw.2015.07.004

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

Chapelle, C. A. (2011a). Validation in language assessment. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. II, pp. 717–730). Routledge.

Chapelle, C. A. (2011b). Validity argument for language assessment: The framework is simple…. *Language Testing, 29*(1), 19–27. https://doi.org/10.1177/0265532211417211

Chapelle, C. A. (2012). Conceptions of validity. In G. Fulcher & F. Davidson (Eds.), *Routledge handbook of language testing* (pp. 21–33). Routledge.

Chapelle, C. A., & Voss, E. (2014). Evaluation of language tests through validation research. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–17). John Wiley & Sons, Inc. doi: 10.1002/9781118411360.wbcla110

Chapelle, C. A., Chung, Y.-R., Hegelheimer, V., Pendar, N., & Xu, J. (2010). Towards a computer-delivered test of productive grammatical ability. *Language Testing, 27*(4), 443–469. https://doi.org/10.1177/0265532210367633

Chapelle, C. A., Cotos, E., & Lee, J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing, 32*(3), 385–405. https://doi.org/10.1177/0265532214565386

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

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

Chen, H., & Chen, J. (2016). Retrofitting non-cognitive-diagnostic reading assessment under the Generalized DINA model framework. *Language Assessment Quarterly, 13*(3), 218–230. https://doi.org/10.1080/15434303.2016.1210610

Cheng, L., & Fox, J. (2017). *Assessment in the language classroom: Teachers supporting student learning*. Palgrave Macmillan.

Chiu, C.-Y., Sun, Y., & Bian, Y. (2018). Cognitive diagnosis for small educational programs: The general nonparametric classification method. *Psychometrika, 83*(2), 355–375. https://doi.org/10.1007/s11336-017-9595-4

Cho, D. (2008). Investigating EFL writing assessment in a classroom setting: Features of composition and rater behaviors. *The Journal of AsiaTEFL, 5*(4), 49–84. https://www.earticle.net/Article/A182195

Cizek, G. J. (2012). Defining and distinguishing validity: Interpretations of score meaning and justifications of test use. *Psychological Methods, 17*(1), 31–43. https://doi.org/10.1037/a0026975

Cizek, G. J. (2016). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice, 23*(2), 212–225. https://doi.org/10.1080/0969594X.2015.1063479

Clauser, J. C., & Hambleton, R. K. (2018). Item analysis procedures for classroom assessments in higher education. In C. Secolsky & D. B. Denison (Eds.), *Handbook on measurement, assessment, and evaluation in higher education* (2nd ed., pp. 296–309). Routledge.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*(1), 155–159. https://doi.org/10.1037/0033-2909.112.1.155

Connor, U., & Mbaye, A. (2002). Discourse approaches to writing assessment. *Annual Review of Applied Linguistics, 22*, 263–278. https://doi.org/10.1017/S0267190502000144

Corder, S. P. (1981). *Error Analysis and Interlanguage*. Oxford: Oxford University Press.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). Sage.

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Creswell, J. W., & Zhou, Y. (2016). What is mixed methods research? In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 35–50). Cambridge University Press.

Cronbach, L. J. (1988). Five perspectives on validity argument. In H. Wainer & H. Braun (Eds.), *Test validity* (pp. 3–17). Erlbaum.

Cumming, A. (2013). Validation of language assessments. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 1–10). Blackwell Publishing Ltd. doi: 10.1002/9781405198431.wbeal1242

Cumming, A. (2015). Design in four diagnostic language assessments. *Language Testing, 32*(3), 407–416. https://doi.org/10.1177/0265532214559115

Cumming, A. (2016). Theoretical orientations to L2 writing. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 65–88). Walter de Gruyter.

Cumming, A., Kantor, R., & Powers, D. (2001). *Scoring TOEFL essays and TOEFL 2000 prototype writing tasks: An investigation into raters' decision making and development of a preliminary analytic framework* (TOEFL Monograph Series N 22). Educational Testing Service.

Cumming, A., Kantor, R., & Powers, D. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86*(1), 67–96. https://doi.org/10.1111/1540-4781.00137

Davis, L. (2015). Designing and using rubrics. In J. D. Brown & C. Coombe (Eds.), *The Cambridge guide to research in language teaching and learning* (pp. 238–246). Cambridge University Press.

DeVellis, R. E. (2017). *Scale development: Theory and applications* (4th ed.). Sage.

di Gennaro, K. K. (2011). *An exploration into the writing ability of generation 1.5 and international second language writers: A mixed methods approach* (Order No. 3484222) [Doctoral dissertation, University of Columbia]. ProQuest Dissertations and Theses Global.

Doe, C. (2014). Diagnostic English Language Needs Assessment (DELNA). *Language Testing, 31*(4), 537–543. https://doi.org/10.1177/0265532214538225

Doe, C. (2015). Student interpretations of diagnostic feedback. *Language Assessment Quarterly, 12*(1), 110–135. https://doi.org/10.1080/15434303.2014.1002925

Doe, C. D. (2013). *Validating the Canadian academic English language assessment for diagnostic purposes from three perspectives: Scoring, teaching, and learning* (Publication No. NS27834) [Doctoral dissertation, Queen's University]. ProQuest Dissertations and Theses Global.

Douglas, D., & Hegelheimer, V. (2007). *Strategies and use of knowledge in performing New TOEFL listening tasks.* Final Report to the Educational Testing Service. Department of English, Iowa State University.

Eckes, T. (2015). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Peter Lang.

Eckes, T. (2019). Many-facet Rasch measurement: Implications for rater-mediated

language assessment. In V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 152–175). Routledge.

Elder, C, Knoch, U, Barkhuizen, G., & von Randow, J. (2005). Individual feedback to enhance rater training: Does it work? *Language Assessment Quarterly, 2*(3), 175–196. https://doi.org/10.1207/s15434311laq0203_1

Elder, C. (2017). language assessment in higher education. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 271–286). Springer International Publishing. doi: 10.1007/978-3-319-02261-1_35

Elder, C., & Read, J. (2015). Post-entry language assessments in Australia. In J. Read (Ed.), *Assessing English proficiency for university study* (pp. 25–46). Palgrave Macmillan.

Elder, C., & Read, J. (2015). Post-entry language assessments in Australia. In J. Read (Ed.), *Assessing English proficiency for university study* (pp. 70–92). Palgrave Macmillan.

Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007). Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing, 24*(1), 37–64. https://doi.org/10.1177/0265532207071511

Elder, C., Knoch, U., & Zhang, R. (2009). Diagnosing the support needs of second language writers: Does the time allowance matter? *TESOL Quarterly, 43*(2), 351–360. http://www.jstor.com/stable/27785015

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

Engelhard, J. G., & Wind, S. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data*. MIT Press.

Esfandiari, R., & Myford C. M. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*(2), 111–131. https://doi.org/10.1016/j.asw.2012.12.002

Fan, J., & Bond, T. (2019). Applying Rasch measurement in language assessment: Unidimensionality and local independence. V. Aryadoust & M. Raquel (Eds.), *Quantitative data analysis for language assessment volume I: Fundamental techniques* (pp. 83–102). Routledge.

Flower, L., & Hayes, J. (1980). The Cognition of Discovery: Defining a Rhetorical Problem. *College Composition and Communication, 31*(1), 21–32. doi:10.2307/356630

Flower, L., & Hayes, J. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication, 32*(4), 365–387. doi:10.2307/356600

Fulcher, G., & Davidson, F. (2007). Constructs and models. In G. Fulcher & F. Davidson (Eds.), *Language testing and assessment: An advanced resource book* (pp. 36–51). Routledge.

Fulcher, G. (1993). *The construction and validation of rating scales for oral tests in English as a foreign language* [Unpublished doctoral dissertation]. University of Lancaster.

Fulcher, G. (1996). Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*(2), 208–238. https://doi.org/10.1177/026553229601300205

Fulcher, G. (2003). *Testing second language speaking*. Pearson.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher & F. Davison (Eds.), *The Routledge handbook of language testing* (pp. 378–392). Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests: Performance decision trees. *Language Testing, 28*(1), 5–29. https://doi.org/10.1177/0265532209359514

Fung Y. M., & Mei, H. C. (2015). Improving undergraduates' argumentative group essay writing through self-assessment. *Advances in Language and Literary Studies, 6*(5), 215–224. http://dx.doi.org/10.7575/aiac.alls.v.6n.5p.214

George, D., & Mallery, P. (2018). *IBM SPSS statistics 25 step by step: A simple guide and reference* (15th ed.). Routledge.

Glaser, B., & Strauss, A. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Sociology Press.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing: An applied linguistics perspective.* Longman.

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge University Press.

Greene, J. C. (2007). *Mixed methods in social inquiry*. Jossey-Bass.

Guest, G. S., MacQueen, K. M., & Namey, E. E. (2012). *Applied thematic analysis*. Sage.

Guetterman, T. C., & Salamoura, A. (2016). Enhancing test validation through rigorous mixed methods components. In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 153–176). Cambridge University Press.

Han, T. (2017). Scores assigned by inexpert EFL raters to different quality EFL compositions, and the raters' decision-making behaviors. *International Journal of Progressive Education, 13*(1), 136–152.

Harding, L. (2014). Communicative language testing: Current issues and future research. *Language Assessment Quarterly, 11*(2), 186–197. https://doi.org/10.1080/15434303.2014.895829

Harding, L., Alderson, J. C., & Brunfaut, T. (2015). Diagnostic assessment of reading and listening in a second or foreign language: Elaborating on diagnostic principles. *Language Testing, 32*(3), 317–336. https://doi.org/10.1177/0265532214564505

Harris, L. R., & Brown, G. T. L. (2018). *Using self-assessment to improve student learning*. Routledge.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell (Eds.), *The science of writing* (pp. 1–27). Erlbaum.

Hayes, J. R., & Flower, L. (1980). Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing: An interdisciplinary approach* (pp. 3–30). Lawrence Erlbaum.

Heidarian, N. (2016). Investigating the effect of using self-assessment on Iranian EFL learners' writing. *Journal of Education and Practice, 7*(22), 80–89.

Hirvela, A., Hyland, K., & Manchón, R. M. (2016). Dimensions in L2 writing theory and research: Learning to write and writing to learn. In R. M. Manchón & P. K. Matsuda (Eds.), *Handbook of second and foreign language writing* (pp. 45–63). Walter de Gruyter.

Huang J., Han, T., Tavano, H., & Hairston, L (2014). Using generalizability theory to examine the impact of essay quality on ESOL writing assessment: A Turkish case study. *China- US Education, 3*, 3–20.

Huang, J. (2009). Factors affecting the assessment of ESL students' writing. *International Journal of Applied Educational Studies, 5*(1), 1–17.

Hung, Y.-J., Samuelson, B. L., & Chen, S.-C. (2016). Relationships between peer- and self-assessment and teacher assessment of young EFL learners' oral presentations. In M. Nikolov (Ed.), *Assessing young learners of English: Global and local perspectives* (pp. 317–338). Springer.

Hymes, D. (1972). On communicative competence. In J. B. Pride, & J. Holmes (Eds.), *Sociolinguistics* (pp. 53–73). Penguin.

Im, G.-H., Shin, D., & Cheng, L. (2019). Critical review of validation models and practices in language testing: their limitations and future directions for validation research. *Language Testing in Asia, 9*(14), 1–26. https://doi.org/10.1186/s40468-019-0089-4

Jamieson, J. (2014). Defining constructs and assessment. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–17). John Wiley & Sons, Inc. doi: 10.1002/9781118411360.wbcla062

Jamieson, J., & Poompon, K. (2013). Developing analytic rating guides for TOEFL iBT's integrated speaking tasks. *ETS Research Report Series, 2013*(1), i–93. doi: 10.1002/j.2333-8504.2013.tb02320.x

Jang, E. E. (2009). Cognitive diagnostic assessment of L2 reading comprehension ability: Validity arguments for applying Fusion Model to LanguEdge assessment. *Language Testing 26*(1), 31–73. doi:10.1177/0265532208097336

Jang, E. E. (2012). Diagnostic assessment in language classrooms. In G. Fulcher & F. Davison (Eds.), *The Routledge handbook of language testing* (pp. 120–133). Routledge.

Jang, E. E., & Wagner, M. (2014). Diagnostic feedback in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–19). John Wiley & Sons. doi: 10.1002/9781118411360.wbcla081

Jeong, H. (2017). Narrative and expository genre effects on students, raters, and performance criteria. *Assessing Writing, 31*, 113–125. https://doi.org/10.1016/j.asw.2016.08.006

Jin, L. Y., Linacre, J. M., & Chyr, Y. O. (1999). *KIDMAP - A diagnostic tool for teachers*. Retrieved from Institute of Education Science website: https://eric.ed.gov/?id=ED440111

Jing, M. J. (2017). Using formative assessment to facilitate learner self-regulation: A case study of assessment practices and student perceptions in Hong Kong. *Taiwan Journal of TESOL, 14* (1), 87–118.

Jiuliang, L. (2014). Examining genre effects on test takers' summary writing performance. *Assessing Writing, 22*, 75–90. http://dx.doi.org/10.1016/j.asw.2014.08.003

Johnson, R. J., & Morgan, G. B. (2016). *Survey scale: A guide to development, analysis, and reporting*. Guilford Press.

Kane, M. (2012). Articulating a validity argument. In G. Fulcher & F. Davison (Eds.), *The Routledge handbook of language testing* (pp. 34–47). Routledge.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527–535. http://dx.doi.org/10.1037/0033-2909.112.3.527

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

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

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

Kane, M. T. (2016a). Explicating validity. *Assessment in Education: Principles, Policy & Practice, 23*(2), 198-211. https://doi: 10.1080/0969594X.2015.1060192

Kane, M. T. (2016b). Validation strategies: Delineating and validating proposed interpretations and use of tests cores. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of test development* (2nd ed., pp. 64–80). Routledge.

Kane, M. T., & Wools, S. (2020). Perspectives on the validity of classroom assessments. In S. M., Brookhart & J. H. McMillan (Eds.), *Classroom assessment and educational measurement* (pp. 11–26). Routledge.

Kenyon, D. M., & MacGregor, D. (2012). Pre-operational testing. In G. Fulcher & F. Davison (Eds.), *The Routledge handbook of language testing* (pp. 295–306). Routledge.

Kim, J. (2019). Effects of rubric-referenced self-assessment training on Korean high school students' English writing. *English Teaching, 74*(3), 79-111. doi: 10.15858/engtea.74.3.201909.79

Kim, Y.-H. (2010). *An argument-based validity inquiry into the empirically-derived descriptor-based diagnostic (EDD) assessment in ESL academic writing* [Doctoral dissertation, University of Toronto]. TSpace. https://tspace.library.utoronto.ca/bitstream/1807/24786/1/Kim_Youn-Hee_201006_PhD_thesis.pdf

Kim, Y.-H. (2011). Diagnosing EAP writing ability using the Reduced Reparameterized Unified Model. *Language Testing, 28*(4), 509–541. doi:10.1177/0265532211400860

Knoch, U. (2007). *Diagnostic writing assessment: The development and validation of a rating scale* [Unpublished doctoral dissertation]. University of Auckland.

Knoch, U. (2009a). Diagnostic assessment of writing: A comparison of two rating scales. *Language Testing, 26*(2), 275–304. https://doi.org/10.1177/0265532208101008

Knoch, U. (2009b). *Diagnostic writing assessment: The development and validation of a rating scale*. Peter Lang.

Knoch, U. (2011). Rating scales for diagnostic assessment of writing: What should they look like and where should the criteria come from? *Assessing Writing, 16*(2), 81–96. https://doi.org/10.1016/j.asw.2011.02.003

Knoch, U. (2016). Validation of writing assessment. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 1–6). Blackwell Publishing Ltd. doi: 10.1002/9781405198431.wbeal1480

Knoch, U., & Chapelle, C. A. (2018). Validation of rating processes within an argument-based framework. *Language Testing, 35*(4), 477–499. https://doi.org/10.1177/0265532217710049

Knoch, U., & Elder, C. (2013). A framework for validating post-entry language assessment (PELAs). *Papers in Language Testing and Assessment, 2*(2), 48–65.

Knoch, U., & Macqueen, S. (2017). Assessment in the L2 classroom. In S. Loewen & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 181–202). Routledge Taylor and Francis Group.

Knoch, U., Deygers, B., & Khamboonruang, A. (in press). Re-visiting rating scale development for rater-mediated language performance assessments: Modelling construct and contextual choices made by scale developers. *Language Testing*.

Knoch, U., Read, J., & von Randow, J. (2007). Re-training writing raters online: How does it compare with face-to-face training? *Assessing Writing, 12*(1), 26–43. https://doi.org/10.1016/j.asw.2007.04.001

Koizumi, R., Saka, H., Ido, T., Ota, H., Hayama, M., Sato, M., & Nemoto, A. (2011). Development and validation of a diagnostic grammar test for Japanese learners of English. *Language Assessment Quarterly, 8*(1), 53–72. https://doi.org/10.1080/15434303.2010.536868

Kunnan, A. J., & Jang, E. E. (2009). Diagnostic Feedback in Language Assessment. In M. H. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 610–627). Blackwell.

Lallmamode, S. P., Mat Daud, N., & Abu Kassim, N. L. (2016). Development and initial argument-based validation of a scoring rubric used in the assessment of L2 writing electronic portfolios. *Assessing Writing, 30*, 44–62. http://dx.doi.org/10.1016/j.asw.2016.06.001

Lane, S. (2019). Modeling rater response processes in evaluating score meaning. *Journal of Educational Measurement, 56*(3), 653-663. https://doi.org/10.1111/jedm.12229

Lantolf, J., & Poehner, M. (2011). Dynamic assessment in the classroom: Vygotskyan praxis for second language development. *Language Teaching Research 15*(1), 11–33. doi: 10.1177/1362168810383328

Larson-Hall, J. (2015). A guide to doing statistics in second language research using SPSS and R (2nd ed.). Routledge.

Lee, I (2011). Formative assessment in EFL writing: An exploratory case study. *Changing English, 18*(1), 99–111. https://doi.org/10.1080/1358684X.2011.543516

Lee, I. (2017). *Classroom writing assessment and feedback in L2 school contexts*. Springer.

Lee, Y.-W. (2015). Diagnosing diagnostic language assessment. *Language Testing, 32*(3), 299–316. https://doi.org/10.1177/0265532214565387

Lee, Y.-W., & Sawaki, Y. (2009a) Application of three cognitive diagnosis models to ESL reading and listening assessments. *Language Assessment Quarterly, 6*(3), 239–263. https://doi.org/10.1080/15434300903079562

Lee, Y.-W., & Sawaki, Y. (2009b). Cognitive diagnosis approaches to language assessment: An overview. *Language Assessment Quarterly, 6*(3), 172–189. https://doi.org/10.1080/15434300902985108

Li, X., Wang, W.-C., & Xie, Q. (2020). Cognitive diagnostic models for rater effects. *Frontiers in Psychology, 11*, 1-12. https://doi.org/10.3389/fpsyg.2020.00525

Linacre, J. M. (1989). *Many-Facet Rasch measurement*. MESA Press.

Linacre, J. M. (2018). *Facets computer program for many-facet Rasch measurement*, version 3.80.4. Winsteps.com.

Lockwood, J. (2013). The Diagnostic English Language Tracking Assessment (DELTA) writing project: a case for post-entry assessment policies and practices in Hong Kong universities. *Papers in Language Testing and Assessment, 2*(1), 30–49.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*(3), 246–276. https://doi.org/10.1191/0265532202lt230oa

Lumley, T. (2005). *Assessing second language writing: The rater's perspective*. Peter Lang.

Mackey, A., & Gass, A. (2016). *Second language research: Methodology and design* (2nd ed.). Routledge.

Matsuno, S. (2009). Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. *Language Testing, 26*(1), 75–100. https://doi.org/10.1177/0265532208097337

McMillan, J. H. (2014). *Classroom assessment: Principles and practice for effective standard-based instruction* (6th ed.). Pearson.

McNamara, T. (1996). *Measuring second language performance*. Longman.

McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: The role of measurement*. Oxford University Press.

Mekala, S., & Ponmani, M. (2017). The Impact of direct written corrective feedback on low proficiency ESL learners' writing ability. *The IUP Journal of Soft Skills, XI*(4), 23–54. https://ssrn.com/abstract=3220360

Menold, N. (2020). Rating-scale labeling in online surveys: An experimental comparison of verbal and numeric rating scales with respect to measurement quality and respondents' cognitive processes. *Sociological Methods & Research, 49*(1), 79–107. https://doi.org/10.1177/0049124117729694

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. http://dx.doi.org/10.1037/0003-066X.50.9.741

Milanovic. M., Saville, N., & Shuhong, S. (1996). A study of decision-making behaviour of composition markers. In M. Milanovic & N. Saville (Eds.), *Studies in Language Testing 3: Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge and Arhem* (pp. 92–115). Cambridge University Press.

Mislevy, R. J., & Yin, C. (2012). Evidence-centered design in language testing. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 208–222). Routledge.

Mohamadi, Z. (2018). Comparative effect of online summative and formative assessment on EFL student writing ability. *Studies in Educational Evaluation, 59*, 29–40. https://doi.org/10.1016/j.stueduc.2018.02.003

Montee, M., & Malone, M. E. (2014). Writing scoring criteria and score reports. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–13). John Wiley & Sons, Inc. doi: 10.1002/9781118411360.wbcla112

Moors, G., Kieruj, N. D., & Vermunt, J. K. (2014). The effect of labeling and numbering of response scales on the likelihood of response bias. *Sociological Methodology, 44*(1), 369–399. https://doi.org/10.1177/0081175013516114

Moss, P. A. (2003). Reconceptualizing validity for classroom assessment. *Educational Measurement: Issues and Practices, 22*(4), 13–25. https://doi.org/10.1111/j.1745-3992.2003.tb00140.x

Moss, P. A. (2013). Validity in action: Lessons from studies of data use. *Journal of Educational Measurement, 50*(1), 91–98. https://doi.org/10.1111/jedm.12003

Moss, P. M. (2016). Shifting the focus of validity for test use. *Assessment in Education: Principles, Policy & Practice, 23*(2), 236–251. https://doi.org/10.1080/0969594X.2015.1072085

Murnane, R. J., Sharkey, N. S., & Boudett, K. P. (2009). Using student-assessment results to improve instruction: Lessons from a workshop. *Journal of Education for*

Students Placed at Risk (JESPAR), 10, 269–280. https://doi.org/10.1207/s15327671espr1003_3

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 371–398.

Naghdipour, B. (2017). Incorporating formative assessment in Iranian EFL writing: A case study. *The Curriculum Journal, 28*(2), 283-299. https://doi.org/10.1080/09585176.2016.1206479

North, B. (1996). *The development of a common framework scale of descriptors of language proficiency based on a theory of measurement* [Unpublished doctoral dissertation]. Thames Valley University.

North, B., & Schneider, G. (1998). Scaling descriptors for language proficiency scales. *Language Testing, 15*(2), 217–262. https://doi.org/10.1177/026553229801500204

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Oscarson, M. (2014). Self-assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–18). John Wiley & Sons. doi: 10.1002/9781118411360.wbcla046

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

Park, H., & Yan, X. (2019). An investigation into rater performance with a holistic scale and a binary, analytic scale on an ESL writing placement test. *Papers in Language Testing and Assessment, 8*(2), 34–64.

Poehner, M. E. (2014). Dynamic assessment in the classroom. In A. J. Kunnan (Ed.), *The companion to language assessment* (pp. 1–16). John Wiley & Sons. doi: 10.1002/9781118411360.wbcla046

Poehner, M. E., & Infante, P. (2016). Dynamic assessment in the language classroom. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 27–290). Walter de Gruyter.

Purpura, J. E. (2008). Assessing communicative language ability: models and their components. In E. Shohamy & N. H. Hornberger (Eds.), *Encyclopedia of language*

and education, Vol. 7. Language testing and assessment (2nd ed., pp. 53–68). Kluwer.

Ranalli, J., Link, S., & Chukharev-Hudilainen, E. (2017). Automated writing evaluation for formative assessment of second language writing: investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology, 37*(1), 8–25. https://doi.org/10.1080/01443410.2015.1136407

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danmarks Paedagogiske Institut.

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

Read, J. (2015). *Assessing English proficiency for university study*. Palgrave Macmillan

Reckase, M. D. (1979). Unifactor Latent Trait Models Applied to Multifactor Tests: Results and Implications. *Journal of Educational Statistics, 4*(3), 207–230. https://doi.org/10.3102/10769986004003207

Reid, J. (1998). Responding to ESL student language problems: Error analysis and revision plans. In P. Byrd & J. Reid (Eds.), *Grammar in the composition classroom* (pp. 118-137). Heinle and Heinle.

Riazi, A. M. (2017). *Mixed methods research in language teaching and learning*. Equinox.

Ross, J. A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment Research and Evaluation, 11*(10). 1–13. https://doi.org/10.7275/9wph-vv65

Şahan, Ö. (2018). *The impact of rating experience and essay quality on rater behaviour and scoring* [Unpublished doctoral dissertation]. Çanakkale Onsekiz Mart University.

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing.* Advance online publication. https://doi.org/10.1177/0265532219900228

Sakyi, A. A. (2000). Validation of holistic scoring for ESL writing assessment: How raters evaluate compositions. In J. J. Kunnan, (Ed.), *Fairness and validation in language assessment* (pp. 129–152). Cambridge University Press.

Salehi, M., & Masoule, Z. S. (2017). An investigation of the reliability and validity of peer, self-, and teacher assessment. *Southern African Linguistics and Applied Language Studies, 35*(1), 1–15. doi: 10.2989/16073614.2016.1267577

Saville, N. (2016). Managing language assessment systems and mixed methods. In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 17–31). Cambridge University Press.

Schoonen, R. (2011). How language ability is assessed. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (Vol. II, pp. 701–716). Routledge.

Schreier, M. (2012). *Qualitative content analysis in practice*. Sage.

Schreier, M. (2014). Qualitative content analysis. In U. Flick (Ed.), *The SAGE handbook of qualitative data analysis* (pp. 170–183). Sage.

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

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

Sireci, S. G. (2016). On the validity of useless tests. *Assessment in Education: Principles, Policy & Practice, 23*(2), 226–235. https://doi.org/10.1080/0969594X.2015.1072084

Smith, D. (2000). Rater judgments in the direct assessment of competency-based second language writing ability. In G. Brindley (Ed.), *Studies in immigrant English language assessment* (Vol. 1, pp. 159–189). Sydney: National Centre for English Language Teaching and Research, Macquarie University. Li Giblin Library

Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. In J. W. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–49). Sage.

Tsagari, D., & Banerjee, J. (2015). Language assessment in the educational context. In M. Begelow & J. Ennser-Kananen (Eds*.), The Routledge handbook of educational linguistics* (pp. 339–52). Routledge.

Turner, C. E. (2012). Classroom assessment. In G. Fulcher & F. Davidson (Eds.), *The Routledge handbook of language testing* (pp. 65–78). Routledge.

Turner, C. E. (2013). Rating scales for language tests. In C. A. Chapelle (Ed.), *Encyclopedia of applied linguistics* (pp. 1–7). Blackwell Publishing Ltd. doi: 10.1002/9781405198431.wbeal1045

Turner, C. E., & Purpura, J. E. (2016). Learning-oriented assessment in second and foreign language classrooms. In D. Tsagari & J. Banerjee (Eds.), *Handbook of second language assessment* (pp. 255–273). Walter de Gruyter.

Turner, C. E., & Upshur, J. A. (2002). Rating scales derived from student samples: Effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly, 36*(1), 49–70. https://doi.org/10.2307/3588360

Ünaldı, İ. (2016). Self and teacher assessment as predictors of proficiency levels of Turkish EFL learners. *Assessment & Evaluation in Higher Education, 41*(1), 67–80. http://dx.doi.org/10.1080/02602938.2014.980223

Upshur, J. A., & Turner, C. E. (1995). Constructing rating scales for second language tests. *ELT Journal, 49*(1), 3–12. https://doi.org/10.1093/elt/49.1.3

Upshur, J. A., & Turner, C. E. (1999). Systematic effects in the rating of second-language speaking ability: test method and learner discourse. *Language Testing, 16*(1), 82–111. https://doi.org/10.1177/026553229901600105

Wagner, M. (2015). *The centrality of cognitively diagnostic assessment for advancing secondary school ESL students' writing: A mixed methods study* (Order No. 3744197) [Doctoral dissertation, University of Toronto]. ProQuest Dissertations and Theses Global.

Wang, W. (2017). Using rubrics in student self-assessment: student perceptions in the English as a foreign language writing context. *Assessment & Evaluation in Higher Education, 42*(8), 1280–1292. https://doi.org/10.1080/02602938.2016.1261993

Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach.* Palgrave Macmillan.

Xi, X., & Davis, L. (2016). Quality factors in language assessment. In D. Tsagari & J. Baberjee (Eds.), *Handbook of second language assessment* (pp. 61–76). Walter de Gruyter.

Xi, X., & Sawaki, Y. (2017). Methods of test validation. In E. Shohamy, I. G. Or, & S. May (Eds.), *Language testing and assessment* (pp. 193–209). Springer International Publishing. doi: 10.1007/978-3-319-02261-1_14

Xiao, Y., & Yang, M. (2019). Formative assessment and self-regulated learning: How formative assessment supports students' self-regulation in English language learning. *System, 81*, 39–49. https://doi.org/10.1016/j.system.2019.01.004

Xie, Q. (2019). Error analysis and diagnosis of ESL linguistic accuracy: Construct specification and empirical validation. *Assessing Writing, 41*, 47–62. https://doi.org/10.1016/j.asw.2019.05.002

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education, 42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

Yates, R., & Kenkelb, J. (2002). Responding to sentence-level errors in writing. *Journal of Second Language Writing, 11*(1), 29–47. https://doi.org/10.1016/S1060-3743(02)00051-6

Yu, S., & Lee, I. (2016). Understanding the role of learners with low English language proficiency in peer feedback of second language writing. *TESOL Quarterly, 50*(2), 483–494. https://doi.org/10.1002/tesq.301

Yu, S., & Hu, G. (2017). Can higher-proficiency L2 learners benefit from working with lower-proficiency partners in peer feedback? *Teaching in Higher Education, 22*(2), 178–192. https://doi.org/10.1080/13562517.2016.1221806

Zhang, J. (2016). Same text different processing? Exploring how raters' cognitive and meta-cognitive strategies influence rating accuracy in essay scoring. *Assessing Writing, 27*, 37–53. https://doi.org/10.1016/j.asw.2015.11.001

Zhang, S., & Thompson, N. (2004). DIALANG: A diagnostic language assessment system (review). *The Canadian Modern Language Review, 6*(2), 290–293. https://doi.org/10.1353/cml.2005.0011

Ziegler, N., & Kang, L. (2016). Mixed methods designs. In A. J. Moeller, J. W. Creswell, & N. Saville (Eds.), *Second language assessment and mixed methods research* (pp. 51–83). Cambridge University Press.

Appendix A. First-Draft Diagnostic Rating Scale

Essay Topic: _____**Student ID**: _____ **Classroom Section**: ____

Instruction: Tick "**0"** if an essay has no evidence or does not satisfy the descriptor requirement or tick "**1"** if an essay satisfies all the descriptor requirements.

Subskills	No.	Descriptors	0	1
1. Organisation				
• Introduction paragraph	01	The introduction paragraph introduces the topic of the prompt.		
	02	The introduction paragraph attracts the readers' interest.		
	03	The introduction paragraph states a thesis that responds to the prompt.		
• Main body paragraph	04	The topic sentences state the topic related to the thesis statement.		
	05	The topic sentence has a specific controlling idea guiding supporting ideas.		
	06	Supporting ideas are given to support the topic sentence.		
	07	The paragraph concluding sentence restates the topic sentence in different words.		
	08	The paragraph concluding sentence summarises supporting points in different words.		
• Conclusion paragraph	09	The conclusion paragraph restates the thesis in different words.		
	10	The conclusion paragraph summarises all main ideas in different words.		
	11	The conclusion paragraph ends an essay with a final thought.		
2. Coherence				
• Within paragraph	12	Supporting ideas relate to a single main idea in a paragraph.		
	13	Supporting ideas are sufficient to support a main idea in a paragraph.		
	14	The main idea in a paragraph relates to the thesis statement.		
• Within essay	15	All main ideas in an essay relate to a single topic.		
	16	All main ideas in an essay relate to the thesis statement.		
	17	All main ideas in an essay are unique.		
3. Cohesion				
• Within paragraph	18	Supporting ideas in a paragraph are arranged logically.		
	19	Supporting ideas in a paragraph are linked by transition signals.		
• Within essay	20	All main ideas in an essay are arranged logically.		
	21	All main ideas in an essay are linked by transition signals.		
4. Content				
• Understanding*	22	Content is understandable enough.		
• Redundancy	23	Content is not redundant.		
• Logic***	24	Content is logical.		
• Completion	25	Content meets all prompt requirements.		

Subskills	No.	Descriptors	0	1
5. Grammar				
• Part of speech	26	Accurate part of speech is used.		
• Subject-verb agreement	27	Subject-verb agreement is used accurately.		
• Tense	28	Tenses are used appropriately.		
• Passive voice	29	Passive voice is used accurately.		
• Transition signals	30	Transition signals are used accurately.		
• Article	31	Articles are used with nouns accurately.		
• Pronoun	32	Pronouns are used with noun phrases accurately.		
• Parallel	33	Parallel structure is used appropriately.		
6. Sentence				
• Simple sentence	34	Simple sentences are built accurately.		
• Compound sentence	35	Compound sentences are built accurately.		
• Complex sentence	36	Complex sentences are built accurately.		
• Variety	37	Various types of sentences are used.		
7. Vocabulary				
• Choice	38	Words are used appropriately for contexts.		
• Variety	39	Various words (e.g. synonyms, word types) are used.		
• Collocations	40	Collocations are used appropriately.		
8. Mechanics				
• Punctuation	41	Punctuations are used accurately.		

**** Transition signals*** *include any connectors, transitions, or phrases that link ideas, words, phrases, sentences, or paragraphs.*

Other comments

_____
_____
_____
_____
_____

Number of unsatisfied skills: _____
Number of satisfied skills: _____

Teacher Name: _____
Date of Rating: _____

Appendix B. Revised Diagnostic Rating Scale

Essay Topic: _____**Student ID**: _____ **Classroom Section**: ____

Instruction: Tick "**0"** if an essay has no evidence or does not satisfy the descriptor or tick "**1"** if an essay satisfies all the descriptor requirements.

Domains		Descriptors	1	0
Organisation				
• Introduction paragraph	01	The introduction paragraph introduces the topic of an essay.		
	02	The introduction paragraph states the thesis that responds to a prompt.		
• Main body paragraph	03	The body paragraph has the topic sentence related to the topic.		
	04	The body paragraph has the topic sentence specifying a controlling idea.		
	05	The body paragraph has supporting ideas related to the topic sentence.		
• Conclusion paragraph	06	The conclusion paragraph restates the thesis in different words.		
	07	The conclusion paragraph summarises all main ideas in different words.		
	08	The conclusion paragraph signals the end of an essay appropriately.		
Coherence				
• Within a paragraph	09	Supporting ideas relate to a single main idea in a paragraph.		
	10	Supporting ideas are enough to support the main idea in a paragraph.		
	11	The main idea in a paragraph relates to the thesis statement.		
• Within an essay	12	All main ideas in an essay relate to a single topic.		
	13	All main ideas in an essay relate to the thesis statement.		
Cohesion				
• Within a paragraph	14	Supporting ideas in a paragraph are arranged appropriately or logically.		
	15	Supporting ideas in a paragraph are linked by transition signals.		
• Within an essay	16	All main ideas in an essay are arranged appropriately or logically.		
	17	All main ideas in an essay are linked by transition signals.		
Content				
• Comprehension	18	Content is understandable enough.		
• Completion	19	Content meets all prompt requirements.		
• Length	20	The essay has an appropriate length.		
Grammar use				
• Part of speech	21	Accurate part of speech is used.		
• Subject-verb agreement	22	Subject-verb agreement is used accurately.		
• Tense and voice	23	Tense and passive voice are used appropriately.		
• Transition signals	24	Transition signals are used accurately.		

Domains		Descriptors	1	0
• Article	25	Articles are used with nouns accurately.		
• Pronoun	26	Pronouns are used with noun phrases accurately.		
• Parallel	27	Parallel structure is used appropriately.		
Sentence use				
• Simple sentence	28	Simple sentences are built accurately.		
• Compound sentence	29	Compound sentences are built accurately.		
• Complex sentence	30	Complex sentences are built accurately.		
• Sentence variety	31	Various types of sentences are used.		
Vocabulary use				
• Choice	32	Words are used appropriately for contexts.		
• Variety	33	Various words (e.g. synonyms, word types, difficult words) are used.		
• Collocation	34	Collocations are used appropriately.		
Mechanic use				
• Punctuation	35	Punctuations are used accurately.		
• Capitalisation	36	Capitalisation is used accurately.		
• Spelling	37	Spelling is accurate.		

**** Transition signals*** *include any connectors, transitions, or phrases that link ideas, words, phrases, sentences, or paragraphs.*

Other comments

_____
_____
_____
_____
_____

Number of unsatisfied skills: _____
Number of satisfied skills: _____

Teacher Name: _____
Date of Rating: _____

Appendix C. Finalised Diagnostic Rating Scale

Essay Topic: _____**Student ID**: _____ **Classroom Section**: _____

Instruction: *Tick (✔) "**0**" if an essay **has no evidence or does not satisfy** the descriptor or "**1**" if an essay **satisfies** the descriptor.*

Domains		Descriptors	0	1
1. Organisation				
• Introduction paragraph	01	The introduction paragraph introduces the topic of an essay.		
	02	The introduction paragraph states the thesis that responds to a prompt.		
• Body paragraph	03	The body paragraph has the topic sentence related to the thesis statement		
	04	The body paragraph has the topic sentence specifying the topic and controlling idea		
	05	The body paragraph has supporting ideas related to the topic sentence.		
• Conclusion paragraph	06	The conclusion paragraph restates the thesis in different words.		
	07	The conclusion paragraph summarises all main ideas in different words.		
	08	The conclusion paragraph ends an essay appropriately.		
• Essay length	09	The essay has an appropriate length.		
2. Coherence				
• Paragraph coherence	10	All supporting ideas in a body paragraph relate to the single main idea or topic sentence.		
	11	All supporting ideas in a body paragraph are convincing and enough.		
• Essay coherence	12	All main ideas in all body paragraphs relate to the thesis statement.		
3. Cohesion				
• Paragraph cohesion	13	All supporting ideas in a body paragraph are arranged appropriately.		
	14	All supporting ideas in a body paragraph are linked by appropriate transition signals.		
• Essay cohesion	15	All main ideas in an essay are arranged appropriately.		
	16	All main ideas in an essay are linked by appropriate transition signals.		
4. Content				
• Comprehension	17	Content is understandable enough.		
• Completion	18	Content meets all prompt requirements.		
• Distribution	19	Contents in all paragraphs are well-balanced.		
5. Grammar				
• Part of speech	20	Part of speech is used accurately or with a few errors.		
• Subject-verb agreement	21	Subject-verb agreement is used accurately or with a few errors.		
• Tense and voice	22	Tense and voice are used accurately or with a few errors.		
• Article	23	Articles are used accurately or with a few errors.		
• Pronoun	24	Pronouns are used accurately or with a few errors.		

Domains		Descriptors	0	1
6. Sentence				
• Simple sentence	25	Simple sentences are used.		
• Compound sentence	26	Compound sentences are used.		
• Complex sentence	27	Complex sentences are used.		
• Sentence problem	28	Sentences are built accurately or with a few errors (e.g., fragment and run-on).		
7. Vocabulary				
• Choice	29	Words are used appropriately for contexts.		
• Variety	30	Various words (e.g. synonyms, word types, difficult words) are used.		
8. Mechanics				
• Punctuation	31	Punctuations are used accurately or with a few errors.		
• Capitalisation	32	Capitalisation is used accurately or with a few errors.		
• Spelling	33	Words are spelled accurately or with a few errors.		

*** *Transition signals*** *include any connectors, transitions, or phrases that link ideas, words, phrases, sentences, or paragraphs.*

Other comments

_____
_____
_____
_____
_____

Number of unsatisfied skills: _____
Number of satisfied skills: _____

Teacher Name: _____
Date of Rating:_____

Appendix D. Scale Evaluation Form

Domains		Descriptors	Is the descriptor relevant to the course contents?	Is the descriptor taught in the course?	Is the descriptor suitable for criteria inclusion?	Other Comment
Organisation						
• Introduction paragraph	01	The introduction paragraph introduces the topic of an essay.				
	02	The introduction paragraph states the thesis that responds to a prompt.				
• Main body paragraph	03	The body paragraph has the topic sentence related to the topic.				
	04	The body paragraph has the topic sentence specifying a controlling idea.				
	05	The body paragraph has supporting ideas related to the topic sentence.				
• Conclusion paragraph	06	The conclusion paragraph restates the thesis in different words.				
	07	The conclusion paragraph summarises all main ideas in different words.				
	08	The conclusion paragraph signals the end of an essay appropriately.				
Coherence						
• Within a paragraph	09	Supporting ideas relate to a single main idea in a paragraph.				
	10	Supporting ideas are enough to support the main idea in a paragraph.				
	11	The main idea in a paragraph relates to the thesis statement.				
• Within an essay	12	All main ideas in an essay relate to a single topic.				
	13	All main ideas in an essay relate to the thesis statement.				
Cohesion						
• Within a paragraph	14	Supporting ideas in a paragraph are arranged appropriately or logically.				
	15	Supporting ideas in a paragraph are linked by transition signals.				
• Within an essay	16	All main ideas in an essay are arranged appropriately or logically.				
	17	All main ideas in an essay are linked by transition signals.				
Content						
• Comprehension	18	Content is understandable enough.				
• Completion	19	Content meets all prompt requirements.				
• Length	20	The essay has an appropriate length.				
Grammar use						
• Part of speech	21	Accurate part of speech is used.				
• Subject-verb agreement	22	Subject-verb agreement is used accurately.				

Domains		Descriptors	Is the descriptor relevant to the course contents?	Is the descriptor taught in the course?	Is the descriptor suitable for criteria inclusion?	Other Comment
• Tense and voice	23	Tense and passive voice are used appropriately.				
• Transition signals	24	Transition signals are used accurately.				
• Article	25	Articles are used with nouns accurately.				
• Pronoun	26	Pronouns are used with noun phrases accurately.				
• Parallel	27	Parallel structure is used appropriately.				
Sentence use						
• Simple sentence	28	Simple sentences are built accurately.				
• Compound sentence	29	Compound sentences are built accurately.				
• Complex sentence	30	Complex sentences are built accurately.				
• Sentence variety	31	Various types of sentences are used.				
Vocabulary use						
• Choice	32	Words are used appropriately for contexts.				
• Variety	33	Various words (e.g. synonyms, word types, difficult words) are used.				
• Collocation	34	Collocations are used appropriately.				
Mechanic use						
• Punctuation	35	Punctuations are used accurately.				
• Capitalisation	36	Capitalisation is used accurately.				
• Spelling	37	Spelling is accurate.				

Appendix E. Coding Guideline

Definitions of the coding categories	Subcategories	Example quotes
Scale functioning		
The scale comprehensibility refers to how the scale criteria affect the way in which raters comprehend or judge the descriptors, thus subcategorised into the criteria clarity and criteria judgement. • A coding unit belongs to the **criteria clarity** if any responses of a participant indicates or implies that the descriptors ***are or are not clear*** to understand. • A coding unit belongs to the **criteria judgement** if any responses of a participant indicates or implies that the descriptors ***are easy or difficult*** to judge. The criteria clarity focuses on the linguistic understanding of the scale criteria while the criteria judgement focuses on the rater decision-making or judgement on the descriptors. **Do not apply** this category if a participant expresses that the scale properties are easy or difficult to use or follow and this is instead applied to the scale applicability.	1. Criteria clarity	*Ivey: [Are the scale descriptors easy to understand? If not, specify descriptors that were ambiguous or not clear?] Yeah, it is easy to understand.*
	2. Criteria judgement	*Sara: [Any descriptors you think are ambiguous or not clear or you find it difficult to judge?] Number 17 (content is understandable enough). [Ok why is that?] Em, as a Thai teacher, yes I understand what they try to tell the audience. If a foreigner, native-English speaker, have to tick ZERO or ONE on Number 17, it might be difficult for them [So, when you rated the essays, you kind of think of native-speaker readers right?] Yes.*
The scale comprehensiveness refers to how well the scale criteria/scoring capture specific, discrete writing skills and detailed writing quality and represent learning contents and assessment criteria, thus subcategorised into to the criteria specificity and criteria coverability. • A coding unit belongs to the **criteria specificity** if any expressions of a participant indicates or implies that the scale criteria/scoring ***capture or do not capture*** specific, discrete writing skills and detailed writing quality • A coding unit belongs to the **criteria coverability** if any expressions of a participant indicates or implies that the scale criteria/scoring ***cover or does not cover*** essential learning contents and assessment criteria in the classroom. **Do not apply** this category if a participant expresses that the rating format affects the rater application of the scale and this is instead applied to the scale applicability.	3. Criteria specificity	*Nana: [Do you think the descriptors are specific enough to capture I mean detailed skills or several writing skills?] I think yes.*
	4. Criteria coverability	*Ivey: [Do you think that the rating scale cover all productive skills and textual features of the expository writing in classroom? If no, specify descriptors that were irrelevant to expository writing] I think it all covers essential skills according to the course objectives, for example five paragraph essay, what is a paragraph.*

Definitions of the coding categories	Subcategories	Example quotes
The scale applicability refers to how well the scale criteria and properties are organised and structured in the way that facilitates raters' application of the scale and it is subcategorised into the scale organisation and rating format. • A coding unit belongs to this the **scale organisation** if any expressions of a participant indicates or implies that the scale criteria or properties (e.g., scale layout, scale criteria, scale length) ***are or are not*** well organised, arranged, or ordered. • A coding unit belongs to this the **rating format** if any expressions of a participant indicates or implies that the binary rating ***is or is not*** well organised, arranged, or ordered. The scale applicability focuses on the impact of the structure, organisation, and arrangement of the scale properties on the application of the scale rather than the clarity and judgement of the scale criteria.	5. Scale organisation	*Cali: For me, I think it's better if everything will be on the same page because sometimes you need to [So that it's easier for you because you don't have to flip] yeah.*
	6. Rating format	*Ken: [Is the judgement or scoring of the scale descriptors appropriate? I mean only two options ZERO and ONE or strong and weak] Yes, it is I think you have or we have more it's gonna be complicated. Two is enough.*
Assessment usefulness		
The teaching usefulness refers to the extent to which the diagnostic rating scale can provide any useful information that generally supports ongoing teaching and learning in the classroom and it is divided into seven subcategories: diagnostic information, diagnostic feedback, student improvement, diagnostic result report, summative assessment, teaching guideline, and scale practicality. • A coding unit belongs to the **diagnostic information** if any expressions of an interviewee indicates or implies that the scale provide diagnostic information which ***is or is not*** indicative of writing strengths and weaknesses. • A coding unit belongs to the **diagnostic feedback** if any expressions of an interviewee indicates or implies that the scale provide diagnostic information which ***is or is not*** supportive of diagnostic feedback quality. • A coding unit belongs to the **student improvement** if any expressions of an interviewee indicates or implies that the	7. Diagnostic information	*Ivey: [Anything else?] Students often use spelling and punctuation incorrectly, but I understand that and when I gave feedback, I talked about this but not that much. Yeah, it's interesting. If we don't have the scale, we will not see these mistakes.*
	8. Diagnostic feedback	*Nana: When I give students' feedback, I always show I mean I show this with the written comments in the students' essays together like this. So, I show them and let them see what point they loose and what point they gain so that the students can see clearly which parts or which skills they should put more focus on or improve more on that. So, I think it's useful when I use this with students' writing and when I gave students' feedback, so I can like group and clearly show students the quality of their writing ability, writing skill in each essay.*
	9. Student improvement	*Ivey: When they write the first and second drafts, they did not do well on the first draft but when I rated the second draft I can see that they improved though it is not as good as I expected. This shows that students listened to feedback and then go back to revise their writing. So many students did better on the second draft better than the first draft though there are some mistakes.*
	10. Diagnostic report	*Nana: [Anything else you want to say in terms of the scale help you to improve your teaching?] Yeah, I mentioned it already. So, I looked at the scale I gave to*

Definitions of the coding categories	Subcategories	Example quotes
scale provide diagnostic information which ***is or is not*** supportive of student writing improvement. • A coding unit belongs to the **diagnostic result report** if any expressions of an interviewee indicates or implies that the scale provide diagnostic results which ***is or is not*** supportive of the meaningful interpretation of diagnostic result. • A coding unit belongs to the **summative assessment** if any expressions of an interviewee indicates or implies that the scale ***is or is not*** supportive of the summative assessment. • A coding unit belongs to the **teaching guideline** if any expressions of an interviewee indicates or implies that the scale ***is or is not*** supportive of teaching preparation or activities. • A coding unit belongs to the **scale practicality** if any expressions of an interviewee indicates or implies that the scale ***is or is not*** practical for multiple-round assessment		*the students each time and I am not a kind of statistic person but if I can do it I can like put it on the programme so that we know what are the weakness of the students. [You mean there should be something like a report] Yeah so that we know right so we can analyse right the students' strengths weakness whatever so it would be very useful for. [Yeah I'll will produce that report I mean diagnostic profile reports] Yeah, ah profile report right it will be useful for every teacher.*
	11. Summative assessment	*Sara: [Even though you used another scale for the midterm and final exams, do you think at some point the criteria in the diagnostic scale helped you to kind of better judge midterm exams even though you used another rating scale] Yeah it reminded me actually most of the items on your scale yeah, we include them in another criteria we use yes. [But the descriptions] is different a bit different.*
	12. Teaching guideline	*Cali: You know when Teacher Ken and I received this scale, we teach exactly like the scale because we want students to learn exactly like the scale [So, the scale is part of your teaching materials] Yes, it's part of our teaching materials. We talked to each other if we want students to get ONE for all of these, what kind of things that we should teach them. We train them on each point according to this scale.*
	13. Scale practicality	*Nana: But I found out that maybe it's too many descriptors here but another but again because I do understand that you have to cover every writing skill right you provide a lot of descriptors and information here but in terms of I mean it's good but in terms of practical [The number of descriptors is too many?] Yes, too many yes.*
The learning usefulness refers to the extent to which the scale is useful and provide useful information for supporting students' self-learning, self-assessment, and writing development from the teachers' perspectives. It is subcategorised into the self-assessment and self-regulation and writing development. • A coding unit belongs to the **self-assessment and self-regulation** if any expressions of an interviewee indicates or implies that the scale ***is or is not*** useful to support students' self-learning and self-assessment activities. • A coding unit belongs to the **writing development** if any expressions of an interviewee indicates or implies that the scale ***is or is not*** useful to promote students' writing development.	14. Self-assessment and self-regulation	*Ken: Em, I think descriptors in my opinion because students have this with them and then they have to follow all the descriptors. So, this might have something in their mind yes, for example, the introduction paragraph, they should have a clear topic they should have a clear thesis statement for their writing or for their introduction.*
	15. Writing development	*Cali: [Do you think that the self-diagnostic assessment helped and or hindered the students' writing improvement? If so, why and how?] Sure, I think it does not hinder it helped the students a lot because you know when they know their goal, it's easy for them to reach the goal.*

Definitions of the coding categories	Subcategories	Example quotes
Assessment impact		
The awareness raising refers to how the use of the diagnostic rating scale impacts the teachers' awareness and it is subcategorised into assessment fairness, self-assessment, and feedback. • A coding unit belongs to the **assessment fairness** if any expressions of an interviewee indicates or implies that the use of the scale raise any awareness about assessment fairness or transparency. • A coding unit belongs to the **self-assessment** if any expressions of an interviewee indicates or implies that the use of the scale raises any awareness about self-assessment. • A coding unit belongs to the **feedback** if any expressions of an interviewee indicates or implies that the use of the scale raises any awareness about feedback.	16. Assessment fairness	*Ivey: Ah I think it is good guideline to develop a scale, but it should be adjusted according to the teaching course and subject so that it is useful for teachers and learners and the teacher team. So, when we use the same scale, there will be no question about bias judgement or score assignment. [Anything else?] no.*
	17. Self-assessment	*Ken: [Should self-assessment of writing be used in classroom teaching and assessment?] You mean [The way students rate their own essays] Yes yes. [Why can you explain a little bit more about this] If we want to do something like try to ah ah ride a motorbike and then you don't know how but if there is a manual for you to follow. I think it's the same thing with that I think if they want to have good writing, they should have this criteria they should have these descriptors as a guideline for them to follow.*
	18. Feedback	*Nana: I think that feedback is very important after I've been through this project, I think feedback is very important, both teacher feedback and student feedback and I look forward to seeing the report coz the result would be juicy right. [You mean the diagnostic outcomes score report] Right yeah, I really want to see coz that would be very useful for my future class [teaching preparation] right yeah right.*
The future plan refers to how the use of the scale impacts the teachers' future plan which contribute to instructional and other professional development and it is subcategorised into the scale adaptation and professional development. • A coding unit belongs to the **scale adaptation** if any expressions of an interviewee indicates or implies that the use of the scale makes the interviewee interested in, want to, or plan to use or modify the scale for future teaching. • A coding unit belongs to the **professional development** if any expressions of an interviewee indicates or implies that the use of the scale makes the interviewee come up with any ideas to improve future teaching or have any interest in doing research.	19. Scale adaptation	*Sara: My writing course I think in the future, if I have a chance to teach expository or argumentative writing again, I would use the scale and give it to my students and explain some major points they need to acquire.*
	20. Professional development	*Nana: Even this class, I know that the student they lack the models, the examples, coz I learned that from using this rating scale. So, I know that the students they lack they know how to write individually they know how to write but they don't know how to em student were focused predominantly on the element of the essay but lack good essay examples or they did not much analyse the whole input essay or analyse the elements in good essay models [So do you think it's important for students to analyse good essays models or learn from the characteristics of good essays] Yes coz normally we always teaches separate elements of essays like each part of the essay [you mean you focus on teaching the skills necessary to write] right receptive skills.*

Appendix F. Teacher Perception Interview

I would like to thank you very much for your participation in this interview. I would like to ask you some questions about your perception of the scale use and your reflection on participation in this research. This interview should take about 30 minutes and it will be recorded to make it easier for me to concentrate on what you say. Are you okay with this? (ask consent from the participants).

Scale usability (focus on scale characteristics)

1. Is the scale appropriate for identifying students' writing strengths and weaknesses in an ongoing classroom instruction?
2. Do you think the scale is user-friendly for classroom assessment?
3. Are the scale descriptors easy to understand? If not, specify descriptors that were ambiguous or not clear.
4. If you want to improve the rating scale for diagnostic and practical purposes in classroom, what would you want to improve the most? Why/how?
5. Is the judgement/scoring of the scale descriptors appropriate? If not, explain why?
6. If you have any positive or negative comments about the use of the rating scale please tell me.

Teaching/ learning contents

7. Do you think that the rating scale cover all productive skills and textual features of the expository writing in classroom? If no, specify descriptors that were irrelevant to expository writing
8. Are there any particular descriptors that you think most or least important in developing students' expository writing?
9. Do the diagnostic rating criteria target all the writing skills/contents you teach in class?

Teaching and assessment

10. Do you think that the diagnostic rating scale provides useful information for improving the way you teach expository writing? If so, why and how?
11. Do you think that the diagnostic rating scale provides useful information for improving the way you assess students' expository writing? If so, why and how?

Diagnostic feedback

12. Do you think that the diagnostic rating scale provides useful diagnostic information about the strengths and weaknesses of students' expository writing? If so, why/how?
13. How did you use the diagnostic rating scale as part of your feedback?
14. In your opinion, what influenced how students receive and use feedback?

Self-assessment of writing

15. Should self-assessment of writing be used in classroom teaching and assessment?
16. Do you think that the self-diagnostic assessment helped and/or hindered the students' writing improvement? If so, why/how?
17. Do you think that the diagnostic rating scale is useful for students' self-assessment of writing? If so, why/how?

Study experience and reflection

18. After participating in the study, how do you feel about the research project and are there any changes in your writing instruction, and assessment, and any other aspects you have made because of this experience?

Appendix G. Student Self-Assessment Interview

I would like to thank you very much for your voluntary participation in this interview. I would like to ask you some questions about your self-assessment of your own essays in order to better understand, for example, how you went about it, how you rated your own essays, what you found difficult and other questions related to your self-assessment of writing. This interview should take about 30 minutes and it will be recorded to make it easier for me to concentrate on what you say. Are you okay with this? (ask consent from the participants).

1. Have you ever done self-assessment of writing in any courses before? If so, where, when and how did you do this?
2. After the self-rating training, did you better understand how to use the diagnostic rating scale to rate your essay?
3. How much time on average did you put into your self-assessment of each essay?
4. Did you understand the assignment task instruction?
5. Did you reread the assignment task instruction? If so, why?
6. Did you try to understand the essay topic?
7. Did you reread the essay topic? If so, why?
8. Did you read the rating scale descriptors before you rated your essay?
9. Were you thinking about the rating scale descriptors as you rated your essay?
10. Did you read the rating scale descriptors again while rating your essay? If so, why?
11. Did you think of the self-rating training paper as the model for rating your essay?
12. Did you change your rating decision on any descriptors? If so, what influenced you to change a particular score?
13. Were there any particular rating scale descriptors you found difficult to understand and judge? If yes, please identify:
14. Did the rating scale descriptors help you to know your strengths and weaknesses in writing?
15. What did you do after you knew your writing strengths and weaknesses? For example, did you compare your writing strengths and weaknesses with teachers' rating results and feedback or did you use the information to improve your learning and writing skills?
16. Do you think the diagnostic rating scale is useful for self-assessment of your essay?
17. Do you think the diagnostic rating scale is user-friendly for self-assessment?
18. Do you think self-assessment should be included in writing classroom?
19. If you want to change or improve the diagnostic rating scale, what do you want to improve and why?
20. Any other opinion or comments?

Appendix H. Student Perception Interview

I would like to thank you very much for your participation in this interview. I would like to ask you some questions about your perception on the use of the diagnostic rating scale for your self-assessment and your reflection on research participation. This interview should take about 30 minutes and it will be recorded to make it easier for me to concentrate on what you say. Are you okay with this? (ask consent from the participants).

1. Do you think that the use of the diagnostic rating scale for self-assessment helped you improve your writing ability?
2. Do you think that the use of the diagnostic rating scale for self-assessment helped you reflect on your writing strengths and weaknesses?
3. Do you think that the use of the diagnostic rating scale for self-assessment helped you become more engaged and motivated in writing learning?
4. Did you compare your self-rating results with teacher rating results? If so, were the results similar or different on the whole?
5. Did you find your teacher's feedback from the teacher rating scale helpful? How and give examples?
6. During the writing course, did you feel that you were motivated or involved in learning and assessment processes?
7. After participating in this research, how do you feel about the research project and are there any changes in your writing learning and improvement and any other aspects that happened because of this research?

Appendix I. Teacher Background Questionnaire

The purpose of this questionnaire is to collect your background information for research purposes. The study aims to develop and validate a diagnostic English writing rating scale for university classroom diagnostic writing assessment. Please note that the aim of the research is not to judge your performance and the information you provide will remain confidential. Your identity will remain confidential. Please answer all of the following questions to the best of your ability.

1) Full name: _____

2) Preferred pseudonym: _____

3) Age: _____years or ☐ 25-30 years ☐ 31-35 years ☐ 36-40 years ☐ 41-45 years

4) Gender: ☐ Male ☐ Female

5) First language _____

6) Additional language(s): _____

7) Educational background (If you are pursuing, please specify after a degree)

- Bachelor's degree in _____Country _____
- Master's degree in _____Country _____
- Doctoral degree in _____ Country _____
- Other degree _____Country _____
- Other certificate _____Country _____
- Other certificate _____Country _____
- Other certificate _____Country _____

8) Have you ever taught any English writing courses before? ☐ No ☐ Yes
If yes, please specify the courses: _____
_____

9) How many times have you taught the Expository Writing course? _____ Years

10) Have you ever been trained in rating writing or essays? ☐ No ☐ Yes
If yes, please specify: Year(s) _____, Total hours: _____ or minutes: _____

11) Are there any assessment experiences that you think might have influenced your English writing assessment or scoring judgement? ☐ No ☐ Yes
If yes, please specify: _____

Thank you very much for your time and responses!

Appendix J. Student Background Questionnaire

This questionnaire aims to collect your background information which will be used for research purposes. The research aims to develop and validate a diagnostic rating scale for an EFL university writing classroom. The information you provide will only be used for research purposes and your name will be kept confidential. Please answer all of the following questions to the best of your ability.

1) Full Name: _____ 2) Student ID: _____
3) Major: _____ 4) Section: _____
5) Age: _____ years 6) Gender: ☐ Male ☐ Female
7) How long have you learned English so far? _____ years
8) Have you ever studied English abroad? ☐ No ☐ Yes
If yes, please specify:

Year	Courses	Country	Approximate total hours

9) Have you ever taken any English writing courses before? ☐ No ☐ Yes
If yes, please specify:

Year	Courses	Where	Approximate total hours

10) How often do you write in Thai?
☐ Every day ☐ Three times a week ☐ Once a week ☐ Less than once a week

11) How often do you write in English?
☐ Every day ☐ Three times a week ☐ Once a week ☐ Less than once a week

12) Have you ever taken any of the following English tests? ☐ No ☐ Yes
If yes, please specify and if possible specify the score:
☐ MSU-EXIT Year _____ Score _____ ☐ IELTS Year _____ Score _____
☐ CU-TEP Year _____ Score _____ ☐ TOEFL iTP Year _____ Score _____
☐ TU-GET Year _____ Score _____ ☐ TOEFL iBT Year _____ Score _____
☐ TOEIC Year _____ Score _____

13) How would you assess your overall English language proficiency? (Circle one)

Basic									*Very proficient*
1	*2*	*3*	*4*	*5*	*6*	*4*	*8*	*9*	*10*

14) How would you assess your overall English writing proficiency? (Circle one)

Basic									*Very proficient*
1	*2*	*3*	*4*	*5*	*6*	*4*	*8*	*9*	*10*

Thank you very much for your time and responses!

Appendix K. Characteristics of Writing Assignment Tasks

0105306 English Expository and Argumentative Composition

Classroom A1 (Sara)
Task 1: Expository essay (cause-and-effect essay).

- Which one would you choose between losing both legs and losing both hands? As a person with disabilities, what would you do to make the world a better place? Use specific reasons to support your idea.

Task 2: Expository essay (problem-solution essay). Choose one of the following topics.

1. As countries become more industrialized, more people move from the countryside into urban areas. What are some problems caused by this and how can we solve them?
2. With easier access to the internet, many students are relying heavily on online sources instead of libraries. State some of the problems caused by this and the methods to address them.
3. With the advent of the internet, an increasing number of people are shopping online. What issues arise from this and how they can be tackled?
4. More and more families are choosing fast food over home-cooked meals. State the possible problems that may occur from this. What are your solutions to this problem?
5. Movies and TV programs have become saturated with violent content. What are some social problems resulting from this and how they can be dealt with?

Task 3: Argumentative essay (opinion essay). Do you agree or disagree with the following statement? Use specific details or examples to support your answer. Choose one of the following topics.

1. All strayed dogs should be killed to stop the breakout of rabies in our country.
2. Gay marriage should be legalized in the Thai society.
3. Mobile phones should be banned when driving.
4. School uniforms should no longer be mandatory as students should have to the right to choose their own style of clothing or use fashion to express themselves.
5. More people view that living together before marriage isn't as a taboo as it used to be and it is a good way to practice.

Classroom A2 (Nana)
Task 1: Expository essay (cause-and-effect essay). Choose ONE topic.

1. Factors driving the Thai political crisis
2. What causes Google to be the most popular search engine?
3. What are the causes and effect of insomnia during exam week?
4. What are the reasons of popularity of Korean pop singer and what are the effects of the K-POP on teenagers?
5. Reasons why online shopping makes internet users spend more money

Task 2: Expository essay (problem-solution essay). Choose ONE topic.

1. Global warming is one of the biggest threats humans face in the 21st Century and sea levels are continuing to rise at alarming rates. What problems are associated with this and what are some possible solutions.
2. With easier access to the internet, many students are turning to online sources to study instead of libraries. State some of the problems caused by this and methods to address them.
3. In some countries the average weight of people is increasing and their levels of health and fitness are decreasing. What do you think are the causes of these problems and what measures could be taken to solve them?

4. Many small, local shops are closing as they are unable to compete with the big supermarkets in the area. How does this affect the community? How could this situation be improved?

Task 3: Argumentative essay

- Choose your own topic (e.g., Should We Stop Eating Instant Noodles?; Should Student Take A Gap Year?)

0105333 English Expository Composition Writing

Classroom B1 (Ivey)

Task 1: Descriptive essay

- My life in English Major in the Faculty of Education at Mahasarakham University

Task 2: Process essay

- Choose your own topic (Example topic from student essays: *A Process of Preparing Your Face Before Wearing Makeup; The Process of Planting Rose*)

Task 3: Compare-contrast

- Choose your own topic (Example topic from student essays: *The Differences Between Fans and Air Conditioners; What is the Better Places to Live Between Countryside and City*)

Classroom B2 (Ken and Cali)

Task 1: Descriptive

- Choose your own topic (Example topic from student essays: *The Best Places for One Day Trip; Sakon Nakhon, the City of Ancient Civilization*)

Task 2: Cause/effect

- Choose your own topic (Example topic from student essays: *Effects of Skipping Breakfast; Effect of Meditation*)

Task 3: Compare/contrast

- School Life versus University Life

Appendix L. FACETS Specification File

```
Title        = TASK123COURSEAB
Facets       = 5
Inter-rater  = 1
Positive     = 2,3
Noncentered  = 2
Unexpected   = 2
Vertical     = 1A,2A,3A,4A,5A,S
Arrange      = aN
Pt-biserial  = Measure
Models       = ?,?,?,?,?,D
*

Labels=
1, Teacher  ; (element = 6)
1=1S
2=2N
3=3I
4=4K
5=5C
6=6A
*
2, Student  ; (element = 80)
01=01L
02=02L
03=03M
04=04M
05=05M
06=06M
07=07M
08=08M
09=09M
10=10L
11=11M
12=12L
13=13L
14=14L
15=15L
16=16L
17=17M
18=18L
19=19L
20=20M
21=21H
22=22L
23=23M
24=24H
25=25L
26=26H
27=27H
28=28L
29=29H
30=30L
31=31H
32=32L
33=33M
34=34L
35=35H
36=36L
37=37M
38=38L
39=39L
40=40L
41=41L
42=42L
43=43M
44=44M
45=45M
46=46L
47=47M
48=48L
49=49L
50=50L
51=51L
52=52L
53=53L
54=54L
```

```
55=55L
56=56M
57=57M
58=58M
59=59L
60=60M
61=61M
62=62H
63=63M
64=64M
65=65H
66=66M
67=67H
68=68M
69=69M
70=70H
71=71M
72=72H
73=73H
74=74M
75=75M
76=76M
77=77H
78=78M
79=79H
80=80M
*
3, Task     ; (element = 3)
1=T1
2=T2
3=T3
*
4, Descriptor , G      ; (element = 33)  group-anchor
1=01OR    ,0,1         ; items grouped by category number
2=02OR    ,0,1
3=03OR    ,0,1
4=04OR    ,0,1
5=05OR    ,0,1
6=06OR    ,0,1
7=07OR    ,0,1
8=08OR    ,0,1
9=09OR    ,0,1
10=10CR   ,0,2
11=11CR   ,0,2
12=12CR   ,0,2
13=13CS   ,0,3
14=14CS   ,0,3
15=15CS   ,0,3
16=16CS   ,0,3
17=17CT   ,0,4
18=18CT   ,0,4
19=19CT   ,0,4
20=20GM   ,0,5
21=21GM   ,0,5
22=22GM   ,0,5
23=23GM   ,0,5
24=24GM   ,0,5
25=25ST   ,0,6
26=26ST   ,0,6
27=27ST   ,0,6
28=28ST   ,0,6
29=29VC   ,0,7
30=30VC   ,0,7
31=31MC   ,0,8
32=32MC   ,0,8
33=33MC   ,0,8
*
5, Category          ;  , D   ; (element = 8)
1=1OR
2=2CR
3=3CS
4=4CT
5=5GM
6=6ST
7=7VC
8=8MC
*
data=ABT123DATA.xlsx
; enter in format:
```

Appendix M. Supplementary Materials for Qualitative Results

Teacher perceptions		Teachers					
		N	Sara	Nana	Ivey	Ken	Cali
01	Most descriptors are largely clear and understandable.	*5*	✓	✓	✓	✓	✓
02	Some descriptors are difficult to judge.	*4*	✓	✓	-	✓	✓
	· *D03 Topic sentence relevancy*	*1*	-	✓	-	-	-
	· *D11 Supporting idea convincing*	*1*	-	✓	-	-	-
	· *D12 Main idea relevancy*	*1*	-	✓	-	-	-
	· *D17 Content comprehension*	*3*	✓	✓	-	-	✓
	· *D28 Sentence problem*	*1*	✓	-	-	-	-
	· *D29 World choice*	*2*	-	✓	-	✓	-
	· *D30 Word variety*	*2*	-	✓	-	✓	-
03	Criteria capture discrete and specific writing skills.	*5*	✓	✓	✓	✓	✓
04	Error counting does not capture the quality of writing skills.	*2*	-	-	-	✓	✓
05	Binary rating does not capture the granularity of writing skills.	*2*	-	-	✓	-	✓
06	Rating format should have more than two options.	*2*	-	-	✓	-	✓
07	Criteria largely cover core writing skills and learning contents.	*5*	✓	✓	✓	✓	✓
08	Other skills should be assessed added to the criteria.	*5*	✓	✓	✓	✓	✓
	· *Academic language*	*1*	-	-	-	-	✓
	· *Consistent use of English style*	*1*	✓	-	-	-	-
	· *Genre-specific features*	*1*	-	✓	-	-	-
	· *More grammar features*	*1*	-	-	✓	-	-
	· *Overall impression*	*1*	-	-	-	✓	-
	· *Standard English*	*2*	✓	-	-	-	✓
09	The scale layout is largely arranged in a way easy to use.	*5*	✓	✓	✓	✓	✓
10	The scale length should be a single page.	*2*	-	✓	-	-	✓
11	Micro-skill descriptors should come before macro-skill descriptors.	*1*	-	-	-	-	✓
12	Binary rating is largely practical and easy to judge.	*5*	✓	✓	✓	✓	✓
13	More rating options are easier to judge.	*2*	-	-	✓	-	✓
14	The 1-point option should come before 0-point option.	*1*	-	✓	-	-	-
15	The scale provides useful information about students' writing strengths and weaknesses.	*5*	✓	✓	✓	✓	✓

Teacher perceptions		Teachers					
		N	Sara	Nana	Ivey	Ken	Cali
16	The scale provides useful information for detailed and digestible feedback.	*3*	✓	✓	-	-	✓
17	The scale provides useful information about students' writing improvement.	*2*	-	-	✓	-	✓
18	The scale should include a concise report of idividual students' diagnostic profiles.	*1*	-	✓	-	-	-
19	The scale helps to better assess students' summative exam essays.	*5*	✓	✓	✓	✓	✓
20	The scale is useful as teaching resources and guidelines.	*3*	✓	-	✓	-	✓
21	The scale has a lot of descriptors and thus is time-consuming to use in ongoing assessment.	*1*	-	✓	-	-	-
22	The scale is generally useful for self-learning and self-assessment.	*5*	✓	✓	✓	✓	✓
23	The scale is not fully useful for low-proficiency students' self-learning and self-assessment.	*2*	✓	-	✓	-	-
	• *Show low-engagement with learning and feedback.*	*1*	-	-	✓	-	-
	• *Unable to identify strong or weak skills and problems or errors.*	*1*	✓	-	-	-	-
24	Need further support from higher-proficiency peers and teachers.	*1*	✓	-	-	-	-
25	The scale is largely useful for writing development.	*5*	✓	✓	✓	✓	✓
26	The scale is not useful for supporting students' idea development.	*1*	-	-	-	✓	-
27	Assessment should be fair or unbiased to students.	*2*	✓	-	✓	-	-
28	Self-assessment is necessary for students' writing development.	*5*	✓	✓	✓	✓	✓
29	Feedback is important for students' writing development.	*1*	-	✓	-	-	-
30	Teachers want to adopt and adapt the scale for future teaching and assessment.	*4*	✓	✓	✓	✓	-
31	Teachers have new ideas from diagnostic assessment to improve future teaching.	*1*	-	✓	-	-	-
32	Teachers is interested in doing research as inspired by the diagnostic outcomes.	*1*	-	✓	-	-	-

Student perceptions		Students ID																				
		N	03	04	09	12	17	21	23	26	31	32	39	46	54	56	57	61	62	63	64	65
01	Some descriptors are not clear and difficult to understand and judge	*19*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	-	✓	✓	✓	✓	✓	✓	✓	✓	✓
	· *Organisation*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓
	· *Essay length*	*2*	-	-	-	✓	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-
	· *Main idea summary*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
	· *Supporting idea & topic sentence relation*	*2*	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-
	· *Topic sentence specificity*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
	· *Coherence*	*4*	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	✓	-	-

Student perceptions		Students ID																				
		N	03	04	09	12	17	21	23	26	31	32	39	46	54	56	57	61	62	63	64	65
	· *Cohesion*	*4*	-	-	✓	-	-	-	-	-	-	-	-	-	✓	-	✓	✓	-	-	-	-
	· *Supporting idea convincing*	*1*	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-
	· *Main idea arrangement*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-
	· *Transition signals*	*2*	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	✓	-	-
	· *Supporting idea arrangement*	*3*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	✓	✓	-
	· *Main and supporting idea differentiation*	*1*	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-
	· *Content*	*2*	-	-	-	-	-	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-
	· *Content comprehension*	*2*	-	-	✓	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-
	· *Content distribution*	*4*	-	-	-	✓	✓	✓	-	✓	-	-	-	-	-	-	-	-	-	-	-	-
	· *Grammar*	*2*	-	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-
	· *Sentence problem*	*3*	✓	-	-	-	-	-	-	-	✓	-	-	-	✓	-	-	-	-	-	-	-
	· *Sentence variety*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-
	· *Punctuation*	*2*	✓	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-	-
	· *Error counting*	*2*	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
	· *Word variety*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-
	· *Compound sentence*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-
	· *Complex sentence*	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-
02	Binary rating is not detailed and more rating options should be added.	*9*	✓	-	✓	-	-	✓	-	✓		-	-	-	✓	✓	✓	-	✓	-	-	✓
03	Additional criteria should be included in the scale.	*2*	-	-	-	-	-	-	-	✓	-	-	✓	-	-	-	-	-	-	-	-	-
	· *Sources citation*	*1*	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-	-	-	-
	· *Native-like or standard English*	*1*	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-	-
04	The scale is well organised and easy to use.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
05	Micro-skill descriptors should precede macro-skill descriptors.	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-
06	The rating option label should be words instead of numbers.	*2*	-	-	-	-	-	-	-	-	-	✓	✓	-	-	-	-	-	-	-	-	-

Student perceptions		Students ID																				
		N	03	04	09	12	17	21	23	26	31	32	39	46	54	56	57	61	62	63	64	65
07	The scale helps to know and realise writing strengths/weaknesses.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
08	Self-assessment helps to guide and revise writing.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
09	Self-assessment helps to become attentive to or engaged in learning and writing.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
10	Self-assessment motivates to learn and write.	*9*	✓	-	-	-	-	✓	-	✓	-	✓	✓			✓		✓	-	✓	-	✓
11	Scale-assisted self-assessment does motivate to learn and write	*11*	-	✓	✓	✓	✓	-	✓	-	✓	-	-	✓	✓	-	✓	-	✓	-	✓	-
	・ *It's the writing topic that motivates.*	*4*	-	✓	✓	✓	✓	-	-	-	-	-		-	-	-	-	-	-	-	-	-
	・ *It's the teacher that motivates*	*2*	-	-	-	-	-	-	✓	-	-	-	-	-	-	-	-		-		✓	-
12	Peer-assessment should complement self-assessment for unbiased assessment.	*3*	-	✓	✓	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-	-	-
13	Teacher's scale-assisted feedback is useful for writing revision and improvement.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
14	Using the scale for self-assessment helps to improve writing.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
15	The scale should have the overall evaluative description of diagnostic results.	*1*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	✓	-	-	-
16	Self-assessment is useful and should be included in writing classroom.	*20*	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓