Efficacy of Student Assessment as Part of English Writing Instruction for Japanese High School Students

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Education
in the Graduate School of Education

Yoko Oi

Waseda University

2021

Table of Contents

List of Tables

List of Figures

Acknowledgements

I would like to express my deepest appreciation to the following people for their endless empathy, encouragement and advice. Their wide-ranging and generous support ultimately made this doctoral thesis complete.

First, I wish to give my utmost sincere and heartfelt thanks to my supervisor, Professor Dr Yasuyo Sawaki, who endlessly guided me with academic and insightful comments throughout my work on this study. Also, her graceful encouragement has always motivated me to strive for the best. I am also most grateful to Professor Emeritus Dr Michiko Nakano for the opportunity she gave me to embark on this study. Her professional and inspiring comments have always helped me keep the end goal in sight and stay on track.

I would also like to thank Sawaki-seminar and Nakano-seminar for their ongoing encouragement. I have been blessed with such exceptional and generous friendships and cooperation. A special thanks must go to my colleagues, friends and all of the students who participated in the study. This study could not have been completed without their cooperation and participation.

Finally, I will always be deeply grateful to my dear generous husband, Katsuhiro, who constantly encouraged and supported me throughout the study. I am also personally indebted to my family for their faith and affection.

I hope that this study will contribute to shedding light on the improvement of English writing classes, especially for teachers and young English learners.

Chapter 1

INTRODUCTION

1.1Background and Study Aims

Writing in a foreign language is a difficult skill, especially for high school students. This is because writing is a complex process involving communicating with others, building social relationships, changing the writer's social presence, keeping shared meanings and completing social action. Compared with receptive skills such as listening and reading, writing is considered to be a more cognitively complicated task because writing and speaking involve the use of working memory more than reading and listening (Tindle & Longstaff, 2015). Although both speaking and writing are productive skills, their outcomes are different. For example, written production comprises the unit of sentences and needs much greater lexical density and nominalization than spoken production does (Halliday, 2005). Spoken discourse is usually carried out one clause at a time, and longer utterances comprise coordinated clauses with simple linking expressions. Even when spoken discourse consists of many more words, or even clauses, it is not necessarily found to have greater lexical density and grammatical intricacy.

Chafe (1982) pointed out the following common differences between writing and speech. Written language is characterized by "its integrated quality that contrasts with the fragmented quality of spoken language" (p. 38). To be precise, writing generally requires writers to control, plan and monitor the writing process more than with speaking. Writing also tends to require a higher level of accuracy, intricacy and diversity of vocabulary than speaking (Kuiken & Vedder, 2008). Thus, to second language (L2) learners, writing has features that may appear more complicated than speaking.

For many people, writing is closely related to learning in school, and writing performance is evaluated according to criteria enforced in individual educational institutions (Christie & Derewianka, 2008; VanDerHeide & Newell, 2013). Furthermore, most high school students in Japan are in the process of learning English and have not yet completed a regular English educational curriculum. Despite the fact that writing in English is a challenge for them, effective methods of learning and teaching writing in the high school classroom have not yet been established in Japan. In addition, it appears that writing has not been sufficiently emphasized in English classes in Japanese high schools (The Japan Ministry of Education, Culture, Sports, Science and Technology [MEXT], 2015), because students have less interest in learning how to write in English and because only a limited amount of time is allocated to it. It is only natural, therefore, that Japanese high school students who are learning English as a foreign language have difficulties in writing in English. Thus, English writing classes require improvements in terms of their effectiveness in developing students' writing skills.

Another reason why the quality of L2 writing instruction in Japan should be improved is because globalization demands it. The course of study for senior high schools in Japan set out by MEXT stresses the importance of the ability to communicate with others through foreign languages (MEXT, 2013). In this document, writing and speaking skills in particular are considered important requirements for understanding information accurately and conveying ideas appropriately. However, the language proficiency level of English learners in Japan has not attained the national goal. The aim is for more than half of all high school students nationwide to achieve EIKEN Grade Pre-2 (Eiken Foundation of Japan) by 2020 (MEXT, 2013). Grade Pre-2 corresponds to level A2 of the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001), which is defined as the ability level where learners can deal with simple, straightforward information and begin to express themselves in familiar contexts. However, a recent national survey (the English Education Implementation Status Survey [*Eigo kyoiku jisshi chosa*], 2018) conducted by MEXT reported that only 39.3% of 703,380 third-year high school students had achieved this goal, an increase of just 2.9 points from the 2016 survey. Moreover, MEXT's (2018b, 2018c) survey also showed the imbalance in the development of learner proficiency across the four skills: the proportion of high school students who had attained level A2 of the CEFR was 33.6% for listening, while it was 33.5% for reading, 12.9% for speaking and 19.7% for writing. Thus, none of the results for the four skills met the national goal of more than 50% attaining level A2. Moreover, these percentages for productive skills were relatively low. Another issue of concern raised in the survey was the decrease in students' interest in learning to write in English. This was consistent with the results of Goto's (2005) survey, where students indicated a greater interest in acquiring skills in speaking and listening than in reading and writing. It was reported that only 29% of the students surveyed showed a preference for improving their writing skills, while the figure was over 71% for speaking skills. In Takanashi's (2004) study, only 9% of participants reported an interest in taking English writing classes. Finally, the MEXT survey also suggested that Japanese students are not actively engaged in writing in English, with as many as 80% claiming that they had no opportunities to write their opinions or ideas in English at home. Instead, they mainly spent their time looking up the meanings of words in a dictionary (Benesse Educational Research and Development Institute, 2015).

The educational circumstances described above have prompted us to address the situation. It is necessary to overcome the difficulty in acquiring writing skills for Japanese learners of English and to find a solution that could fill the gap between the national goal of English language proficiency and the current reality. In an attempt to do so, the present study aims to illuminate the role of the student as an assessor in the development of English writing in classroom assessment. This study takes the position that student assessment activities such as self-assessment and peer assessment should be introduced into school writing curricula, because they present opportunities for learners to acquire new perspectives that enable learners to go beyond traditional assessments and relationships (Bazerman, Applebee, Berninger, Brandt, Graham, Matsuda, & Schleppegrell, 2017).

This study proposes that students be involved in classroom assessment as a means of breaking the deadlock. What is called *formative assessment* can then be carried out. Formative assessment is defined as "the ongoing instructional process during daily classroom activities for adjusting and continuing instruction" (Brookhart, 2003, p. 7) and connects assessment with improvements in writing proficiency. It could have a positive effect on the development of L2 writing skills and learner affect toward L2 writing such as writing anxiety and learner autonomy.

To make writing classes more effective and active, the focus of classroom assessment should move away from the teachers and towards the students. A number of factors have caused the present stagnant state of writing classes. For instance, one of the noticeable concerns is that writing classes are monotonous (Goto, 2005; Takahashi, 2004). Consequently, the relationship between individual students' roles, their writing ability and their affect should be analysed in terms of classroom assessment, so that writing classes might then become both reflective and participative.

The reason why writing assessment in the classroom is being considered here is that classroom assessment is closely connected with daily language learning in the classroom. In other words, classroom assessment should function in a formative way (though classroom assessment can also be summative). The next section provides an overview of the key terms, in order to clarify the present study's aims.

1.1.1 Learners as Agents of Formative Classroom Assessment

Classroom assessments generally take the form of summative and formative assessments (Brookhart, 2003, p. 7). Summative assessment is carried out at the end of a unit, term or year.

It is used to evaluate students' progress and/or study outcome (Davison & Leung, 2009). Formative assessment is conducted during the teaching process, and the data and information acquired from it are used to adapt teaching and learning to attaining study goals (Garrison & Ehringhaus, 2013). In short, classroom-based assessment is closer to students' daily study and is designed to encourage learners to develop themselves.

To increase the effectiveness of writing instruction, what happens inside the classroom should be analysed and the findings should be offered to teachers, along with information about how teachers have helped students to learn, because it is mainly the teachers who use assessment data to put information into practice (Mandinach & Gummer, 2015). In other words, the analysis of classes, that is, the assessment of how instruction is proceeding, is used for observing students' progress towards achieving pedagogical goals and for amending teaching content and strategies. Teachers should be given multifaceted data about classroom assessment to help them adjust their classes and give students valuable feedback on what they are learning. Thus, classroom assessment should be conceptualized as formative assessment. The knowledge gained from such assessment gives teachers a better understanding of how they should use the information in their subsequent teaching. Therefore, they should obtain information on multiple aspects from formative assessment procedures in order to develop students' writing proficiency effectively.

Everyday assessment is a fundamental part of teaching and learning; teachers need to acquire knowledge about their students' needs, achievements and abilities, and they can do so through the use of formative assessment. In the classroom, teachers are often expected to be the main assessing agent, while students are regarded as the recipients of instruction. The term "teacher assessment" is commonly used to describe both everyday assessment, which is performed throughout the important stage of learning, and the judgements made by teachers at the end of the key stage. However, it should not be overlooked that students might also be able to act as important agents in promoting formative assessment. Thus, to advance this process, the role of students in the classroom should be reconsidered. The transformation of the student role from a passive to an active agent of assessment would broaden perspectives and could be reflected in a more beneficial formative assessment. Harlen and Winter (2004) defined it as the type of classroom assessment activity in which students are involved in terms of learning English learning skills. In the present instance, this would revitalize writing instruction and learning and enhance formative assessment. Moreover, teacher instruction and assessment would be re-energized by the augmentation of student assessment.

The present study focuses on how student assessment may contribute to formative assessment. Additionally, it aims to elucidate the differences between self- and peer assessment so that each assessment mode can increase the effect of teacher assessment on the development of students' writing ability. The efficacy of student assessment has been discussed in applied linguistics mainly in terms of its effect on performance development and the reliability of assessment results. Yet, many studies have focused on young adults. Adolescent learners such as junior and senior high school students have not been targeted, though the implementation of student assessment in regular classes is expected to invigorate writing classes.

In addition, learner affect related to writing has not received much attention in previous studies. Thus, the relationship between writing and those affective conditions should be unpacked. The present study focuses on two types of learner affect: writing anxiety and learner autonomy. They were chosen because they have been considered to be affected by student assessment in previous studies (MacIntyre, Noels, & Clément, 1997; McDonald & Boud, 2003; Tsui & Ng, 2000).

In sum, the present study focuses on the effect of student assessment on writing performance, its reliability and learner affect in the context of formative assessment. The following sections describe these assumptions.

1.2 Why Student Assessment Influences Writing Ability

It is believed that student assessment has some effect on the development of writing ability. The reason is that the interchange of roles in language assessment affects learning, because students learn by taking on the roles of assessors and examiners of others. For instance, taking time to allow students to understand scoring rubrics helps them to improve their learning. Students could be given a simpler modified rubric used by the teachers, or they could be asked to revise them or even to tailor their own rubrics (Black, Harrison, Lee, Marshall, & Wiliam, 2004). It is important for learners to understand the criteria before being assessed. Harlen and Winter (2004) claimed that understanding the criteria is important for learning because it places the students in a position of controlling their learning. Knowing the assessment criteria makes it easier for them to understand their position in the learning process without the need for the teacher to tell them where they need to improve. Learners should understand the quality of judgement as well as the goal of the task to employ this effectively. Harlen and Winter stated that "knowing what they are aiming to do is one thing, but knowing whether they have done it well is another" (2004, p. 404). Therefore, student assessment helps them to become aware of their strengths and weaknesses.

This study focuses specifically on two types of student assessment: students' self-assessment and peer assessment. Here, self-assessment is defined as "students' reflection on, and engagement in their own work as a means to evaluate their own performance" (Boud, 1992; Boud & Falchikov, 1989, p. 395). Peer assessment refers to "an arrangement in which individuals consider the amount, level, value, worth and quality of success of the products or outcomes of the learning of peers" (Topping, 1998, p. 250). Both types of assessment have been adopted by teachers for many years, and researchers have analysed their reliability and effectiveness in student learning.

For instance, the reliability of student assessment has been investigated in order to look into whether student assessment could be a substitute for teacher assessment. Yet many of them have not examined specific aspects of reliability such as its strictness and consistency. Therefore, it is meaningful to look closely into these aspects of reliability of student assessment. As regards its effectiveness in learning, student assessment has been found to help learners increase their sense of responsibility in managing the assessment process, to make students sensitive to assessment criteria, to make them aware of individual strengths and weaknesses, and to increase their responsibility for learning (Orsmond, Merry, & Reiling, 2000; Saito & Fujita, 2004; Sung, Chang, Chiou, & Hou, 2005).

Self- and peer assessment are categorized as one type of formative assessment (Black, Harrison, Marshall, & Wiliam, 2003; Wiliam, 2000), since student assessment is closely connected with student learning and is also considered to be effective for improving English language ability. According to previous research, it can: (1) promote student learning; (2) encourage students to be more reflective; and (3) yield formal accreditation and accountability of knowledge (Boud, 1990; Harvey & Knight, 1996; Kwan & Leung, 1996). In other words, self-assessment and peer assessment give teachers and students the opportunity to reconsider how student learning produces outcomes from an assessed assignment. In particular, teachers could have meaningful communication with their students, and reflect on the process of completing the assignment rather than just the product (Orsmond et al., 2000). To make the introduction of self- and peer assessment methods more efficacious, it is helpful to concentrate on the process rather than the outcome. Hence, formative assessment should partly involve students' self- and peer assessment (Black et al., 2003; Wiliam, 2000, 2007).

The main goal of the present study is to explore the effectiveness of self- and peer assessment for formative assessment purposes. Learning and assessment are inextricably interconnected, and they make formative assessment impactful. In other words, formative assessment lies at the confluence of two processes. The first is activated by learners' perception of a gap between their desired goal and their present position (in terms of knowledge, understanding and skill). The second is accomplished by learners achieving the desired goal (Black & Wiliam, 1998a; Ramaprasad, 1983; Sadler, 1989). Black and Wiliam (1998a) argued that the promotion of self- and peer assessment in a classroom would allow learners to fulfil the gap between the learning goal and the learner's present stage. It would also show how a learner might work to close such a gap, because students need to understand the learning goals and the assessment criteria, and have the opportunity to reflect on their work. Nicol and Macfarlane-Dick (2006) stated that self-assessment tasks drive learners to generate systematized opportunities for self-monitoring and track the degree of their progress toward goals. On the other hand, peer assessment might be a helpful device for enhancing a variety of social and communication skills, such as negotiating and diplomatic skills, verbal communication skills, exchanging criticism, justifying one's situation and objectively judging suggestions (Topping, 2000).

In sum, self- and peer assessment can function as means for meta-cognitive training. Student assessment has been implemented to develop greater learner responsibility when it is conducted in conjunction with teacher assessment (Scriven, 1967). Therefore, student assessment is considered to have some effect on the development of writing ability. However, it could also be used for the design and interpretation of assessments, an issue that has not been thoroughly investigated (Stiggins, 2002). Self- and peer assessment would not succeed as an effective formative assessment tool if they just depended on students' ability to evaluate according to criteria (Orsmond, Merry, & Reiling, 1996). In other words, students are expected to possess introspection, objectivity and metacognition. Therefore, one of the purposes of the present study is to clarify the specific function of student assessment in the teaching of writing ability.

1.3 The Influence of Student Assessment on Learner Affect

As noted above, the present study focuses on two types of learner affect, namely writing anxiety and learner autonomy, because both can be positively influenced by student assessment. Furthermore, the relationship between student assessment and those types of affect has not been discussed in previous studies.

Writing anxiety is a significantly different affect from other forms of anxiety, is unique to written communication (Bline, Lowe, Meixner, Nouri, & Pearce, 2001; Burgoon & Hale, 1983a, 1983b; Daly & Wilson, 1983), and the interplay between writing anxiety and writing performance is more intricate (Daly & Shamo, 1976, 1978). Writing anxiety has been found to be adversely connected with writing processes (Bannister, 1992; Bloom, 1980), though many previous studies have focused on young adults. Moreover, research on the writing anxiety of Japanese English learners is scant. Hence, the writing anxiety of Japanese high school students in particular merits investigation.

Previous studies on the effect of anxiety have shown that student assessment is hypothesized to reduce learners' anxiety, which plays a central role in foreign language learning in the classroom (Clément, Gardner, Smythe, 1977; Gregersen & Horwitz, 2002; MacIntyre, 1992; MacIntyre et al., 1997). Heywood mentioned that students tend to feel learner anxiety when teachers assess their work (2000). According to Esfandiari and Myford (2013), this is because teacher assessment is the strictest assessment out of self-, peer and teacher assessment. Self-assessment tends to be more lenient than peer assessment, and it has the potential to have positive effects on English learning and self-confidence (Butler & Lee, 2010). Previous studies also show that it enhances students' attainment, turns their consciousness toward measured and assessed purposes, and stimulates them as motivators (Ross, Rolheiser, & Hogaboam-Gray, 1998). Peer assessment may also help learners to share knowledge and experience in a less anxiety-inducing atmosphere, which can be conducive to more effective learning (Zarei & Usefli, 2015). Finally, it encourages learners to work together and thus enhances cooperative skills (Tsui & Ng, 2000).

It has been reported that both self-assessment and peer assessment have some influence on language learners' anxiety (Cho, Schunn, & Wilson, 2006; Ross, 1998); however, writing anxiety has rarely been discussed in terms of its relationship with different forms of student assessment. Thus, the present study aims to examine how writing anxiety relates to the type of student assessment, namely self- or peer student assessment. The present study also explores learner autonomy. Because student assessment may limit the power teachers have over students (Searby & Ewers, 1997), the effect of student assessment on learner autonomy should be investigated in terms of the efficacy of introducing it in writing classes. The definition of learner autonomy in this study follows that of Benson (2001), who stated that it is the "capacity to take control of one's own learning" (p. 61). Autonomy in language learning should be recognized in the context of the three levels at which learner control may be exercised: control over learning management, control over the cognitive process and control over learning content (Benson, 2001, p. 61). Students' language proficiency is influenced by learner autonomy (Dafei, 2007) because of the close connection between autonomy and effective learning (Benson, 2001). Therefore, this study examines the influence of student assessment on learner autonomy and explores the effectiveness of the implementation of student assessment in class.

Student assessment has been associated with moves towards self-regulated learning as well as developing greater student autonomy and responsibility in learning (Lew, Alwis, & Schmidt, 2010). Student assessment gives students opportunities to internalize criteria of professional expertise and reflect on their progress, which will in turn help them to regulate the effectiveness of their learning. As well as fostering learner autonomy, self- and peer assessment can help students develop evaluation skills and take responsibility for their own learning (Cho et al., 2006). Students are expected to bear the aim of their study in mind in assessing the progress they make toward achieving study goals. They should then be able to steer their way toward becoming independent learners (Black et al., 2004).

It is believed that student assessment has an effect on language learning anxiety and learner autonomy. Thus, students would gain greater satisfaction from their learning by being more deeply involved in the assessment process (Sluijsmans, Brand-Gruwel, & Merriënboer, 2002). However, in its formative mode, self- and peer assessment is contingent upon educational and learners' individual conditions because it can be influenced by other factors, for instance the classroom environment, individual motivation, confidence and relationships with peers. Therefore, one is bound to examine the extent to which individual students are affected by learners' anxiety and autonomy when they write in English; the relationship between student assessment and those affects can then be ascertained.

In summary, it is imperative that writing classes be made more formative in the context of the national reform of English education and assessment. Cooperative and independent learning would lead to more dynamic and motivational classes. Self- and peer assessment would encourage teachers and students to change the form and substance of their writing activities. It would be an effective approach to developing writing ability as well as facilitating positive learner affect by, for example, reducing writing anxiety and enhancing learner autonomy.

1.4 Research Questions

The present study aims to examine the effects of student assessment on reliability, writing ability and learner affect. Student assessment is considered to have a positive influence on writing performance, though its reliability and learner affect (i.e., writing anxiety and learner autonomy) have not been analysed in terms of Japanese adolescents. This needs to be done so that more effective strategies can be introduced into their classrooms. The effect of self- and peer assessment should be considered, compared and contrasted separately in order to clarify the advantages and disadvantages of the two types of student assessment in relation to teacher assessment.

Furthermore, both types of student assessment contribute to the reinforcement of formative assessment. Learners will be able to frame a perception of their study goals, which they are making efforts to achieve. Also, learners will be able to share their study goals with their peers and compare the level of desirable performance with their real performance (Ellis, 2003). Student assessment can help students identify and take the next steps in their learning, because it helps students to understand the gap between their actual performance and the desired performance. In addition, students' involvement in classroom assessment would enhance teacher assessment because formative assessment might be constructed by both teacher and student.

As its research methodology, the present study adopts a mixed-method – qualitative and quantitative – approach to data collection and results interpretation in order to reinforce the advantage and disadvantage of both methods. To be specific, a mixed-method approach makes it possible to provide a more synergistic integration between the detailed statistical assessment of quantitative analysis and the deep understanding of survey responses of qualitative analysis than a separate approach. A mixed-method approach "is premised on the idea that the use of quantitative and qualitative approaches in combination provides a better understanding of research problems than either approach alone" (Creswell & Clark, 2017, p. 18). Multiple perspectives can be gleaned from the aggregated data in a mixed-method investigation. For example, the effects of self-assessment on individual students' affect can be explored by analysing multiple data sources in combination. The present study also employs a group comparison study design to test the difference in effect after the implementation of the two assessment modes in order to identify distinct qualities of each assessment mode.

The following two research questions are proposed:

(1) How do self- and peer assessment compare with each other in terms of:

a) the reliability of scores against teacher assessment;

b) the effects on writing ability; and

c) the effects on writing anxiety and learner autonomy?

(2) How can student assessment work as formative assessment in the classroom?

The first question comprises three sub-questions that are designed to explore specific differences between self- and peer assessment. They are addressed for the most part using quantitative methods. The first sub-question considers the matter of reliability. Student assessment as an alternative or complement to teacher assessment has been researched extensively (e.g., Orsmond et al., 1996; Ross, 2006; Topping, 2003; Weaver, Keer, Schellens, & Valcke, 2011). Several studies have attempted to examine the reliability of student assessment and to implement it as a subordinate assessment. For instance, Ross (2006) observed the high internal consistency of self-assessment across items, times and tasks, and also higher student–teacher agreement. Peer assessment has also been reported as presenting a high level of reliability (Weaver et al., 2011) and as being more efficacious than teacher assessment (Zarei & Mahdavi, 2014). However, previous findings on the reliability of self-assessment have varied. This is possibly because student assessment is influenced by other factors, such as learners' maturity, academic level and cultural background. As stated above, adolescent Japanese high school students have not been widely targeted as research subjects, so the present study attempts to adjust this imbalance, in the hope that it facilitates the implementation of new assessment policies in writing classes in high school.

The second sub-question looks at the difference between self- and peer assessment with regard to writing ability, because their positive effect on writing ability has also been reported in previous studies (Black et al., 2004; Sluijsmans & Prins, 2006). What has not been researched is the extent to which they influence it, and in what particular areas. This is especially so in the case of Japanese adolescents.

Many previous studies have used holistic rather than analytic scoring methods. Analytic scoring rubrics yield more information about a student's performance than single holistic scores. Analytic scoring methods present various perspectives of writing performance, such as syntactic control, vocabulary depth and organizational control (Wiseman, 2012). Holistic composite scores do not provide explicit information on how much additional instruction is needed for learners (Becker, 2011). Therefore, the present study employs analytic scoring methods.

The third sub-question focuses on the effects of student assessment on learner affect (i.e., writing anxiety and learner autonomy), which have not been researched adequately in the context of student assessment. In particular, there has been little research into the area of the relationship between anxiety and writing (E. Horwitz, M. Horwitz, & Cope, 1986; Philips, 1992; Woodrow, 2006). In addition, as already noted, most studies have targeted young adults (Cheng, 2004; Cho el al., 2006). Therefore, it would be useful to examine the extent to which student assessment influences writing anxiety among high school students (Butler & Lee, 2010).

It is believed that student assessment has positive effects on the development of learner autonomy (Cho et al., 2006). However, similarly to writing anxiety, previous studies have focused on young adults, not adolescents. Furthermore, self- and peer assessment have not been compared with each other in terms of their effects on the development of learning autonomy. As Wenden (1987) stated, learner autonomy is generally recognized as an important "pedagogical goal" (p. 8). Moreover, Cotterall (1995) opined that autonomous learners tend to try to overcome obstacles in the way, such as educational background, culture norms and former experiences. In short, students are expected to become autonomous learners. Student assessment is regarded as an effective driver of autonomous learning, so the distinct (and common) effects of different assessment need to be investigated.

The first main research question focuses on the effects of student assessment mode on writing ability, reliability and learner affect. Those effects are integrated and discussed to find out how student assessment could contribute to formative assessment. However, those effects that are found using quantitative methods would not sufficiently address individual differences in values and nuances across students. Thus, the second research question will mainly be addressed by conducting qualitative data analyses where these values and nuances can be investigated.

The second main research question is aimed at obtaining evidence from the qualitative analysis to discuss how self- and peer assessment could make classroom assessment more functionally formative. Specifically, it asks how individual students are involved in the activity of self- or peer assessment and learning instead of being passive recipients of information. Both assessment types are assumed to be influential measures in formative assessment, but it remains unclear how student assessment should be embedded in the process of formative assessment. In order to elucidate the relationship between student assessment and formative assessment, the ideas and opinions of students and teachers should be qualitatively analysed. This is because qualitative analysis is effective for "clarify[ing] inner experiences of participants and [exploring] how meanings are formed and transformed" (Corbin & Strauss, 2015, p. 5).

Qualitative data such as learner responses to open-ended questionnaire items and interviews are analysed using the grounded theory method. In order to analyse data, a set of inductive strategies are used. To put it sequentially, first, the qualitative data are synthesized; second, patterned relationships within data are identified; and finally, more progressively abstract conceptual categories are extracted (Charmaz, 1996). The purpose of formative assessment is to connect assessment with learning, so this study is designed to find how to involve student assessment in the form of formative assessment. Also, the relationship between teacher assessment and student assessment is considered, because teachers take the main roles in classroom assessment.

Finally, the quantitative and qualitative results are all integrated into a comprehensive whole, and the efficacy of student assessment is discussed based on combined results. It is also considered how and when each student assessment mode should be implemented in class to make formative assessment successful.

1.5 Organization of the Dissertation

The dissertation comprises six chapters in addition to this introduction. The literature review in Chapter 2 presents the background to student assessment, its purpose, agents and effect. It also presents hypotheses for the research questions. Chapter 3 provides an overview of the study's methodology for the mixed-method investigation. Chapter 4 summarizes the results of the quantitative analysis that was conducted to address the first main research question, while Chapter 5 presents qualitative study results in relation to the second research question. Chapter 6 synthesizes the results of the quantitative and qualitative analysis from Chapters 4 and 5 and discusses new possibilities for student formative assessment based on the study results. Finally, Chapter 7, the conclusion, summarizes the study's findings and outlines its limitations and pedagogical implications, with suggestions of directions for further research.

Chapter 2

LITERATURE REVIEW

2.1 Introduction

This chapter mainly comprises five main sections. In the first section, the background of the current study is explained in terms of the difference between formative and summative assessment as well as student learning of classroom-based assessment. The second section focuses on previous studies related to student assessment, specifically on its reliability and effectiveness, in order to clarify the advantages and disadvantages of the two types of student assessment in relation to teacher assessment. The third section discusses second/foreign language learning anxiety, with a special focus on writing anxiety, and then the fourth section looks at learner autonomy, referring to the relationship between learner autonomy and student assessment. Finally, the hypotheses corresponding to the research questions are presented based on the literature review.

2.2 Background of the Study

2.2.1 The Framework of Assessment in the Classroom: Classroom-based Assessment

The main purpose of the current study is to explore the efficacy of student assessment, so it is necessary, first, to clarify the context of classroom assessment, where students are expected to work as agents of such assessment. Classroom assessment is multifaceted, with properties connected with learning and teaching. The assessment in the classroom is carried out by key stakeholders, namely teachers and students. In classroom assessment, it is important to link teaching content with students' learning to meet learning needs and assist teachers' decision-making. In the previous literature, the role of students in classroom assessment is not necessarily clear, as teacher assessment has been taken for granted, though students are also important agents. Therefore, it is meaningful that this study focuses on student assessment, such as self- and peer assessment, and clarifies the distinct features of those assessment types as well as specific roles that students can play in classroom assessment. This is because the goal of classroom assessment is to adjust teaching and learning in order to help students to achieve a learning goal. It is also significant that classroom assessment should function formatively in order to link learning to assessment.

In a classroom assessment, there is a psychological context between teachers and students, because classroom assessment and teachers' instruction are closely related to each other. Therefore, classroom assessment should be formative (Benson, 2000; Brookhart, 2003). These principles can operate in both student- and teacher-centred classrooms. Even in a teacher-centred classroom, the students' point of view matters because of its effect on learning. From a student's point of view, classroom assessment information is not merely information about the students themselves. Rather, it constitutes a major part of students' learning life, learning content and their relationship with peers and teachers (Black & Wiliam, 1998a; Brookhart, 2003; Crooks, 1988).

Classroom-based assessment has been described in conjunction with the term "assessment bridge", which means "the place where assessment, teaching and learning interweave in the classroom" (Colby-Kelly & Turner, 2007, p. 12), so classroom assessment should be organized focusing on these influences on assessment and planned in terms of how it should be conducted. In other words, good classroom practice is not conducted without assessment strategies related to informal assessment opportunities, and they are regularly embedded in teachers' instruction (Rea-Dickins, 2000). In good classroom practice, teachers ask students questions, interaction between learners and their teacher is encouraged, and learners even collaborate effectively with one another (Rea-Dickins, 2000). These classroom activities all help to develop the learners' awareness of, for instance, the language being used and learning goals in class, and to stimulate reflection on what is being taught. This information may be a significant source to guide students to develop themselves, and to inform teachers about learners' progress and achievement (Rea-Dickins, 2000). In short, classroom-based assessment encourages teachers to be aware of new aspects of students' learning.

Compared to large-scale testing, it might be felt that classroom-based assessment does not need to be "as reliable" (Brookhart, 2003). One important thing is that learners' errors can be instantaneously pointed out and corrected, and more information about learning content can be provided by teachers in class. Classroom-based assessment mainly stresses the educational role it plays in the progress of individual students. In addition, it is said that classroom-based assessment might be on a smaller scale than large-scale testing. DePascale (2003) stated that the purpose of large-scale tests is to provide schools, teachers and students with accountability. For instance, large-scale testing gives information on the success of schools in order for teachers to be accountable for student achievement (Landau, Vohs, & Romano, 1999). Large-scale testing also allows students to ensure that they have acquired the knowledge and skills they require before promotion and graduation (DePascale, 2003). In brief, the purposes of classroom assessment and large-scale testing are different, so applying the goal of large-scale testing to classroom assessment is questionable.

Messick (1989) stated that the main difference between classroom-based assessment and large-scale testing lies in the presence or absence of a classroom context. Classroom-based assessment has been discussed in terms of the effect of classroom-based assessment on students. For example, Crooks (1988) categorized the impact of classroom assessment on students into three major perspectives: (1) the impact of normal classroom assessment on students; (2) the impact of the relationship with teaching content in class on assessment; and (3) the impact of classroom assessment on the motivational aspects of students. This is because classroom-based assessment tends to have a narrower focus than large-scale testing, and such a limited focus might have an effect on students.

On the other hand, Harlen and Winter (2004) elaborated on Crooks's (1998) first point, which is the improvement of student learning in classroom-based assessment. For example, they proposed five key points aimed at maximizing the success of classroom-based assessment in terms of promoting students' attainment of the study goal: (1) information about the learning process and products should be collected with a view to improving teaching and students' learning; (2) feedback should be given to students to enable them to improve their skills and go forward with their learning; (3) teachers and students should share their understanding of learning goals in every study; (4) students should be involved in classroom-based assessment such as self- and peer assessment; and (5) students should not be passive recipients but rather active learners when they learn (Black & Wiliam, 1998a; Harlen & Winter, 2004, p. 392). Hence, previous research has indicated that classroom-based assessment has different perspectives, but the current study aims to find effective ways to implement student assessment in class. In other words, it is believed that student assessment could be useful for accelerating formative assessment rather than summative assessment, so the current study adopts the perspectives that were proposed by Harlen and Winter (2004).

2.2.2 Formative Assessment

One of the main aims of the current study is to investigate the effectiveness of student assessment in terms of formative assessment. Therefore, this section discusses the meaning of formative assessment in contrast to summative assessment with a specific focus on the following four points: (1) definitions and purposes of formative and summative assessment; (2) the impact of formative assessment on teaching and learning; (3) the process of formative assessment; and (4) the difference between formative and summative assessment.

First, definitions and purposes of formative and summative assessment are elucidated. The term “formative evaluation” was first coined by Scriven (1967). Since then, formative assessment has meant that learners learn through ongoing assessment procedures aimed at supporting learning. In other words, it enhances learning rather than just being used to calculate final grades (as in summative assessment). Specifically, formative assessment is defined in terms of action taken as a consequence of an assessment activity. Rea-Dickins (2000) stated that formative assessment might be performed by the key stakeholders in the assessment process, such as teachers, learners and the school, because formative assessment is based on perspectives of learner performance in assessment. To be specific, as Black and Wiliam (2003) mentioned, formative assessment assists teachers in choosing what they should teach in the next stage as well as helping teachers decide what to do next. However, at the same time, learners are also expected to understand what they have learned and what they need to learn next. The learner’s role is important because it is the learner who does the learning (Black & Wiliam, 2003). The term “formative” itself is open to a variety of interpretations. It often means that assessment is planned at the same time as teaching. Such assessment does not necessarily have all the characteristics identified above as helping learning. For instance, if teachers can identify points on which students need more explanation or practice supported by assessment, it might be called “formative assessment”. However, formative assessment should be interpreted as not only indicating learners’ success or failure but also informing students how to move ahead toward the learning goals. In short, information that is produced from formative assessment should be effectively used to contribute to making changes to students’ learning (Bloom, 1969; Scriven, 1967).

On the other hand, summative assessment is generally defined as formal planned assessments at the end of a unit, term or year, which are used to evaluate student progress and/or grade students (Davison & Leung, 2009, p. 397). Large-scale tests are more frequently categorized as summative assessment than classroom-based assessment. Summative assessment is usually administered at the end of episodes of teaching (units, course, etc.) in order to grade or certify students, or assess the usefulness of a curriculum (Bloom, Hastings, & Madaus, 1971). In other words, some classroom assessments are intrinsically summative, serving the purposes of grading and accountability (Saito & Inoi, 2017). If the tests are not used to give students feedback about learning, if they are no more than indicators of a final high-stakes summative test or if they are just components of a continuous assessment scheme in order to facilitate a high-stakes interpretation, classroom-based assessment will amount to no more than frequent summative assessment (Black & Wiliam, 1998a).

Secondly, the impact of formative assessment on teaching and learning is summarized here. According to Black and Wiliam (2009), a formative interaction could lead learners toward mutual interaction between external stimulation and internal production. External stimulation refers to, for instance, teacher instruction, feedback and assessment. Examples of internal production include students’ learning attainment and learners’ affect. Such an interactive relationship between an external stimulus and internal feedback has an effect on learners’ cognition. In a formative process, formative interaction depends on the formative interaction, because students' attainment of educational goals and teachers' assessment constantly influence classroom-based assessment. In other words, teachers tend to change their teaching contents and strategies in response to those students' daily achievement. In short, teachers' initial prompts are designed to encourage students to think more. By so doing, the students are more actively involved in formative learning (Black & Wiliam, 2009; Davis, 1997).

Thus, it is possible that formative assessment could have an innovative impact on teaching and learning, because formative assessment intrinsically functions to interact between teaching and learning in terms of innovation. However, it is difficult to find any innovation in formative assessment when the latter works as a marginal bridge between teaching and learning. Therefore, it is critical to focus on the nature of these interactions between teachers and students, and between students and their peers, and the outcomes of any change between them. However, it might be difficult to acquire data about formative assessment from many of the published reports because the classroom is treated as a black box, as Black and Wiliam mentioned (1998b, pp. 1–2). Learning is promoted by how teachers and students interact with one another, but teachers are asked to deal with various kinds of complicated and challenging matters, such as the personal, social and emotional pressure around teachers and students. Teachers are also expected to have initiatives to determine standards and accountability in class, though they might not be sure how new input such as teaching resources and tests will produce better outputs (Black & Wiliam, 1998b). As Black and Wiliam stated, teachers are expected to take responsibility for making the classroom work better and for achieving the required standards in class: however, the interaction between learning and teaching has been ignored. Therefore, it is necessary to elucidate the inside of a black box, that is, the interaction between learning and teaching. Doing so would provide teachers with the direction they need to improve teaching and learning.

As the third point, it is considered that the formative assessment process works like a cycle in learning. Teachers are also key components of formative assessment. Viewed from a student standpoint, the students' formative task is to compare ideal and actual performance, and attempt to close the gap. Three stages of students' learning improvement are presented by Brookhart (2003) in terms of formative learning. First, students hold an intention toward their learning goal. The learning goal is regularly prepared for by teachers, however internalization of the learning goal functions meaningfully in the learning process. Such a growing concept, namely the clarification of the learning goal, shapes part of the learning itself. In the second stage, students are expected to monitor their ability and study by comparing their reality with the ideal. Finally, students could enhance their ability and strategies to close the gap. Students are also asked to build their own goals as independent learners. During this learning process, teachers are also responsible for providing the feedback that helps students to approach their goals. Such provision of feedback by teachers indicates formative assessment (Brookhart, 2003).

As previously mentioned, both teachers and students are key components in formative assessment. Above all, the student occupies a central role in the formative process because it is students who can take the actions required to improve. Furthermore, the formative assessment process partially constructs learning and students come to understand the standard of high-quality work. It is also important to focus on the role of teachers and students in terms of assessment in the classroom in order to promote formative assessment. In the classroom, teachers generally take on the role of assessor, while students are the assessed. Black and Wiliam (2009) stated that the interaction between the assessor and the assessed may be particularly effective, partly because the quality of interactive feedback is a critical point in determining the quality of the learning activity, and is therefore a central feature of pedagogy (2009, p. 100).

Finally, it is stated how formative assessment functions differently to summative assessment in terms of learning and assessment. In contrast to summative assessment, formative assessment changes the relationship between the measures and the measured, because the assessor is the only one who necessarily has to understand the standard in summative assessment (Wiliam, 1998). Thus, formative assessment has different perspectives from summative assessment (Leung & Mohan, 2004). In short, the difference between them is evident in terms of learning and evaluation. In summative assessment, evaluation – or grading – tends to be influential in the learning process. Moreover, summative assessment generally means an examination that is given at the end of periods such as units or terms, so that students are graded or certified (Bloom et al., 1971). On the other hand, as mentioned above, formative assessment puts more emphasis on the potential of classroom assessment to assist learning. In particular, feedback to students is considered to be regular, relevant and specific to the task.

Scriven (1967) used the terms "formative" and "summative" assessment to distinguish between the two roles that evaluation may play in education (pp. 40−43). Since there is a difference between formative and summative assessment, some research has discussed the need to add a formative process to summative assessment. For instance, Shinn and Good III (1993) stated that a paradigm shift is needed in assessment, that is, from summative functions of assessment to the formative functions of assessment. Shinn and Good III also commented that formative assessment works in summative assessment like problem solving. Specifically, information from language tests can be useful for the purpose of formative assessment, to help students reflect their own subsequent learning, or for helping teachers change their teaching methods and materials so as to make them more appropriate for their students' needs, interests and abilities (Genesee & Upshur, 1996, p. 49). As well as performing a formative function, a summative function is also employed in order to grade for purposes of certification or placement of students on to the next level. In short, it is possible that summative and formative assessment mutually compensate for each other's advantages and limitations.

Hence, as Black and Wiliam (2003) stated, summative assessments can even be used formatively to provide students and teachers with constructive feedback and improve learning and teaching, although formative and summative assessments are considered to be different in terms of form and function. This is because the quality of the information in summative and formative assessment is basically the same in a broad sense. In other words, everyday formative assessment provides small amounts of information that can be used as part of the data for the overall summative assessment (Teasdale & Leung, 2000). Brookhart (2003) also supported the combination of summative and formative assessment but in a different way. For instance, if teachers used evaluation in regard to students' development and improvement in class and made decisions about the result of an educational process, it would be called "summative".

It is worth noting that both summative and formative assessment are aimed at providing students and teachers with supportive information related to different aspects, so it is important to clarify the goals and requirements of classroom assessment for students and teachers. Also, it is vital that students and teachers are involved in the learning process and share improvement in either summative or formative way. However, many previous studies have explored the efficacy of formative assessment using a qualitative method such as class observation and interviews with teachers rather than quantitative analysis. Moreover, few studies have focused on the relationship between specific assessment types, such as student assessment and formative assessment. Therefore, it is worthwhile for the current study to examine the efficacy of student assessment using a mixed-method approach.

2.2.3 Teacher Assessment

The main agent or assessor of classroom-based assessment is the teacher, and teachers are expected to evaluate students both summatively and formatively. Therefore, this section discusses the role of teacher assessment in classroom-based assessment. The term "teacher assessment" is commonly used to describe both everyday assessment and the judgements made by teachers at the end of an important stage of instruction. Everyday assessment is an integral part of teaching and learning in terms of formative assessment, because teachers can gain knowledge about their students' needs, attainments and abilities from it. Statutory teacher assessment involves teachers using the knowledge gained from everyday assessment to make and record their judgements on students' overall attainment at the end of a key stage (School Curriculum & Assessment Authority [SCAA], 1995, p. 4; Teasdale & Leung, 2000).

According to Teasdale and Leung (2000), there are three notable characteristics of a teacher's assessment: (1) teachers do not understand formative assessment very well, so teachers practise it weakly; (2) formative assessment is conducted under strong influence of the context of national or local requirements for certification and accountability; (3) teachers are expected to innovate their perceptions of their own role in relation to their students and in their classroom practice.

In classroom-based assessment, a teacher plays a more predominant role as an "agent" in the "washback", compared to the roles in assessment used for student placement or the measurement of achievement (Harlen, 1996). "Washback" means the impact of a test on what teachers and students do in the classroom (Buck, 1988, p. 17). Berry (1994) notes the effect of a test on classroom practice (p. 31). This is because the washback has the most powerful effects on teaching and learning behaviours where participants see the tests as challenging and the results as important. Therefore, washback has a micro-level impact on the contexts of the classroom assessment. In that sense, teacher assessment is defined more precisely as relating "to the agent of the assessment, while the formative/summative distinction refers to the purpose of the assessment" (Harlen, 1996, p. 129). Classroom-based assessment depends on the consistency with which teachers subjectively understand what it is they are actually assessing and how they make decisions.

It has been believed that classroom-based teacher assessment does not represent a high-stakes context unlike formal tests and examinations. However, this view of classroom assessment – as having low stakes – is not supported by research findings (Rea-Dickins & Gardner, 2000). Rea-Dickins and Gardner (2000) investigated the nature of formative assessment in a language learning context. They found that the classroom context represents a major high-stakes setting, and that wrong decisions by teachers have very serious implications for individual students (p. 238). The gap between teachers and students regarding the interpretation of low- or high-stakes assessment might be caused by the misunderstanding of formative classroom assessment and summative classroom assessment (Schneider, Egan, & Julian, 2013). In spite of the wide professional support for teacher assessment (Daugherty, 1996), classroom communication between teachers and students has not been interpreted enough for assessment to be made (Teasdale & Leung, 2000). Teachers clearly face difficult problems in controlling the balance between their formative and summative roles, and they are confused over how to manage to bridge the gap between formative and summative roles in class. This confusion in teachers' minds could impede the integration of teaching and assessment (Davison, 2007; Oi, 2018).

With a view to developing classroom-based assessment in terms of formative assessment, not only teacher assessment but also student assessment should be positively accepted, because classroom assessment comprises both agents (teachers and students) in order to make classroom interaction more productive. Student assessment supports teachers in making them aware of new perspectives of classroom-based assessment. Therefore, the following section discusses the feasibility of student assessment in the classroom.

2.2.4 The Implication of Student Assessment

As previously mentioned, teacher assessment is an integral part of classroom assessment, but the current study aims to explore the feasibility of student assessment in terms of formative assessment. So the categories of language assessment are presented first. Early work on formative assessment focused on five main types of activity suggested by the evidence of their latent effectiveness, and developed with and by teachers in regular classroom work. The five main types are as follows: (1) learners share assessment criteria; (2) learners are questioned in the classroom; (3) learners receive assessment; (4) learners are involved in assessment; and (5) summative tests are used as formative functions (Black et al., 2003; William et al., 2000, 2007). The effectiveness of formative work depends not only on the content of the feedback and associated learning opportunities, but also on the motivations and self-perceptions of students within which it occurs. Accordingly, student assessment such as self-assessment and peer assessment is expected to promote formative assessment.

In other words, self- and peer assessment are considered to be effective in placing the students at the centre of the process of formative assessment and providing a rationale for classroom assessment. It is also thought that teachers will know the students' "existing frameworks" for the criteria. Furthermore, students can adapt learning experiences to be within the capacity of students as learners and connect with these frameworks by knowing the results of their self- and peer assessment (Harlen & Winter, 2004, p. 392).

Black and Wiliam (1998a) stated that classroom assessment practices give students opportunities to restructure separate pieces of learning content that they have almost forgotten. Classroom assessment practices for students are associated with formative assessment, since formative assessment practices lead to a greater understanding of what is being taught, rather than learning focused on evaluation. If teachers and students are motivated to assess the ability of students, self- and peer assessment will be implemented in class (Black et al., 2004). This is because students should be encouraged to bear in mind what they study and what they assess to see learning progress. Then students can manage their studying by themselves and become independent learners (Black et al., 2004). Self-assessment and peer assessment make distinct contributions to the development of students' learning. Hence, they confirm their learning goals, which cannot be achieved in any other way.

In short, students should be encouraged to apply criteria to understand how their performance might be improved through self- and peer assessment (Black et al., 2004). Engaging in self- and peer assessment not only means checking for errors or weaknesses but also allowing students to learn actively. Involving themselves in self- and peer assessment will enable students to make explicit what is usually ambiguous (Black et al., 2004), because understanding the assessment criteria can activate learners in assessing their own work. It is also important for students to control their learning by making sure where they are in their learning process without the teacher's guidance. To do this effectively, students and teachers need to share their learning goals, and teachers should identify and tell students what the next step is in order to improve. Information concerning students' learning and progress towards goals should be identified and gathered (Harlen & Winter, 2004). Hence, the interchanging of roles between teachers and students would support teachers' instruction. Students' involvement in self- and peer assessment could help teachers to collect information to enable them to make formative assessment effective.

In order to deepen the mutual understanding between teachers and students to make self- and peer assessment effective, several approaches have been suggested. For instance, taking time to assist students in understanding scoring rubrics is very helpful. Students can be given simplified versions of the rubrics teachers use, or they can modify them or even create their own original rubrics (Black et al., 2004). Furthermore, teachers are expected to reconsider the relationship that teachers and students have developed in order to implement formative assessment efficiently (Brousseau, 1984; Perrenoud, 1998). As one of the strategies, dialogue or discussion is useful for teachers to respond to and redirect students' ideas (Black & Wiliam, 1998a). Teaching and learning must have an interactive relationship, so teachers need to know about their students' progress and difficulties with learning. This information would help teachers to meet students' needs, which are often beyond prediction and varied depending on each student, more quickly (Davison, 2007).

In sum, if self- and peer assessment are used in conjunction with teacher assessment, teachers will make their assessment process more worthwhile. It opens up the opportunity for teachers to support students by sharing information with them and by sharing taking the lead.

It allows the teacher to give immediate and constructive feedback to students. It stimulates continuous evaluation and adjustment of the teaching and learning program. It complements other forms of assessment, including external examination.

The goal of classroom-based assessment is to improve student learning, so self- and peer assessment should be essential components of all assessment activity (Davison & Leung, 2009). Students are expected to play an active role in the assessment process, particularly in the process of self- and peer assessment.

2.3 Previous Studies on Student Assessment

One of the main aims of the current study is to investigate the efficacy of student assessment, so this section discusses why student assessment plays an important role in classroom assessment. Assessment is basically carried out to give learners information about their capabilities and weaknesses in the learning process (Zarei & Usefli, 2015). Teacher assessment is considered to be the traditional system in which teachers are responsible for students' performance and learning outcome (Brown & Hudson, 1998). On the other hand, student assessment is capable of activating the classroom, because student assessment provides teachers with new perspectives and makes students one of the leading players in the class. This section presents the results of a meta-analysis of 101 previous studies on student assessment to elucidate key findings on issues corresponding to the research questions for this study. .

2.3.1 Meta-analysis Results

A total of 101 previous studies published from the 1980s to 2018, which comprise 38 studies on self-assessment, 42 studies on peer assessment and 21 studies on the comparison between self-assessment and peer assessment, were meta-analytically compared by the present author (Table 2.1). Individual assessment types of previous studies were coded in terms of study year, sample size, sample age, method and purpose. The number of comparative studies on self-assessment and peer assessment (21 out of 101, or 20.8% of the total) is fewer than that of studies focusing only on self- or peer assessment (38% and 42% of the total, respectively). Therefore, it is meaningful to conduct a comparative study to clarify the individual strengths and weaknesses of each assessment mode.

Table 2.1

A Comparison of Student Assessment Studies from the 1980s to 2018

Year	Self-assessment	Peer assessment	Self-assessment & peer assessment	Total
1980−1989	5/38 (13.1 %)	1/42 (2.3 %)	1/21 (4.7 %)	7
1990−1999	8/38 (21.0 %)	13/42 (30.9 %)	1/21 (4.7 %)	22
2000−2009	16/38 (42.1 %)	10/42 (23.8 %)	10/21 (47.6 %)	36
2010−2018	9/38 (23.6 %)	18/42 (42.8 %)	9/21 (42.8 %)	36
Total	38	42	21	101

As can be seen in Table 2.2, most of the examinees in the studies were found to be university students. In contrast, few studies involved middle and high school students (7 out of 101 or 6.9% of the total). Above all, the number of studies focusing on high school students was the lowest (3 out of 101 or 2.9% of the total). This finding supports the current study's focus on investigating the effect of student assessment on high school students.

Table 2.2

The Ages of Examinees in Studies on Student Assessment

	Young learners	Middle school	High school	University (undergraduate)	Adult learners	Others
Self-assessment	5/38 (13.1 %)	2/38 (5.2 %)	2/38 (5.2 %)	23/38 (60.5 %)	5/38 (13.1 %)	1/38 (2.6 %)
Peer assessment	1/42 (2.3 %)	1/42 (2.3 %)	1/42 (2.3 %)	37/42 (88.0 %)	1/42 (2.3 %)	1/42 (2.3 %)
Self- & peer assessment	3/21 (14.2 %)	1/21 (4.7 %)	0/21 (0 %)	15/21 (71.4 %)	0/21 (0 %)	2/21 (9.5 %)
Total	9/101 (8.9 %)	4/101 (3.9 %)	3/101 (2.9 %)	75/101 (74.2 %)	6/101 (5.9 %)	4/101 (3.9 %)

2.3.1.1 Methodology of the Study. In terms of methodology, the previous studies mainly employed quantitative methods; in particular, those studies that compared self- and peer assessment have overwhelmingly used a quantitative method (51 out of 101 or 50.4 % of the total) because researchers have been principally interested in the correlation or agreement between teacher assessment and student assessment results (Table 2.3). In contrast, the studies on self-assessment and studies on peer assessment have also used qualitative methods (42.1 % and 40.4 % of the total, respectively). This is because these studies tended to analyse students' comments on student assessment. Meanwhile, mixed-method approaches were adopted only by 12 out of the 101 studies (11.8 % of the total). Therefore, the current study decided to take a mixed-method approach to identify distinct aspects of each assessment mode.

Table 2.3

The Ratio of Research Methods

	Qualitative Research	Quantitative Research	Mixed-Method Research
Self-assessment	16/38 (42.1 %)	16/38 (42.1 %)	6/38 (15.7 %)
Peer assessment	17/42 (40.4 %)	21/42 (50.0 %)	4/42 (9.5 %)
Self- & peer assessment	5/21 (25.0 %)	14/21 (70.0 %)	2/21 (10.0 %)
Total	38/101 (37.6 %)	51/101 (50.4 %)	12/101 (11.8 %)

2.3.1.2 Sample Sizes in Previous Studies. The number of participants who took part in individual research varied greatly, with the sample size ranging from five to 637 persons for self-assessment studies, and from 2 to 1740 persons for peer assessment studies. Meanwhile, the number of participants in the comparative studies between self- and peer assessment ranged from 15 to 3588. Therefore, among the three types of studies, self-assessment studies tended to involve fewer subjects than the other two types of studies.

In regard to the average number of participants in individual studies, sample sizes varied across them, depending on the study conditions. However, between approximately 100 and 300 students most frequently participated in all types of studies. The average sample size of peer assessment (mean = 76.25; SD = 2.88) is smaller than that of the other two types of studies (mean = 221.8; SD = 1.28 for self-assessment; mean = 272.22; SD = 2.01 for comparative studies), because studies on students' revision behaviours in peer assessment tended to limit the number of participants in the survey (Cho & MacArthur, 2010; Orsmond et al., 2000). In this case, the studies investigated a small number of participants. In view of the conditions of the study, the current study investigated 293 students, which was significantly more than the average number of subjects in previous self-, peer and comparative studies.

2.3.1.3 Study Purpose. In this section, the purposes of previous studies related to the research questions of the current study are reported. Table 2.4 shows the main categories of study purposes, i.e., the number of previous studies that respectively appeared in self-assessment, peer assessment and comparative studies between self- and peer assessment. As Table 2.4 shows, the most common study aims were to examine the reliability and effectiveness of student assessment related to the improvement of writing. Above all, the category of reliability was the most frequent study aim (35 out of 87, or 40.2 % for self-assessment; 20 out of 76, or 38.1 % for peer assessment; 22 out of 41, or 53.6 % for comparative study of self- and peer assessment), and it was basically examined based on the scores of teacher assessment. Also, the effects of student assessment on writing ability and learner affect have been investigated. The study purposes are wide-ranging, so this section summarizes five key findings of these studies: (1) reliability; (2) the relationship with writing ability; (3) the effectiveness of student assessment; (4) learner affect; and (5) comparative studies between self- and peer assessment.

Table 2.4

Study Purpose

Categories	What is assessed	Self-assessment	Peer assessment	Self-assessment & peer assessment
Reliability	Reliability	11	10	4
	Agreement with teacher assessment	10	7	9
	Correlation with teacher assessment	5	3	1
	Rater severity		1	5
	The impact of rater training	2	5	
	Validity	7	3	3
Writing ability	Accuracy of language use	10		1
	The development of English proficiency	1		5
	Perception of grammar awareness		1	
Effectiveness of student assessment	Effectiveness	13	14	6
	Utility	5		
	Benefits	6		
	The impact of revision/process/criteria		9	
Learner affect	Interaction with peers		5	
	Reaction with peers		4	
	Psychological and personality traits of the rater			2
	Attitudes towards self-assessment	4		
	Belief about peer assessment		1	
	Language anxiety	10	1	
	Motivation	2		
	Learner autonomy	1		1
	Self-efficacy		2	
	The role of interpersonal variables (psychological safety, diversity, interdependence)		9	
Implementation	How to implement			4
Gender	Bias of gender		1	
	Total	87	76	41

2.3.1.3.1 Reliability of Self-assessment. According to Table 2.4, 40.2 % of the previous studies on self-assessment focused on its reliability based on the agreement or correlation between self-assessment and teacher assessment. For example, as one of the previous studies reviewed in Table 2.4, Andrade, Du, and Mycek (2010), reported, there is an adequate correlation between them: in particular, the measures of grammatical competence seem to be better indicators than the measures of pragmatic and socio-linguistic competence. However, the results show that factors such as rater training, educational experience, students having a high proficiency in English and students' involvement in establishing criteria are needed to achieve high reliability in self-assessment. In addition, it is believed that the reliability of self-assessment depends on the number of criteria rating scales. For instance, Runnels (2014) included in the present meta-analysis, investigated the reliability of self-assessment focusing on Japanese university students. The results showed that the reliability of self-assessment assigned by the university students tended to be influenced by the number of rating scales (Runnels, 2014). To be specific, a neutral option, namely the scale midpoint, appeared to have a negative effect on the reliability of self-assessment.

Dieten (1989) and Peirce, Swain, and Hart (1993), reviewed in the current meta-analysis, stated that students' higher English proficiency is a key factor in evaluating themselves accurately (Dieten, 1989; Peirce, Swain, & Hart, 1993), while the reliability of self-assessment depends on students' conditions and educational backgrounds. For instance, Boud and Falchikov (1989) subjected 48 quantitative self-assessment studies comparing self- and teacher ratings in higher education courses to a meta-analysis. They identified three factors that influence the reliability of self-assessment and teacher assessment ratings: (1) students and teachers marked more accurately in well-designed studies; (2) students in upper-level classes rated more accurately than students in elementary-level classes; and (3) students in science classes evaluated more accurately than students in other subjects. In this study, effect sizes varied from -0.62 to 1.42, with the mean effect size for this subset being 0.47. The relationship between teacher and student ratings varied from -0.05 to 0.82, with a mean *r* value of 0.39.

Ross (1998) also conducted a meta-analysis on learners' self-assessment of their L2 abilities and analysed 60 correlations obtained from previous studies. The results of contrastive multiple regression analyses showed different validities for self-assessment compared to teacher assessment. Ross (1998) mentioned that the validity of self-assessment is influenced by learners' experience of self-assessment. Ross (1998) also found that the accuracy of self-assessment tended to rise, (1) if criteria had concrete examples that included "can-do" skills; (2) if learners had episodic memory or experience of using skills in class. In addition, it was found that (3) teachers' observations of student performance were gathered and that their cumulative information shaped teachers' generalizable assessment; and (4) self-assessment may shape learners' confidence in attaining their goal. According to Ross's study (2006), which is included in the meta-analysis, the reliability of self-assessment was enhanced by being helped by teachers in terms of tasks and items, and also tended to be positive when students received self-assessment training (Ross, 2006). Additionally, Heine and Hamamura (2007), reviewed in the present meta-analysis, also found that some students tended to underestimate their proficiency and be influenced by culture or gender. Based on the above, Ross (2006) proposed conditions to increase reliability: (1) administering self-assessment over short time periods; and (2) ensuring that students are taught how to assess their work and have sufficient knowledge of the content of the domain. Therefore, careful consideration is needed to achieve high reliability of self-assessment.

2.3.1.3.2 Reliability of Peer Assessment. The results of meta-analysis of the present study (Table 2.4) showed a positive correlation or agreement between peer assessment and teacher assessment. For instance, as one of the examples of the meta-analysis, Weaver et al. (2011) reported that peer assessment indicated a high level of reliability. However, according to Orsmond et al. (1996) and Orsmond et al. (2000) reviewed in Table 2.4, this depends on how students understand the assessment criteria. To be specific, Orsmond et al. (2000) conducted a study about the effect of a peer assessment exercise among university students on the agreement between peer assessment and teacher assessment. The results showed that the maximum percentage of students marking the same as teachers was 51 %, but the others had difficulty in understanding the particular criterion “clear and concise” or expressing peer assessment referring to the assessment criteria (Orsmond et al., 2000, p. 34).

A meta-analysis that was carried out by Topping (1998), which qualitatively analysed 31 peer assessment studies with a focus on the mechanisms and benefits of peer assessment, compared the agreement between teacher and peer markings. Topping (1998) concluded that peer assessment was adequately reliable and valid in a wide range of applications. In addition, peer assessment as reviewed by Topping took three different forms: (1) peer nomination, which is the act of identifying the best and the worst performances in the group; (2) peer scoring, which is evaluating each learner based on a set of performances or assessment criteria; and (3) peer rankings, which is ranking individual learners from the best to the worst against a set of criteria (Pope, 2001; Weaver & Esposto, 2012). Topping (1998) mentioned that all types of peer assessment had shown positive formative effects on student attainment and affect.

As another example, in a meta-analysis conducted by Falchikov and Goldfinch (2000), a total of 48 quantitative peer assessment studies comparing peer and teacher marks were analysed. Falchikov and Goldfinch found a favourable resemblance between peer assessment and teacher assessment in terms of global judgements based on well-understood criteria. However, it was difficult to find a similarity in the assessment of several individual dimensions between teacher assessment and peer assessment. Effect sizes varied from -4.48 to 7.34, with the mean effect size for this subset being 0.24. The relationship between teacher and student assessment varied from -0.14 to 0.99 and was affected by different sample size, with the mean value of *r* being 0.69. Based on the results of a meta-analysis, Falchikov and Goldfinch (2000) also suggested seven reasons for applying peer assessment: (1) a lower number of peers is recommended; (2) peer assessment is desired in academic products and processes; (3) the usage of an overall global assessment and well-understood criteria is suggested; (4) the involvement of students in discussions about criteria is important; (5) the design, implementation and reporting of the study should be given attention; (6) any training genre and level make peer assessment successful; (7) positive levels of agreement between peers and teachers do not necessarily present a measure of validity, because the nature of assessment tasks and students' understanding of criteria have an impact on peer assessment.

In sum, the results of the meta-analysis showed that both types of student assessment had a positive correlation or agreement with teacher assessment. However, the reliability of self-assessment was considered to depend on rater training and learners' educational backgrounds, such as proficiency and experience of self-assessment. On the other hand, the reliability of peer assessment is presumed to be enhanced by the extent to which students can understand assessment criteria. Hence, the results of the meta-analysis showed the difference between self- and peer assessment, but a common suggestion, that is, rater training or comprehension of assessment criteria, was provided to enhance the reliability of both assessment types. However, the number of previous studies reviewed in the present meta-analysis where rater training was conducted for student assessment appeared to be lower, as Table 2.5 presents (Cheng & Warren, 2005; Saito & Fujita, 2004). Therefore, the results of the meta-analysis suggest the need for rater training before the experiment.

Table 2.5

The Number of Studies Where Rater Training was Conducted for Student Assessment

Self-assessment	Peer assessment	Self-assessment & peer assessment	Sum
2/38	5/42	5/21	12/101

The present study also analyses reliability in terms of rater severity; however, the number of previous studies focusing on rater severity is relatively lower, as Table 2.4 shows: none of the studies for self-assessment; one out of 76 or 1.3 % for peer assessment; five out of 41 studies or 12.1 % for comparative studies between self- and peer assessment (Matsuno, 2009; McDonald & Boud, 2003; Ravand & Ravand, 2016). Therefore, rater severity needs to be analysed in the present study. Thus, the results of meta-analysis evoke the need to reveal the hidden parts of each student assessment type – for instance, the relationship with individual student conditions such as English proficiency and backgrounds, even if both assessment types show satisfactory reliability compared to teacher assessment.

2.3.1.3.3 Writing Ability and Student Assessment. The present study also aims to analyse whether student assessment influences writing ability. The results of the meta-analysis in the present study also focused on the effect of student assessment on the development of writing ability: 11 out of 87 studies or 12.6 % for self-assessment; one out of 76 studies or 1.3 % for peer assessment; six out of 41 studies or 14.6 % for the comparative studies between self- and peer assessment (Table 2.4). The present meta-analysis shows that the number of studies of peer assessment focusing on writing ability was smaller than that of self-assessment (Topping, Smith, Swanson, & Elliot, 2000). The studies about self-assessment noticeably tended to investigate the accuracy of language use such as grammar and vocabulary, as demonstrated by the studies included in the current meta-analysis (Crismore, Markkanen, & Steffensen, 1993; McDonald & Boud, 2003). With regard to comparative studies, Lee (2017) and Sadler and Good (2006), reviewed in Table 2.4, examined the difference between two types of student assessment in terms of the development of English proficiency.

Table 2.6 presents the number of descriptions of the effect on writing ability that appeared in the results of the previous studies included in this meta-analysis. As can be seen in Table 2.6, about 30 % of research on student assessment has examined the relationship between the development of writing ability and student assessment, using statistical tests such as correlation coefficient, ANOVA and MFRM. As qualitative studies, interviews and questionnaires, including open-ended questions, were also commonly used. Such qualitative methodology was employed to identify factors other than writing ability development, since it was difficult to observe the change in writing ability only from the results of quantitative studies. Furthermore, in the case of exploratory studies, the effect of student assessment on writing performance tended to be specified among the effects of student assessment and not specified as an independent effect.

Table 2.6

Results of Previous Studies on the Effects of Self- and Peer Assessment on Writing Ability Improvement

	Self-assessment	Peer assessment	Self-assessment & Peer assessment	Sum
Writing ability developed	8/38 (21.0 %)	12/42 (28.5 %)	4/21 (19.0 %)	24/101 (23.7 %)
No change in writing ability	4/38 (10.5 %)	1/42 (2.3 %)	1/21 (4.7 %)	6/101 (5.9 %)
Sum	12/38 (32.4 %)	13/42 (30.9 %)	5/21 (23.8 %)	30/101 (29.7 %)

ANOVA was the most frequently used methodology to investigate the change or difference in writing ability. It was conducted to examine the difference between before and after an experiment by comparing different assessment conditions. As a result, approximately

20 to 28 % of the studies on peer assessment and comparative studies of self-assessment and peer assessment reported that these types of student assessment resulted in an improvement in writing quality. On the other hand, about 2 to 10 % of the studies on self-assessment concluded that self-assessment did not have an effect on the development of writing ability. These results suggest that peer assessment might have a more positive effect on the development of writing ability than self-assessment, but the studies about peer assessment less frequently analysed the effect of peer assessment on the improvement of writing ability. Moreover, the general number of studies that focus on the effect of student assessment on the development of writing ability is relatively much smaller than the studies about reliability. Therefore, it should be examined how the two types of student assessment affect the improvement of student assessment.

2.3.1.3.4 Effectiveness of Self-assessment. According to the results of the meta-analysis of the present study (Table 2.4), 24 out of 87 studies (27.5 %) focused on the effectiveness of self-assessment. In line with the results of previous studies on self-assessment, it was reported that self-assessment increases the sense of achievement, directs students' attention to the study purpose and assessment criteria, and stimulates their motivation. For instance, Ross et al. (1998), included in Table 2.4, stated that self-assessment provides teachers with indispensable information about barriers to overcome. And another study reviewed in the current meta-analysis, Ross et al. (1998), also mentioned that self-assessment provides teachers with inner responses of students about their learning process and outcomes. The reason why the information produced from self-assessment is meaningful for teachers is that students' perception of assessment seems to be different from that of teachers. Therefore, teachers could gain invaluable information about student learning from different perspectives. Zarei and Usefli (2015) from the present meta-analysis echoed the effect of self-assessment, because self-assessment improves learners' goal orientation. In other words, self-assessment helps them to reflect on their work in terms of a set of goals (Black et al., 2004). Students can achieve a learning goal only if they understand that goal and can evaluate what they need to do to attain it. Accordingly, self-assessment is essential to learning. Black et al. (2004) mentioned that teachers should develop their students' self-assessment skills, because the most effective tasks should develop students' ability to think of a difference between their work and the learning goal.

To sum up, self-assessment by students is a useful tool for developing students' ability. Black and Wiliam (1998a) proposed three principles to make self-assessment more successful: (1) understanding the desired goal; (2) concrete indication of students' present learning position; and (3) understanding a strategy to close the gap between the present position and the desired learning goal. In addition, Black and Wiliam (1998a) added to the hints on the implementation of self-assessment. Specifically, students should be trained in self-assessment so that they can understand the main purpose of their learning and thereby grasp what they need to do to achieve it. Furthermore, self-assessment not only improves learners' goal orientation (Zarei & Usefli, 2015) but it also has a positive effect on learners' English learning and other learner affect.

2.3.1.3.5 Effectiveness of Peer Assessment. According to the results of the present meta-analysis (Table 2.4), 23 out of 76 studies (30.2 %) surveyed the effectiveness of peer assessment. For instance, Zarei and Mahdavi (2014), included in the present meta-analysis, suggested that peer assessment might be more effective than teacher assessment. In another review of the present meta-analysis, Tsui and Ng (2000) also supported that peer assessment could have a more positive effect on the improvement of writing skills than teacher assessment. This is because peer assessment promotes revision, as Orsmond et al. (2000) mentioned in the current meta-analysis. Table 2.4 also indicates that the impact of revision from peer assessment has been focused on in previous studies: nine out of 76 studies or 11.8 %. In other words, peer assessment could improve students' critical abilities. It has been stated by McLaughlin and Simpson (2004), reviewed in the present meta-analysis, that peer assessment enables learners to critically evaluate their peers' assigned work. Thus, peer assessment might be more beneficial for peer assessors than for the assessment receiver. According to Cho et al. (2006), included in the meta-analysis, this is because peer assessors cannot evaluate others' work without understanding assessment criteria and study goals. Seen from a different angle, Huerta-Macias (1995), reviewed in the meta-analysis, insisted that peer assessment is a variant of self-assessment, because peer assessors have opportunities to exchange evaluations between them, so mutual peer assessment enables them to reflect on their own performance in the form of self-assessment (Black et al., 2004).

Moreover, peer assessment enables students to build the joint construction of knowledge and the development of a high-order rational judgement and reasoning process while learning cooperatively and communicating mutually (Cheng & Warren, 2005). Thus, it is clear that there is a positive relationship between peer assessment and students' general improvement in their learning performance (Sluijsmans & Prins, 2006).

In sum, self- and peer assessment types are similarly effective in terms of their perception of assessment criteria. Self-assessment is considered to be effective in making students goal oriented, while peer assessment is useful for promoting revision by developing critical abilities. However, previous studies did not analyse how each student assessment type functions to enhance the goal orientation of self-assessment and revision by critical thinking in peer assessment, so the specific effect of student assessment on the improvement of writing ability should be investigated.

***2.3.1.3.6 Learner Affect of Self-assessment**.* This section specifically explains the effect of self-assessment on learner affect included in the present meta-analysis (Table 2.4). In detail, learner anxiety, motivation and attitudes were the foci of self-assessment as reviewed in this study (17 out of 87 studies or 19.5 %). For instance, in the following studies included in Table 2.4, Butler and Lee (2010) found the augmentation of self-confidence, and Brown (2005) justified the effectiveness of self-assessment in terms of autonomy, intrinsic motivation and cooperative learning. Furthermore, Butler and Lee (2010) stated that self-assessment has a positive effect on learners' English learning and self-confidence. They investigated the effectiveness of self-assessment among 254 elementary school students who learned English as a foreign language. This finding is also supported by other reviewed previous studies in the meta-analysis (Andrade & Valtcheva, 2009; Léger, 2009; McMillan & Hearn, 2008). For example, Léger (2009) examined the effect of self-assessment on 32 university students and found that self-assessment heightened self-perception, which helped to develop learners' self-confidence. Moreover, according to Brown (2005), who investigated the effect of self-assessment and self-annotating on university students, self-assessment is theoretically justified to build principles of autonomy, intrinsic motivation and cooperative learning (Brown, 2005).

2.3.1.3.7 Learner Affect of Peer Assessment. A total of 22 out of 76 studies (28.9 %) pointed out the effect of peer assessment on learner affect. Reportedly, peer assessment can develop students' sense of responsibility. For example, Cho et al. (2006), reviewed in the meta-analysis, found that peer assessment can help students develop responsibility for their own learning and evaluation skills. As another example included in Table 2.4, Kwan and Leung (1996) also stated that peer assessment gives students a chance to take responsibility for analysing, monitoring and assessing their learning process and outcomes. And a further study reviewed in the meta-analysis, Saito and Fujita (2004), investigated the characteristics and assessor acceptance of peer assessment of 47 college students. They found that peer assessment enables students to take responsibility for managing the assessment process for their classmates and this leads to increasing responsibility for learning.

It has been stated that peer assessment also has a positive effect on learning affect and behaviour. Table 2.4 shows learner affect such as anxiety, self-efficacy and psychological safety. A total of 22 out of 76 studies (54.5 %) focused on those items. As an example of positive effect on learner affect included in Table 2.4, van Gennip, Segers, and Tillema (2010) investigated the effect of peer assessment on interpersonal variables of 62 vocational students, and found a change in psychological safety, value diversity and trust in the peer assessor group. Other studies also reported the augmentation of psychological safety effected by peer assessment, because peer assessment is employed between people of a similar status and can decrease students' tension (Edmondson, 1999; van Gennip et al., 2010). As Harlen and Winter (2004) stated, peer assessment helps learners to recognize each other's strengths and build situations where they can support and complement each other (Harlen & Winter, 2004). With regard to autonomy, Little (2009) stated that peer assessment may foster learners' autonomy because they learn how to take responsibility in assessment. Thus, peer assessment can help students to promote "a sense of ownership and responsibility, motivation and reflection of the student's own learning" (Falchikov & Goldfinch, 2000; Saito & Fujita, 2009, p. 151).

Unlike self-assessment, peer assessment encourages learners to work together. By doing so, learners improve not only their learning but also their collaborative learning skills (Topping, 1998). In other words, peer assessors become socialized and enhance interpersonal relationships and trust between peers (Earl, 2006; van Gennip & Tillema, 2010).

In sum, the number of studies of peer assessment related to learner affect is larger than that of self-assessment: 19.5 % for self-assessment and 28.9 % for peer assessment. The gap between those numbers might be caused by the studies about interaction with peers, while self-assessment did not focus on the relationship with peers. Instead, self-assessment tended to focus on language anxiety (10 out of 87 studies or 11.4 % for the self-assessment group; one out of 76 studies or 1.3 % for the peer assessment group; and 0 out of 41 studies or 0 % for comparative studies). The number of studies about learner autonomy was also extremely small according to Table 2.4 (one out of 87 studies or 1.1 % for self-assessment; 0 out of 76 studies or 0 % for peer assessment; and one out of 41 studies or 2.4 % for comparative studies). Hence, none of the assessment types have apparently explored the effect of student assessment on language learning anxiety and learner autonomy yet, so it is necessary to analyse those learner affects in order to implement student assessment more effectively.

2.3.1.3.8 Comparative Studies of Self-assessment and Peer Assessment versus Teacher Assessment. Hence, previous studies about self- and peer assessment have respectively presented strengths and different qualities. As Table 2.4 shows, comparative studies between self- and peer assessment tended to focus on the agreement with teacher assessment (nine out of 41 studies or 21.9 %), so the research questions were mostly related to the reliability or severity of student assessment used as an alternative to teacher assessment (14 out of 41 studies or 34.1 %). Also, the effectiveness (six out of 41 studies or 14.6 %) and the development of English proficiency (six out of 41 studies or 14.6 %) were compared through drawing a parallel between self- and peer assessment.

As regards comparing reliability between self-, peer and teacher assessment, Matsuno (2009), reviewed in the meta-analysis, examined the reliability of self- and peer assessment of 91 university students in comparison with assessments by four teachers. The study revealed that teacher and peer assessment are more consistent than self-assessment, and that peer assessment could be a useful tool in writing classes.

For instance, Sung et al. (2005), included in the meta-analysis, found that self-assessment and peer assessment commonly encourage students to reflect upon the amount of effort they make in their study process and how to apply the criteria to their actual performance. Orsmond et al. (2000), in the meta-analysis, proposed a way to make both types of assessment more successful: (1) present a clear definition of learning outcomes and assessment criteria; (2) ensure thoughtful communication with the student; (3) focus on the process of the assignment rather than just the product; and (4) concentrate on the process to overcome difficulties in the learning process. Saito and Fujita (2004), included in the meta-analysis, also mentioned the following common benefits of self-assessment and peer assessment: (1) obtaining information from diversified perspectives; (2) responsibility for controlling the assessment process; (3) directing students' attention to the assessment criteria; and (4) sharing individual strengths and weaknesses with students. Accordingly, both self-assessment and peer assessment could be useful tools for promoting responsibility for learning (Saito & Fujita, 2004). However, it should be noted that the amount of research on the comparison between self-assessment and peer assessment is much less than those studies that examine only self-assessment or peer assessment.

Self-assessment and peer assessment not only have common benefits but also different qualities. For instance, self-assessment has been shown to be somewhat idiosyncratic or unique, therefore its usefulness has been limited as a part of formal assessment. According to Matsuno (2009), reviewed in the meta-analysis, peer assessment, on the other hand, was shown to be relatively internally consistent and its evaluation patterns were not influenced by peer assessors' writing performance. Furthermore, peer assessment tended to produce fewer biased interactions, so it might play a useful role in writing classes. Self-assessment is usually a private activity that may not need knowledge of the study, while many peer assessment studies have been conducted for the assessment of oral presentations or professional practice. Thus, peer assessment is managed within a public domain where comparisons between performances and ranking of peers occur. In this case, according to Falchikov and Goldfinch (2000), students might find peer assessment difficult.

In terms of common shortcomings, self-assessment and peer assessment share weaknesses, though they have individual differences. For example, Matsuno (2009), included in the meta-analysis, stated that the following are common limitations of self-assessment and peer assessment: (1) difficulty in using it as a formal classroom assessment; (2) inter-rater reliability has been stressed rather than intra-rater consistency; (3) most previous studies have employed the traditional true-score approach, but it is difficult to separate the characteristics of examinees from the characteristics of a test. In brief, it is still doubtful as to whether they are evaluating examinees in the same way. The influence of cultures has also been discussed in terms of reliability. According to Blue (1994), some nationalities have effects on overestimation or underestimation of their language proficiency. For instance, Matsuno (2009) found that Japanese university students in EFL writing classes had a tendency to score their writing lower than teacher evaluation.

In sum, it is found that about half of the comparative studies investigated the reliability against teacher assessment in order to make them a substitute assessment. Also, comparative studies have the benefit of clarifying the differences and similarities between self- and peer assessment. Therefore, the present study adopts a comparative method to make the differences and commonalities distinct. In addition, reliability should be examined as many previous studies have done, because the reliability of assessment might be an indicator showing the extent to which students can understand assessment criteria.

2.3.1.4 Research Methods. Many statistical methods have been used in quantitative studies to examine the correlation and agreement between teacher assessment and student assessment (Table 2.7). Also, various types of qualitative methods have been conducted to explore the depth of learner affect (Table 2.7).

Table 2.7

Research Methods

Skills assessed	Self-assessment	Peer assessment	Self-assessment & peer assessment
Multi-trait multi-method	2	2	
Rasch model analysis	1	2	6
Correlation coefficient (Pearson/Spearman)	7		5
Wilcoxon matched-pairs signed-ranks test	1		
Regression analysis	3	1	
Analysis of covariance (ANCOVA)	1		
Analysis of variance (ANOVA)		11	
Multivariate analysis of variance (MANOVA)			1
Covariance analysis	1		
Cronbach's α	1		
t-test			2
Overall percentage agreement			1
SD/mean	5	2	
Observation	2	1	1
Interview	2	2	1
Recording/VTR		5	
Questionnaires			1
Sum	26	26	18

Note. The number of research methods that were explicitly depicted in the previous studies was counted.

In terms of quantitative methods, some studies conducted analysis of variance (ANOVA) to compare a treatment group and control group (Zarei & Usefli, 2015). With regard to qualitative study skills, observations, interviews (Logan, 2009), recordings and questionnaires (Lee, 2017) were used to explore the effectiveness of student assessment and to explore students' perception of writing influenced by peer assessment. According to previous studies on student assessment, both self-assessment and peer assessment showed a positive correlation with teacher assessment if self-assessment and peer assessment were well organized and students were well trained. However, correlation coefficients such as the Pearson product-moment correlation coefficient or Spearman's rho correlation coefficient were mostly used to examine the reliability of student assessment. The current study employs the many-facet Rasch measurement model to examine the difference between teacher assessment and two modes of student assessment, as previous comparative studies most frequently implemented this analytic approach (six out of 18 studies) to investigate how self- and peer assessment function compared with teacher assessment (Matsuno, 2009; Ravand & Ravand, 2016). The Rasch model could be an effective method for analysing differences such as task difficulty. Since the current study also aims to compare the effect of two types of student assessment, Rasch model analysis seems to provide the required information about the differences between the two types of assessment. It is also helpful to see the task difficulty, because the difficulty of tasks used in the research should be equalized. Therefore, it was decided to implement Rasch model analysis in the current study.

Another main purpose of the current study is to examine the effects of student assessment type (self- and peer assessment) on writing anxiety and learner autonomy. However, it was difficult to find previous studies analysing the relationship between learner affect variables such as these and student assessment. Most of the studies that investigated the relationship between student assessment and such learner affect variables depended on qualitative methods, that is, observation and interviews.

2.3.1.5 Limitation of Self- and Peer Assessment. Limitations of both assessment modes have been pointed out in previous studies. Specifically, self-assessors may not have a clear knowledge and understanding of the assessment criteria or they may not be well trained or may not have been explicitly instructed about how to use these criteria (Leach, 2012). Furthermore, Topping (2013) found that self-assessment marks tended to be higher than those of teacher assessment, and self-assessments were conducted based on the perceptions of the extent to which self-assessors marked effort rather than actual levels of achievement. Therefore, Topping (2013) pointed out the unreliability of self-assessment. Leach (2012) suggested that students' reluctance to self-evaluate was caused by the fear of being wrong in their self-assessment. Those negative attitudes toward self-assessment were also commented upon by Evans, Mckenna, and Oliver. (2005). Students are not fond of evaluating themselves and much prefer to receive teacher assessment (Evans et al., 2005). Blanche and Merino (1989) also mentioned that: (1) students' self-assessment ratings could vary depending on different external assessment criteria; (2) the accuracy of most students' self-estimates could vary, depending on their language proficiency; and (3) students who are more proficient tended to underestimate themselves, while less proficient students tended to overrate themselves. Underrating by more proficient students was also supported by other previous studies (Boud & Falchikov, 1989). Ross (2006) referred to the relationship between age and the reliability of self-assessment, i.e., self-assessment was less reliable when young learners were involved.

Several limitations of peer assessment have been highlighted in spite of its effectiveness. For instance, peer assessment has the potential to cause friction and hurt students' feelings because of ethical challenges, such as the racial or ethnic background of a peer assessor (Vu & Dall'Alba, 2007). Vu and Dall'Alba (2007) also mentioned that students may be reluctant to bear such responsibility by themselves, and worried about doing so. Some students may feel they are not qualified to judge others' work, so may hesitate to evaluate others (Orsmond et al., 1996). Freeman (1995) pointed out other negative aspects of peer assessment. For example, there is a possibility that other students do not take their responsibility seriously enough in assessing peers' work, which might lead to scepticism around the meaningfulness of peer assessment. Similarly to the limitation of self-assessment, peer assessors may have difficulty in understanding the assessment criteria or may not have sufficient training in how to apply the criteria (Orsmond et al., 1996). In addition, as Heywood (2000) stated, peer assessment may be less effective than teacher assessment in L2 learning, because students do not necessarily have a positive attitude towards assessing peers' language proficiency (Freeman, 1995; Orsmond et al., 1996; Vu & Dall'Alba, 2007).

In sum, it was found that the reliability and effectiveness of self-assessment and peer assessment are complicated by many factors, such as learner affect, rater training, learners' individual backgrounds and learners' ages. Those interwound conditions should be clarified in order to elucidate the efficacy of respective student assessment types and implement student assessment in class. Therefore, the present study prepares for three perspectives, namely reliability, writing ability and learner affect, to illuminate the effect of student assessment.

2.3.1.6 Summary of Previous Studies on Student Assessment. The results of the meta-analysis showed common or different qualities of self- and peer assessment. Many of the previous studies focused on the reliability and effectiveness of both types of assessment mode. The reliability of self-assessment was said to be mostly acceptable, but it depended on learners' rater training, educational backgrounds and language proficiency. Self-assessment could encourage students to be goal oriented and confident. In addition, the information about self-assessment could be especially valuable for teachers, because it helps teachers to clarify students' obstacles and decide on a second step for learners. On the other hand, it has been reported that the reliability of peer assessment could be generally positive, especially in global evaluation, but it depends on students' level of understanding of assessment criteria. Students can acquire a revision habit, critical ability, a sense of responsibility and cooperative learning skills.

However, many studies have focused on young adults (Table 2.2) and there have been fewer comparative studies than single studies, that is, self- or peer assessment studies. In order to elucidate the features of each assessment type, a comparative study between self- and peer assessment is considered to be more persuasive. Also, it is more meaningful to focus on high school students, because there have been fewer studies focusing on those subjects. Previous studies of comparative studies adopted Rasch model analysis, and this is advocated for investigating how self- and peer assessment work in comparison with teacher assessment. It could clarify the reliability and severity of each student assessment type, and learners' ability.

Most of the studies reviewed above yielded results supporting the effectiveness of student assessment, but they did not specifically discuss its usefulness and implementation. It is considered to be meaningful to discuss the application of two types of assessment based on the findings of each student assessment type.

2.3.2 Writing Assessment and Student Assessment

This section discusses how writing is assessed and one important component in particular, because the current study employs a scoring rubric in order to analyse the effect of student assessment. Weigle (2002) stated that the evaluation process is an essential component of writing assessment because the evaluation reflects raters' ultimate decision-making and inferences about writers. As the term "writing" is defined as "the acts of thinking, composing and encoding language into such texts" (Cumming, 1998, p. 61), writing performance is expected to be assessed in multiple aspects. Previous studies discussed the theoretical view that writing ability consists of multiple traits that should be reflected in many writing assessments. For instance, Wolf-Quintero, Inagaki, and Kim (1998) stated that writing comprises four components: writing fluency, accuracy, grammatical complexity and lexical complexity. Grant and Ginther (2000) proposed four elements in their writing assessment: essay length, grammatical structure, lexical specificity and lexical features. Schoonen et al. (2002) also identified three writing skills: linguistic knowledge, speed of processing linguistic knowledge and metacognitive knowledge.

When constructing writing tasks, especially in class, Hyland (1996) stated that clarification of the purpose of assessment of a task is key, because assessment provides students with opportunities to see what they have learned and to be sure about how their writing is evaluated. Therefore, according to Douglas (2000), a writing task is expected to include five factors: "(1) the specification of the objective; (2) the procedure for responding; (3) the task format; (4) the time allotted or the deadline for submission; and (5) the evaluation criteria" (p. 50). Hence, an assessment rubric should be created in association with the writing task. Oi (2019a) analysed the needs of Japanese high school English teachers in terms of writing assessment components. The results showed that Japanese high school English teachers stressed task fulfilment, grammatical correctness, appropriate usage of vocabulary, and structure and coherence. The current study supports the results of the needs analysis that organize assessment components for writing tasks in the classroom, so the assessment rubric of the current study implements Oi's assessment rubric (Oi, 2019a).

2.3.3 Holistic Assessment and Analytic Assessment

It is necessary to decide which type of rubric is more suitable to let students assess their own writing or their peers' writing and encourage them to develop English proficiency, so this section discusses characteristic features of holistic and analytic assessment. One of the important effects of rubric use is to promote learning. According to Arter, McTighe, and Guskey (2001), it is believed that the explicitness of criteria and standards is essential to encourage students to receive quality feedback, because rubrics can accelerate student learning. In using rubrics, it would be effective to promote learning and improve instruction, so that teachers and students understand the aims of a rubric and its utility. Therefore, rubrics should meet the aims of the assessment and be explicit to enable the provision of helpful feedback. Rubrics are mainly categorized into two types, holistic and analytic.

Firstly, in holistic assessment, the rater makes an overall judgement about the quality of performance. According to Hyland (1996), “it is based on a single, integrated scoring of writing behaviour, so it reflects the idea that writing is a single entity that is best captured by a single scale that integrates the inherent qualities of the writing” (p. 227). White (1985) mentioned that holistic writing is intended to focus the reader’s attention on the strengths of the writing, not on its shortcomings, so that writers are rewarded for what they write well. However, holistic assessment has been criticized because of uncertainty about the exact nature of the constructs of assessment criteria. For example, as Cumming, Kantor, and Powers (2001) stated, holistic assessment cannot identify complex traits and variables in writing, making it difficult to evaluate the quality of students’ writing performance.

On the other hand, analytic assessment helps the rater to assign a score to each element of the assessment criteria. According to Jonsson and Svingby (2007), analytic assessment “rubric seems to have the potential to promote learning and/or improve instruction” (p. 130). This is because analytic assessment makes study aims specific and explicit, and also enhances the utility of assessment. Hyland (1996) clearly defines analytic assessment as “the features to be assessed by separating, and sometimes weighting, individual components and it is therefore more effective in discriminating between weaker texts” (p. 229). However, the limitations of analytic assessment are also indicated, for instance, in the danger of the halo effect, “where results in rating one scale may influence the rating of others” (Cohen, 1994, p. 317; Hyland, 1996, p. 229; McNamara, 1996), though randomizing the order of scoring learner responses for each analytic scale prevents the halo effect of scoring.

Given the characteristics of these two types of rating scale, analytic assessment may be a more useful tool with which to assess high school students’ writing in the classroom because of the need to diagnose students’ weaknesses and strengths, though there are some arguments in favour of holistic assessment in terms of practicality, such as speed and lower cost. Weigle

(2002) commented that analytic assessment clarifies what is going on inside the writing assessment, because it provides teachers and students with more detailed diagnostic information.

Previous studies on student assessment in the classroom that employed analytic rating scales include Davidson and Henning (1985), Saito (2008), Matsuno (2009), Behjat and Yamini (2012) and Esfandiari and Myford (2013). These studies mainly employed five- to seven-point scales. For instance, Matsuno (2009) examined how self- and peer assessment function compared with teacher assessment, using a six-point analytic rating scale, which was a simplified version of the ESL composition profile developed by Jacobs, Zingraf, Wormuth, Hartfiel, and Hughey (1981). Matsumoto adopted this approach to adjust the rating scales to the participants' level that were weighted equally with a six-point rating scale, though the original ESL composition profile is a weighted scale. Matsuno referred to Kondo-Brown (2002) in the following: "It is not clear how the weightings were determined in the original version, and some researchers have questioned the assignment of different weights to evaluation criteria" (p. 9). This example shows that modifying or tailoring authorized assessment criteria to adjust to learners' conditions is a possible practical approach. In so doing, analytic assessment criteria, other information such as the length of writing and learners' educational backgrounds could be considered as key issues. The current study targeted high school students, but the majority of participants in the previous studies were at college or university level (Table 2.2). Therefore, it is necessary to consider the scales and tasks that best fit the characteristics of the student population in the current study.

2.3.4 Student Assessment as Formative Assessment

Since the current study aims to find an effective implementation of student assessment in formative assessment, this section discusses the meaning of student assessment as formative assessment. First, the relationship between student assessment and formative assessment should be considered.

Graesser, Pearson, and Magliano (1995) opined that the essentials of assessment are questioning the learning outcome and assessing the outcome in relation to the questions at a macro and micro level. In order to make self-assessment succeed as an effective formative assessment device, it is not recommended to just depend on a student's ability to evaluate according to criteria (Orsmond et al., 1996). Peer assessment is also formatively employed when students are involved in the learning process (Topping et al., 2000). For instance, peer assessment requires students to think, compare, communicate, review, summarize, clarify, give feedback, diagnose and identify in the learning process. Topping (1998) called these "cognitively demanding activities" that lead students to a deep understanding. Orsmond et al. (2000) also insisted: "To make the introduction of both self-assessment and peer assessment methods more palatable and to overcome potential academic inertia, it is helpful to concentrate on the process rather than the outcome of assessment" (p. 24).

As the reason why the process is stressed, McDonald and Boud (2003) stated that formative assessment is aimed at improving learning rather than judging final achievement, because formative assessment is constructed to give sufficient feedback and help for learning (Falchikov, 1995). According to Topping (2009), formative assessment facilitates ongoing assessment meeting students' needs and develops their learning during a course. Orsmond et al. (2000) stated that formative assessment mainly works to provide information for learners about their progress, whereas summative assessment is an end point assessment for grading students and provides students with information, indicating the extent to which they have been successful in obtaining information.

In sum, the aim of formative assessment is not to identify success or failure at the end of events but to enhance the effectiveness of learning. Such assessment is intended to help students design their own learning, identify their own strengths and weaknesses, target areas for remedial action and develop metacognitive skills (Boud, 1990; Brown & Knight, 1994). Hence, the purpose of formative assessment is in line with self-assessment and peer assessment.

2.4 Writing Anxiety

It is hypothesized that student assessment (self-assessment and peer assessment) would stimulate and change a student's belief about English writing, because self-assessment and peer assessment provide students with opportunities to reflect on their own work and be given new perspectives by peers. To be specific, student assessment encourages a student to write in English and gives a student an objective analysis about themselves. On the other hand, Cheng (2002) argued that the link between low self-confidence and anxiety should be specified because the link helps learners to find a supportive environment where learners' self-confidence is likely to be enhanced (Cheng, 2002).

Previous studies on self-assessment also reported a reduction in anxiety about performance (MacIntyre, Noels, & Clément, 1997; McDonald & Boud, 2003; Tsui & Ng, 2000), but none of them discussed writing anxiety. Previous studies about peer assessment analysed affective factors such as anxiety and learning awareness. However, they did not anlayse writing anxiety in the context of self-assessment. Therefore, it is meaningful to investigate anxiety about English writing and to analyse the different effects that the two types of student assessment may have on writing in order to implement student assessment in the classroom assessment smoothly.

2.4.1 What is Foreign Language Learning Anxiety?

Anxiety is considered to have an impact on learning a foreign language, because it is presumed that anxiety can lead to lower self-efficacy due to the fear of possible failure (Bandura, 1986, 1989; Schunk, 2007). Woodrow (2011) explained that anxiety is negatively linked with self-efficacy, which, however, is positively connected with writing performance. In other words, self-confidence is closely linked with a lack of anxiety (Clément, 1980). In short, self-efficacy and anxiety are important variables in learning English as a foreign language (FL) or second language (L2). Woodrow (2011) examined the relationship between writing self-efficacy, writing anxiety and writing performance using a structural model. Woodrow (2011) found that writing anxiety is not directly linked with writing performance, but self-efficacy could be a predictor of writing performance.

Perspectives of language learning anxiety have also been discussed in many ways. Spielberger (1983) defined trait anxiety as the possibility that a person becomes anxious in any situation. For instance, it threatens cognitive functioning, impairs memory, induces avoidance behaviours and can lead to other consequences (Eysenck, 1979). According to MacIntyre and Gardner (1991), state anxiety is considered to be a mixture of the trait and situational approaches. Situation-specific anxiety can be seen as measures of trait anxiety in some circumstances. For example, public speaking presents a well-defined situation in which speakers might feel apprehensive.

Research into language anxiety has been conducted since the 1970s. Over the past two decades, first Scovel (1978) reported conflicting and worrying results about how anxiety is linked with second language learning, because the studies about language anxiety had been discussed from only one perspective, that is, one affective variable, though affective variables could be complicated in latent and observable quality. Therefore, Scovel suggested that language learning anxiety should be distinctly classified through global and comprehensive investigation. Based on this idea, language learning anxiety has been discussed in many educational contexts (Gardner, 1985; Horwitz, 1986; Horwitz, Horwitz, & Cope, 1986; MacIntyre & Gardner, 1989; Proulx, 1991). For example, MacIntyre and Gardner (1989) investigated the relationship between foreign language anxiety and performance in 104 university students who were studying French, using factor analysis. They found two statistically independent types of anxiety: general anxiety and communicative anxiety. Also, Horwitz et al. (1986) tried to identify the effect of anxiety on language learning focusing on 225 university students studying foreign languages. They employed the Foreign Language Classroom Anxiety Scale (FLCAS) developed by Horwitz (1986) and found that communication apprehension, test anxiety and fear of negative evaluation in the foreign language classroom have effects on language performance.

Language learning anxiety has also been multidimensionally conceptualized. In previous studies, the different aspects of human behaviour and intellectual performance have been investigated as well as the effects of anxiety (Morris, Davis, & Hutchings, 1981; Smith & Smoll, 1990). For instance, test anxiety (Liebert & Morris, 1967; Morris & Engle, 1981; Sarason & Sarason, 1990), speech anxiety (Fremouw & Breitenstein, 1990) and sport performance anxiety (Smith & Smoll, 1990) were enumerated. Anxiety symptoms were categorized into several relatively independent perspectives, including somatic/physiological (e.g., upset stomach, pounding heart, excessive sweating and numbness), cognitive (e.g., worry, preoccupation and negative expectation) and behavioural (e.g., procrastination, withdrawal and avoidance) aspects. Referring to these categories of language learning anxiety, Xiao and Wong (2014) investigated writing anxiety in 87 Chinese language learning students and found that avoidance was the strongest through factor analysis.

Among the different dimensions of language learning anxiety, Clément (1980) proposed that self-confidence is highly placed in the construct of the Social Context Model of second language learning. Clément (1980) also stated that self-confidence is related to second language use, classroom anxiety and self-evaluations of second language proficiency. In brief, Clément (1980) reported that self-confidence might be a subcomponent of the second language classroom anxiety or second language writing anxiety construct. Hence, language anxiety could be evidence of a continuous relationship between low self-confidence and anxiety. The experience of second language anxiety might be an obstacle to the progress of language learning.

MacIntyre and Gardner (1989, 1991) stated that language learning anxiety is unique to the context of language learning and use, so it is distinctly different from general and other conceptualizations of anxiety, such as general anxiety or test anxiety. It was also reported that anxiety could work positively or negatively for language learning. In other words, positive anxiety promotes study achievement and negative anxiety hinders it (Bandura, 1986; MacIntyre, 1999). According to Krashen (1985), anxiety contributes to an affective filter that prevents the individual from being unreceptive to language input, thus “the learner fails to ‘take in’ the available target language messages and language acquisition is not developed” (Horwitz et al., 1986, p. 127). In short, foreign language anxiety is composed of beliefs, perceptions and feelings that are produced in the process of foreign language learning in the classroom, so it does not just mean a collection of other anxieties (Horwitz et al., 1986). In other words, language learning itself might place individuals in a deeply unstable psychological condition because it could directly affect an individual’s self-conception and view of the world (Guiora, 1983). Most research into anxiety has found that anxiety negatively influences language performance (Horwitz, 2001; Horwitz et al., 1986; MacIntyre, 1999; Young, 1986). The majority of research into foreign language anxiety has focused on its relationship with speaking skills (Horwitz et al., 1986; Philips, 1992; Woodrow, 2006). The majority of research into language learning anxiety refers to the learning of languages other than English (Horwitz, 2001). There is relatively little research into the area of the relationship between anxiety and writing.

The present study examines anxiety because it might play an important role in L2 writing among Japanese high school students. As mentioned above, writing apprehension is a distinct form of anxiety, unique to written communication (Bline et al., 2001; Burgoon & Hale, 1983a, 1983b; Daly & Wilson, 1983). As Daly and Shamo (1976, 1978) stated, the relationship between writing anxiety and writing performance is more complicated than other learning anxiety, and the effect of writing apprehension on writing quality is strong in many variables

(Daly & Shamo, 1976, 1978). Therefore, anxiety could influence self-confidence in order to determine one's willingness to communicate in a first language (L1) and in a second language (L2) (MacIntyre, Clément, Dörnyei, & Noels, 1998).

Students with low self-confidence might tend to underestimate their ability when they study a second language and have negative expectations about their performance. They might feel insecure or apprehensive when they are confronted with language-learning tasks (MacIntyre et al., 1997). Cheng (2002) also mentioned that anxiety-producing tasks might represent a vicious cycle for learners, because such tasks threaten their progress in language learning, and thus learners could find their self-confidence further undermined. In summary, language learning anxiety could exert a negative effect on learners if learning anxiety does not support learners.

2.4.2 The Definition of Writing Anxiety

Writing anxiety is of particular importance in this study, because the study aims to analyse the relationship between student assessment and writing anxiety. The negative effect of writing anxiety is most likely to appear when apprehensive learners write compositions under time pressure (Kean, Gylnn, & Britton, 1987) or narrative-descriptive topics that require disclosure of personal feelings, attitudes and experience (Faigley, Daly, & Witte, 1981). Faigley et al. (1981) reported that there were no performance differences between highly apprehensive writers writing about argumentative topics, but a large number of studies have shown that writing anxiety is negatively related to writing processes (Bannister, 1992; Bloom, 1980).

The relationship between anxiety and self-efficacy has previously been discussed by researchers. Self-efficacy is defined as "the perception of abilities to perform actions at a particular level" (Bandura, 1986; Woodrow, 2011, p. 511). As an example, a student may believe that she/he will present high achievement in a test when they have self-efficacy. According to

Bandura (1989), self-efficacy can be a greater determinant than performance ability. Self-efficacy is a strong predictor of academic performance (Bong, 2002). On the other hand, it is presumed that anxiety can impair self-efficacy due to thoughts of possible failure (Bandura, 1986; Schunk, 2007). Hence, self-efficacy influences the choices and actions of individuals. As Woodrow (2011) stated, the relationship between writing performance and anxiety has been mediated by self-efficacy beliefs: negative affect is conceptualized as anxiety and positive affect as self-efficacy.

From a different perspective, Woodrow (2011) indicated that anxious students are more likely to feel parental pressure and underestimate their effort. Littlewood (1999) reported that eastern Asian students have a tendency to expect the teacher to be the holder of authority and knowledge giver and to be responsible for the assessment of students' learning.

2.4.3 Previous Studies about L1 Writing Anxiety

Since the 1970s, research has been conducted on the relationship between writing anxiety and personality characteristics and has justified the distinctiveness of writing anxiety, which is unique to written communication. These studies were conducted within the context of L1 writing. In other words, they mainly focused on the writing anxiety of first language learners, especially native speakers of English in the United States. While there are fewer studies about writing anxiety than about anxiety connected with other skills, Hayes (1996) highlighted the importance of motivation and apprehension in the writing process. It was proposed that writing apprehension is defined as negative feelings that writers experience when they try to produce ideas and words (Bline et al., 2001; Burgoon & Hale, 1983a, 1983b; Daly & Wilson, 1983; Tsao, Tseng, & Wang, 2017). Madigan, Linton, and Johnson (1996) highlighted two effects of writing apprehension: (1) distress related to writing; and (2) a profound aversion to the writing process. On the other hand, it was reported that there were no performance differences between highly apprehensive and not very apprehensive writers in argumentative topics (Faigley et al., 1981; Madigan et al., 1996).

2.4.4 Previous Studies about L2 Writing Anxiety

In contrast to the wealth of studies on L1 writing apprehension above, there have been only a few studies that directly address second or foreign language (L2) writing anxiety, although the role of language anxiety in L2 learning has been the subject of considerable research (Cheng, 2002). In other words, the research on L2 writing anxiety is still underdeveloped and further studies are expected to deepen the understanding of L2 writing anxiety (Cheng, 2002).

Studies on L2 writing anxiety so far have yielded mixed results regarding: (1) the relationship between L2 writing anxiety and L2 writing performance (Masney & Foxall, 1992; Wu, 1992); (2) negative feelings about content (Gungle & Taylor, 1989; Masny & Foxall, 1992); (3) interest in taking more advanced L2 writing courses; and (4) awareness of L2 writing demands in one's final examination (Gungle & Taylor, 1989). For instance, Masney and Foxall (1992) investigated the relationship between writing anxiety and academic achievement focusing on 28 university students who were learners of English as a second language. They found that students with high scores showed lower anxiety, while students with low scores presented higher anxiety. Another important finding is that L2 writing anxiety was relatively different from L2 speaking anxiety. Cheng, Horwitz, and Schallert (1999) showed, for instance, that second language classroom anxiety and second language writing anxiety are mutually linked but they are independent anxiety traits. In addition, they also found that speaking anxiety is strongly interrelated with language classroom anxiety. However, writing anxiety is connected with skill-specific anxiety.

Second language classroom anxiety and second language writing anxiety may be two relatively independent constructs (Burgoon & Hale, 1983a, 1983b). The majority of research into foreign language anxiety has focused on its relationship with speaking skills. While there is some evidence from research of the influence of anxiety on listening, there is relatively little research in the area of anxiety and writing. As previously mentioned, most studies on writing anxiety have been conducted with first language learners, particularly with native speakers of English in the US. In contrast, there have been only a few studies that focus specifically on L2 writing anxiety, particularly on writing anxiety conducted in FL contexts (Cheng, 2002; Negari & Rezaabadi, 2012; Tsao et al., 2017). An example is the investigation into whether the writing anxiety of EFL college students was affected by self-evaluative judgements of corrective feedback (Tsao et al., 2017). Those previous studies reported inconsistent results regarding second language writing performance (Negari & Rezaabadi, 2012), dimensions of writing anxiety (Güneyli, 2016) and predictors of writing anxiety (Goodman & Cirka, 2009; Tsai, 2008). To sum up, the number of research studies about writing anxiety that focused especially on foreign or second language learners is lower than that of other skills; furthermore, it is still difficult to find an established theory on the writing anxiety of foreign or second language learners.

2.4.5 The Development of Scales for L1 Writing Anxiety

Although the research on writing anxiety for L2 learners has yielded important insights into how writing anxiety has an effect on foreign language learning, the methodology employed in many of the few previous L2 writing anxiety studies was not without flaws. First, for example, the Daly-Miller Writing Apprehension Test (WAT), which was developed for native speakers of English by Daly and Miller (1975), was administered to examine the writing anxiety of ESL or EFL learners. Another limitation of previous L2 writing anxiety studies is that instruments devised to examine second language anxiety might be influenced by items addressing anxiety in speaking a second language mainly related to the classroom environment (Dennis, Lowe, Meixner, Nouri, & Pearce, 2001). Therefore, it is doubtful whether these second language anxiety measurements could identify learners' skill-specific anxiety other than speaking anxiety (Dennis et al., 2001). Accordingly, language anxiety should be specified in reference to the characteristics of language skills. This is because communication anxiety can be impacted by the mode of communication (Burgoon & Hale, 1983a, 1983b).

As previously mentioned, the Daly-Miller WAT for L1 learners has been widely accepted as reflecting the anxiety of native language writers; however, the conflicting results of the WAT have also been reported by previous studies focusing on L1 (Table 2.8). All of the items on the WAT are expected to load either on one factor or on one component in a factor analysis (Daly & Miller, 1975). The conflicting results have highlighted: (1) the need to identify the number and quality of dimensions in the WAT; (2) the construct validity of the WAT (Bline et al., 2001; Shaver, 1990).

Table 2.8

The Conflicting Results of L1 Writing Anxiety using the WAT

Researchers (Year)	Samples	Results
Burgoon & Hale (1983a, 1983b)	257 L1 university students in the U.S.	3 dimensions: (1) discomfort or ease in writing; (2) enjoyment of writing; and (3) rewards of writing
Shaver (1990)	1100 L1 high school students in the U.S.	Principal component analyses of the WAT on L1 learners: (1) writing self-concept; (2) affective performance reaction; and (3) reaction to evaluation →low self-confidence is a major component of the WAT The WAT as a specific measure of writing anxiety may be problematic
Bline et al. (2001)	1128 L1 university students in the U.S.	2-factor solutions have been obtained The WAT is still subject to debate

Despite the conflicting results and being originally developed for L1 writing anxiety, the Daly-Miller WAT has been widely accepted for L2 writing anxiety. This is because, according to some previous L2 writing anxiety studies, the WAT has manifested satisfactory reliability with internal consistency as well as stable and predictive validity (Cheng et al., 1999; Lee, 2001; Wu, 1992). Yet, the need for further improvement has been indicated as follows: (1) the WAT was originally developed for L1 learners and is not appropriate for second language writing anxiety (McKain, 1991); (2) questions have been raised about the construct validity of the WAT (McKain, 1991); and (3) the WAT has difficulty in discriminating the causal links between anxiety and self-confidence (Horwitz et al., 1986).

Therefore, several studies and inventories have been devised to develop a scale for L2 writing anxiety based on the WAT. For instance, McKain's Writing Anxiety Questionnaire (WAQ, 1991) was used as the measurement instrument for second language writing anxiety. McKain (1991) developed an L1 writing anxiety scale by borrowing 12 items from the WAT and Holland’s (1978) Writing Problem Profile. The 12 items were chosen on the basis of coding that was employed by at least three out of four independent raters. When coding, “anxious feelings” associated with writing, as well as other aspects of writing such as behaviour (e.g., avoidance of writing) or cognition (e.g., beliefs about one’s writing ability), were included. As another example, the WAQ was presumed to have a unidimensional structure, but it needs validation through factor analysis. Also, the WAQ, like the WAT, was not developed specifically for L2 learners, so its applicability in the L2 context is questionable. Therefore, the scale for measuring L2 writing anxiety still has room for development and much work is needed to gain a better understanding of L2 writing anxiety.

2.4.6 Scales of L2 Writing Anxiety

Scales of second language learning anxiety have been developed to examine anxiety related to specific language skills because of increasing attention to identifying language learning anxiety (Table 2.9). However, language skill-specific anxiety seems to be more associated with oral aspects of L2 use (Cheng et al., 1999; Horwitz, 2001). As can be seen in Table 2.9, questionnaires about L2 language learning anxiety have been devised to investigate speaking or listening anxiety related to L2 use (Fremouw & Breitenstein, 1990; Horwitz, 2001; Kim, 2000; Vogely, 1998). Therefore, Cheng et al. (1999) and Cheng (2004) tried to invent a scale of L2 writing anxiety through a series of trials (Table 2.9). Above all, the Second Language Writing Apprehension Inventory (SLWAI), which was developed by Cheng (2004), is the only questionnaire designed to investigate the writing anxiety of L2/FL learners. Cheng et al. (1999) translated the Daly-Miller WAT and extended it for L2 learners. It was called the “Second Language Writing Apprehension Test” (SLWAT) and consisted of 29 items analysing how anxious an individual feels when they write in an L2. However, the SLWAT also has the same problem in terms of construct validity.

Table 2.9

Questionnaires about L2 Learning Anxiety

Year	Researchers	Name of questionnaire	Aims & features
1986	Horwitz et al.	Foreign Language Classroom Anxiety Scale (FLCAS)	To conceptualize 2nd/foreign language anxiety as a unique form of anxiety specific to the L2 learning context To measure the degree to which a student feels anxious in a foreign language class Several instruments developed based on FLCAS 33 items: primary measures of anxiety related to speaking situations (Aida, 1994)
1990	Fremouw & Breitenstein		To examine speech anxiety
1998	Vogely		To examine anxiety associated with listening
1999	Cheng et al.	Second Language Writing Apprehension Test (SLWAT)	To examine anxiety associated with writing
2000	Kim		To examine anxiety associated with listening
2004	Cheng	The Second Language Writing Apprehension Inventory (SLWAI)	To examine anxiety associated with L2 writing

Cheng (2004) developed the Second Language Writing Anxiety Inventory (SLWAI) to assess the level of anxiety that students feel toward second language writing. The SLWAI has been depicted as a language skill-specific anxiety scale due to a higher correlation with writing achievement. The SLWAI was carefully developed in Cheng's study to investigate the effects of writing anxiety among second language learners on human beheviours and academic writing performance. The study was designed for a total of 421 freshmen majoring in English from seven different universities in Taiwan, and determined the construct, which consisted of three subscales: somatic anxiety, cognitive anxiety and avoidance anxiety.

In the pilot phase of Cheng's (2004) study, 65 English foreign language (EFL) learners' responses about L2 writing anxiety were applied for a stock of scale items. Then, these items were analysed to help create a preliminary version of the L2 writing anxiety scale for the main study. In order to design the final SLWAI, exploratory factor analysis was conducted and the results were refined and elaborated. All scales and subscales were assessed by means of correlation and factor analysis to establish the reliability and validity of the SLWAI. Two separate principal axis factoring analyses extracted three factors with adequacy. The solutions of the three-factor amount to 47 % and 48 % of the common variance in the two analyses, respectively. The correlations among these three factors were acceptable (r = .32, .32 and .53 for the first set of factors, and r = .37, .38 and .48 for the second set). To analyse validity, correlations with other anxiety-related sales were evaluated to test the convergent and discriminant validity of the 22-item SLWAI. Generally, the correlation between the SLWAI and other measurements associated with writing was higher than L2 anxiety measurements not related to writing. In contrast to the SLWAT and the translated version of the WAT, the SLWAI shows high intercorrelation with other scales, tests and questionnaires. Thus, the development of the SLWAI presented good reliability and adequate validity. The final version of the SLWAI (Cheng, 2004) consists of 22 items and was developed to gauge the level of anxiety that students feel about second language writing.

The total of 22 items of the SLWAI (Cheng, 2004) comprises the three states of anxiety: (1) somatic anxiety (one's awareness of the physiological effects of the anxiety experience as reflected in increased automatic arousal, and uncomfortable feeling states, such as nervousness and tension); (2) avoidance behaviour anxiety (presenting an avoidance pattern toward writing);

and (3) cognitive anxiety (covering negative expectation, fear or worry about negative evaluation, and tests). A Likert-type response format was adopted consisting of a five-choice response scale corresponding to 1 (strongly disagree), 2 (disagree), 3 (no strong feelings either way), 4 (agree) and 5 (strongly agree). The higher the score obtained by the subscales and the total score of the SLWAI, the higher the level of writing anxiety, so the SLWAI is considered to be reliable and valid for measuring L2 writing anxiety in terms of the design and the results of factor analysis.

2.5 Learner Autonomy

Learners' beliefs and experience have a great impact on learning behaviours. In other words, "autonomous learning behaviour may be supported by a particular set of beliefs or behaviours" (Cotterall, 1995, p. 196). Learner autonomy is defined as the ability to take control of one's own learning. Taking control of one's own learning means having the responsibility for all the decisions about all aspects of learning: determining the objectives, defining the contents and progression, choosing the methods and skills to be used, monitoring the acquisition steps and evaluating the content of learning. Benson (2001) also defined learner autonomy as the capacity to take control of one's own learning. Benson also stated that autonomy in language learning should be considered in terms of three key levels at which learner control may be exercised: (1) control of learning management; (2) control of cognitive processes; and (3) control of learning content. According to Little (1995), learner autonomy is based on the acceptance of learning responsibility. In other words, individual learners must determine what and how they want to learn, set their learning targets, reconsider their learning and so on. The keys to learner autonomy are: (1) selecting learning materials; (2) taking private control of learning goals; (3) planning when and how to study each study goal; (4) evaluating progress and attainment; (5) assessing the learning curriculum (Dickinson, 1987; Holec, 1989).

By taking control of their learning, learners learn how to decide for themselves what and how to learn: for instance, they need to understand what they need, reconsider their learning critically and make best use of the opportunities to practise English inside or outside the classroom (Benson, 2001; Dickinson, 1987; Holec, 1981; Little, 1991). These are the essential principles of learner autonomy. As Wenden (1987) stated, learner autonomy is generally recognized as an important educational goal, and students' language proficiency is influenced by learner autonomy (Dafei, 2007). Ho and Crookall (1995) also insisted that autonomy provides learners with the opportunity to take responsibility for their learning (i.e., make decisions about their learning, plan, evaluate, monitor and assess). In brief, autonomous learners can become reflective learners, so they can overcome even temporary demotivation (Little, 2002). According to Benson (2001), an intimate relationship can be seen between autonomy and effective learning (Benson, 2001).

When the learning of the classroom is observed, the traditional teacher feedback on writing does not work well because of the demotivation of students and lack of learner autonomy (Benson, 2001). It is believed that student assessment such as self-assessment and peer assessment can promote learner autonomy. This is because self-assessment can develop learners' awareness of the quality of writing and help students become self-critical writers. Peer assessment can give students multi-perspective feedback and help them develop critical thinking towards their goal, namely writing. Hence, by changing classroom assessment from a teacher-centred to a student-centred approach, it is presumed that students are motivated to make a decision, produce a learning process and product, and become active learners. Holec (1981) stated that an autonomous learner is someone who can control his or her own learning. He also stated that an autonomous learner takes the responsibility for all decisions about all aspects of learning.

Kim and Kim (2005) investigated the relationship between learner autonomy and the practical English of university students. Kim (2009) also investigated how to encourage learner autonomy through journals and found that university students were interested in how to become autonomous learners and develop self-confidence. Spratt, Humphreys, and Chan (2002) analysed the perspectives of responsibility of students and teachers and found that the views on responsibility were connected to learners' confidence.

Deci and Ryan (1985, 1991) stated that no one should feel supervised by power or authority, but rather everyone needs to feel autonomous. Feeling autonomous means that people make decisions about their behaviour at the highest level of self-reflection. According to self-determination theory (SDT; Deci & Ryan, 1985), people feel free in terms of self-determination when they do something interesting, personally important and energizing (Deci & Ryan, 2006). Also, SDT suggests to us that students' internalization of learning motivation should be constructed by satisfaction of students' basic psychological need for autonomy, competence and relatedness. In the classroom context, Niemiec and Ryan (2009) commented that improving autonomy encourages students to be intrinsically motivated and more actively involved in less interesting tasks.

However, only a few previous studies have discussed the influence of student assessment on learner autonomy in the context of L2 learning (Ashraf & Mahdinezhad, 2015; Lee, 2017). They reported that self-assessment and peer assessment had positive effects on the development of learner autonomy, English speaking and writing. Therefore, it is difficult to draw a conclusion as to whether there is a correlation between student assessment and learner autonomy. Moreover, these previous studies focused on university students, so it is beneficial to investigate the effect of student assessment on learner autonomy in high school students in particular. A study targeting adolescent students would be beneficial, because learning autonomously is ideal for all students and promotes a lifelong learning society. Borg and Al-Busaidi (2012) also stated that the benefits of learner autonomy expedite the development of high-quality language learning, democratic societies and individuals' lifelong learning. To sum up, learner autonomy helps students to learn autonomously in the classroom and indeed anywhere else.

2.5.1 Learner Autonomy Scales for L2 Learners

In the early stage of language learner autonomy research, Cotterall (1995, 1999) proposed a questionnaire about learner beliefs in terms of learner autonomy. The list of questions to ask about learners' beliefs was designed based on a series of interviews with ESL students about their experience of language learning. Six factors were identified in Cotterall's factor analysis: "(1) role of the teacher; (2) role of feedback; (3) learner independence; (4) learner confidence in study ability; (5) experience of language learning; (6) approach to studying" (Cotterall, 1999, p. 3). Cotterall's questionnaires served to illuminate the relationship between each factor and autonomous language learning behaviour but could not highlight learners' metacognitive strategies for monitoring and assessing their learning process and outcome.

Chan, Spratt, and Humphreys (2002) conducted a study on learner autonomy with a group of university students learning English in Hong Kong. Chan et al. also designed a questionnaire to investigate the view on responsibilities and decision-making abilities in learning English based on the studies of Deci (1995), Deci and Ryan (1985), Holec (1981) and Littlewood (1999). The reason why Chan et al. were given input from those previous studies was that those studies focused on intrinsic motivation, and autonomy is an essential component of intrinsic motivation. Furthermore, motivation is considered to take a main role as a necessary forerunner of autonomy (Deci & Ryan, 1985). The learner autonomy scale designed by Chan et al. (2002) consisted of 52 questions, divided into four main sections: (1) students' perceptions of the English teacher's responsibilities and their own; (2) their perceptions of their decision-making abilities; (3) their motivation to study English; and (4) how often they carried out different autonomous activities in and outside class (p. 3). In the study, they found that students had firm and clear viewpoints about teachers' role and students' responsibility. However, a total of 13 items among the 52 questions were allotted to asking about the teacher's role in the scale of Chan et al. (2002), so it seems to be difficult to discover an individual learner's views on learner autonomy that are not influenced by the presence of a teacher.

Chang (2007) also developed a questionnaire about learner autonomy to obtain a measure of individual learners' level of autonomy based on Chan el al.'s study (2002). Chang's study aimed to survey the effect of group processes on learner autonomy based on data obtained from 152 Taiwanese university students. As a background to Chang's study, there is an idea that "most learning situations, especially in schools or universities, take place in groups. The context of the group may exert many influences upon the individuals within" (p. 323). Dörnyei and Malderez (1999) also claimed: "[W]e should not underestimate the power of the group: it may bring significant pressure to bear and it can sanction – directly or indirectly – those who fail to conform to what is considered acceptable" (p. 161). In brief, individual learners might be interdependent and are definitely affected by peers in the classroom. Therefore, Chang's study stressed the individual-level autonomy and the interactions between students in the classroom rather than the interdependence between teacher and students.

Chang's autonomy scale comprised a double measurement of learner autonomy, because a single measurement of learner autonomy is considered to be unreliable. This design is considered preferable because, as Little (1991) also claimed, autonomy "can take numerous different forms" and "can manifest itself in very different ways" (p. 4). The questionnaire was composed of 29 questions that were categorized into three sections: (1) autonomy level; (2) group cohesiveness; (3) group norms. The results showed that group processes such as group cohesiveness had a mild effect on individual learners' autonomous behaviours.

In summary, Cotterall's (1995) scale seemed to be short of metacognitive aspects and Chan et al.'s measurement (2002) tended to lean towards the interdependence between teacher and students. On the other hand, Chang's (2007) questionnaire directed attention to individual learner-level autonomy and the relationship between peers in the classroom. As the current study also focused on individual learner-level autonomy in the classroom and student assessment, that is, self- and peer assessment was dealt with, it was considered that Chang's questionnaire (2007) could most suitably ask about learner autonomy for L2 students.

2.6 Study Hypotheses

This chapter presented a literature review on student assessment in relation to teacher assessment in terms of reliability, writing ability and learner affect. The key findings can be summarized in terms of the following six points. First, the meta-analytic review of previous studies on student assessment presented in Section 2.2.1 identified notable trends of previous studies on student assessment. It was found that the number of comparative studies comparing between self-assessment and peer assessment is smaller than that of individual studies on self-assessment and peer assessment. In addition, adolescents (or high school students) have received little attention as research subjects so far. Furthermore, many studies have focused on examining the relationship between teacher assessment and student assessment, particularly in the forms of correlation and agreement of scoring results. Second, while previous studies addressed various issues related to the reliability and validity of student assessment, it is also still worth examining the reliability and validity of student assessment compared to teacher assessment. This is because many studies have examined the reliability in terms of consistency (Ross, 2006; Weaver et al., 2011), but not in terms of severity (Table 2.4), so it is meaningful to analyse the reliability of student assessment in terms of two aspects, i.e., consistency and severity. In addition, investigating the reliability of student assessment allows teachers to receive valuable information about how each student understands teaching content and evaluates themself or their peers. Third, as shown in Sections 2.3 and 2.4, the effect of student assessment on writing ability has not often been investigated (11 out of 87 studies or 12.6 %

for self-assessment; one out of 76 studies or 1.3 % for peer assessment; and six out of 41 studies or 14.6 % for comparative studies). Some previous studies stated the effect of student assessment on the development of writing ability (Black et al., 2004; Sluijsmans & Prins, 2006), but most studies did not independently discuss the effect of student assessment on writing ability. Furthermore, previous studies were conducted using a single method, that is, a quantitative or qualitative method. Therefore, it is difficult to say that student assessment is effective for improving writing ability, because previous studies did not offer multi-perspective analysis. In order to implement student assessment in class, the effect of student assessment on writing ability should be independently examined from various angles. Fourth, previous studies described the effect of student assessment on learner affect. For instance, the enhancement of self-confidence for self-assessment (Butler & Lee, 2010; Ross, 2006) and an increase in responsibility for peer assessment have been reported (Cho et al., 2006; Falchikov & Goldfinch, 2000; Saito & Fujita, 2009). However, language learning anxiety, particularly writing anxiety, has rarely been focused upon compared to speaking anxiety (Horwitz et al., 1986; Philips, 1992; Woodrow, 2006). According to Table 2.4, language anxiety was focused upon in 10 out of 87 studies (11.4 %) for self-assessment, one out of 76 studies (1.3 %) for peer assessment and 0 out of 41 studies (0 %) for comparative studies. With respect to the studies of learner autonomy, Table 2.4 shows fewer studies: one out of 87 studies (1.1 %) for self-assessment, 0 out of 76 studies (0 %) for peer assessment and one out of 41 studies (2.4 %) for comparative studies. Therefore, it is difficult to say that student assessment has a positive effect on learner affect, especially writing anxiety and learner autonomy, because of the shortage of studies. Fifth, most of the previous studies employed a single method, that is, a qualitative (Andrade & Du, 2007) or quantitative method (Esfandiari & Myford, 2013), so more comparative analyses between self- and peer assessment would be required in order to explore the specific relationship between student assessment modes and integrate the quantitative and qualitative results to find a solution to improve writing classes. Finally, it was also found that student assessment is effective in promoting students' self-efficacy of formative assessment (Orsmond et al., 1996; Topping et al., 2000), but how to implement student assessment in class has not been analysed thoroughly enough. Therefore, it is necessary to find effective utilization of student assessment in class formative assessment. As Topping (1998) claimed, teacher assessment has a different function from student assessment. This is because student assessment has unique values and efficacy that differ from those of teacher assessment. Therefore, it is worth elucidating the reliability of each student assessment mode, that is, self-assessment and peer assessment in this study.

Based on the literature review, the next section proposes the following hypotheses corresponding to the research questions stated in Section 1.4 of Chapter 1.

2.6.1 How do Self- and Peer Assessment Compare with Each Other?

First, the reliability of self-assessment and peer assessment is hypothesized to differ from each other. This is because previous studies reviewed in Section 2.3.1 have suggested that the type of student assessment (self- vs peer assessment) affects the reliability of assessment (e.g., Andrade et al., 2010; Boud & Falchikov, 1989; Dieten, 1989; Peirce et al., 1993; Ross, 1998; Runnels, 2014). In detail, self-assessment is considered to heighten self-confidence (Butler & Lee, 2010; Ross, 2006), so such enhanced self-confidence might influence the reliability of assessment. On the other hand, peer assessment is presumed to develop evaluation skills (Cho, Schunn, & Wilson, 2006), so it might affect the reliability of peer assessment. Previous studies on the reliability of self-assessment have generally presented a modest positive correlation with teacher assessment (Ross, 2006), though its reliability tended to be influenced by several conditions, such as rater training (Leach, 2012), cultural backgrounds (Heine & Hamamura, 2007) and language proficiency (Boud & Falchikov, 1989). The reliability of peer assessment has also shown a positive correlation with teacher assessment (Weaver et al., 2011), depending on the degree of learners' understanding of assessment criteria (Orsmond et al., 1996). The comparative studies comparing between self- and peer assessment reported that peer assessment tended to be more reliable than self-assessment (Matsuno, 2009). However, the number of comparative studies of self- and peer assessment is fewer than that of studies that focused only one of the two types, so it is difficult to conclude that peer assessment is more reliable than self-assessment. Moreover, most of those studies targeted young adults. For these reasons, it is meaningful to specifically highlight the reliability of both types of student assessment. Based on previous studies (Orsmond et al., 1996, 2000; Zarei et al., 2014), it can be presumed that peer assessment is closer to teacher assessment, because the reliability of self-assessment appears to be varied, influenced by individual students' attributes such as cultural backgrounds, English proficiency, ages and self-confidence (Blue, 1994; Heine & Hamamura, 2007; Matsuno, 2009).

Secondly, writing performance would also be influenced differentially by student assessment type. Previous studies about self-assessment have indicated that self-assessment could foster students to become goal oriented by directing their attention toward the study aim and assessment criteria. It would lead students to feel a sense of achievement in their learning (Ross et al., 1998; Zarei & Usefli, 2015). In addition, it might work well in enhancing self-confidence (Andrade & Valtcheva, 2009; Butler & Lee, 2010; Léger, 2009; McMillan & Hearn, 2008). According to previous studies (Brown & Hudson, 1998; Huerta-Macias, 1995; Yamashita, 1996), peer assessment is a variant of self-assessment, but it is assumed that peer assessment presents different effects on the change of mentality and English writing from self-assessment. That is because the presence of readers (audience) influences learners as writers (Kwan & Leung, 1996; Saito & Fujita, 2004). Therefore, the second hypothesis is that students who experience peer assessment can improve their writing skills since they are more aware of readers' presence and make efforts to improve than those who experience self-assessment. However, as mentioned above, those studies focused on young adults, not adolescents. Thus, in an attempt to close this research gap, the current study investigated Japanese senior high school students who had not finished the national-level regular English curriculum; in other words, they were still developing English proficiency. While it is difficult to expect the same effect on young adults due to the focus on a different learner population than those of previous studies, it is hypothesized that self- and peer assessment would have a similar effect on the improvement of writing ability, because both assessment types have a positive effect on writing performance in spite of their differences.

Thirdly, it could be hypothesized that learner affect, i.e., writing anxiety and learner autonomy, would be positively affected by student assessment. Previous studies discussed positive effects of student assessment on learner affect (McDonald & Boud, 2003; MacIntyre et al., 1997; Tsui & Ng, 2000). However, they did not discuss the relationship between writing anxiety and student assessment. In addition, previous studies on writing anxiety mostly focused on the effect of self-assessment, and very few studies on the effect of peer assessment on writing anxiety. As regards previous studies about the effect of student assessment on learner autonomy, the number of those studies is also lower, but a positive effect on learner autonomy was reported by some previous authors (Ashraf & Mahdinezhad, 2015; Lee, 2017). However, because those studies did not specifically deal with writing and adolescent students, the present study aims to explore the effect of student assessment on learner affect, that is, writing anxiety and learner autonomy. It is presumed that self-assessment would have a positive effect on writing anxiety and learner autonomy, because self-assessment is believed to encourage students to concentrate on their work and they do not have to be criticized by teachers and peers. In contrast, peer assessors would feel stronger anxiety than self-assessment students, because they would directly receive peer pressure from peer assessment. In terms of learner autonomy, peer assessment would be a positive incentive to develop learner autonomy, because peer assessment produces interaction between peers and peer presence would stimulate students to learn autonomously.

2.6.2 How can Student Assessment Work as Formative Assessment in the Classroom?

It is considered that self- and peer assessment are two of the essential components of formative assessment (Black et al., 2004), because formative assessment helps students to learn through an ongoing assessment process (Scriven, 1967). Specifically, student assessment could provide teachers with information on how individual students evaluate their or their peers' achievement or understand a study purpose through assessment criteria. Such data could encourage teachers to amend everyday instruction, adjusting to students' attainment and obstacles in the study process. In short, formative assessment expects teachers to know students' strengths and weaknesses and adjust their instruction accordingly. However, as previous studies stated, self- and peer assessment have different aspects in terms of reliability, writing ability and learner affect, so it is hypothesized that each assessment type would play a different role in formative assessment. Yet both assessment types also have common benefits and limitations. For instance, both assessment types might enhance learning responsibility and awareness of assessment criteria (Saito & Fujita, 2004). However, it has not been analysed how those different and similar qualities of each assessment type are mutually combined, interdependent or influenced between them. Therefore, the current study pursues the relationship between each assessment type and formative assessment in a qualitative method, because a qualitative method makes it possible to elicit the hidden properties of student assessment based on the participants' personal ideas and experiences. The current study also adopted an exploratory approach to exploring a pedagogical way to establish formative assessment in the classroom. For this reason, no specific hypothesis is stated.

2.6.3 Research Questions and Study Hypotheses

Below is the list of specific hypotheses postulated for Research Question 1 in the present study. With respect to Research Question 2, the hypothesis is not proposed, because the current study was an exploratory study.

Research Question 1: How do self- and peer assessment compare with each other?

Hypothesis 1.1 Peer assessment is more reliable than self-assessment.

Hypothesis 1.2 Self- and peer assessment would have both positive effects on the improvement of writing ability.

Hypothesis 1.3 Self- and peer assessment would positively affect the development of learner autonomy, but writing anxiety would only be affected by self-assessment.

Research Question 2: How can student assessment work as formative assessment in the classroom?

Chapter 3

METHODOLOGY

3.1 Introduction

This chapter describes the methodologies used to explore the research questions. As noted in Chapter 2, a common interest in many previous studies reviewed in this study was to survey the reliability of student assessment. Therefore, the current study employed the many-facet Rasch measurement (Linacre, 1989) to examine the reliability of self- and peer assessment results against teacher assessment (Hypothesis 1.1). This approach was taken because teacher assessment is often considered to be the standard form of assessment in the classroom (Harlen, 1996). Hypothesis 1.2, regarding the effects of assessment type on writing ability development, was tested by means of a many-facet Rasch measurement analysis (Linacre, 1989). For the purpose of testing Hypothesis 1.3, concerning the effect of student assessment on writing anxiety and learner autonomy, a multivariate analysis of covariance (MANCOVA) was conducted. The second main research aim of this study is to analyse how each assessment type can contribute, in a different or similar way, to formative assessment. To this end, qualitative analyses were employed focusing on the responses of students' open-ended questionnaire and student and teacher interviews. Grounded theory and term extraction were conducted to analyse the qualitative data and the results were triangulated to explore the relationship between student assessment and formative assessment. Finally, the aforementioned results were integrated and discussed with a focus on the efficacy of student assessment.

3.2 Participants

A total of 293 students aged 15 to 18 who were enrolled in a public high school in the Kanto area of Japan participated in the present study. Two Japanese English teachers and two native-speaking English teachers also participated in the study as observers and assessors. Academically, the school was ranked at the middle level in the area. The current study adopted an *in situ* study design. *In situ* means “on site” or “in position”, so it is defined as a study that is focused on a specific normal location or is limited to its site (King & Minium, 2003). In other words, students were neither randomly sampled nor randomly assigned to experimental groups. Instead, as described below, existing classes, or homeroom classes, were employed to compare the two types of student assessment with each other. The homeroom classes were balanced in terms of gender and were specifically organized according to entrance examination scores or academic performance in English subjects of the school curriculum. In other words, each class comprised students who had mixed English language ability levels. There were no native English-speaking students. All 293 students took the Benesse Trial Examination, which comprised Listening, Pronunciation and Accent, Language Use, Reading, Colloquial Expressions, and Composition sections. The students took this test before participating in this study because it was regularly administered to students by the school as part of its curriculum at the beginning of the school year. The Benesse Trial Examination was taken almost simultaneously by about 54,000 senior high school students in Japan in 2018 (Benesse, 2018). The maximum score was 100 points with a mean score of 47.5 points and an SD of 18.6. The score distribution of the participants in the current study showed a similar frequency distribution to that of the national test-taking population. The number of participants for each gender was the same, with 146 boys and 147 girls in total. The class size ranged from 36 to 39 students. Among the eight classes, four were designated as the self-assessment group and the other four as the peer assessment group for logistical reasons such as teachers' schedules (Table 3.1).

Table 3.1

Group Organization

	Self-assessment group students				Peer assessment group students			
	Class 1	Class 2	Class 3	Class 4	Class 5	Class 6	Class 7	Class 8
Year	1	1	2	3	1	1	2	3
Subject	I	I	II	III	I	I	II	III
n	36	36	36	39	36	36	36	38
Total	147				146			

Note. I = English Communication I; II = English Communication II; III = English Communication III

As regards teachers, two native-speaking English teachers and two Japanese English teachers also participated in the study. They were experienced English teachers who had taught English at Japanese senior high schools for between seven and 30 years (Table 3.2). Japanese English teacher A and native English teacher D were males, while Japanese English teacher B and native English teacher C were females. JET B and NET C taught English to Class 1, 2 and 5, but the teachers served as observers for the purpose of this study.

Table 3.2

Years of Teaching Experience

	JET A	JET B	NET C	NET D
Age	40	55	38	50
Gender	Male	Female	Female	Male
Years of teaching experience	15	30	7	20

Note. JET = Japanese English teacher; NET = native English teacher.

3.3 The English Curriculum

At this high school, the participants attended English classes that followed the national curriculum and adopted subjects called English Communication I, II, and III. All these subjects were established as compulsory subjects according to the designated course of study for senior high schools (MEXT, 2010). The current study was conducted in English Communication I classes of four 36-person groups, two 36-person groups of English Communication II, and one 38-person and one-39-person group of English Communication III (Table 3.1).

According to the course of study, English Communication I, II, and III all aimed to "develop students' abilities such as accurately understanding and appropriately conveying information, ideas, etc., while fostering a positive attitude toward communication through the English language" (MEXT, 2010). Teachers are expected to teach all four skills in class – that is, reading, listening, speaking, and writing. However, due to the difficulty of allotting class time to all four skills equally, students were mainly taught reading and grammar in English Communication I, II, and III. English composition was taught in regular classes once or twice a month. According to the survey by MEXT (2018b; 2018c) on the English ability of Japanese third-year senior high school students mentioned in Chapter 1, productive skills (speaking and writing) are thought to receive less attention in English language instruction than the other skills (MEXT, 2018b; 2018c). Accordingly, the study sample is also considered to be similar in this respect to the Japanese third-year high school population studied in the MEXT survey.

3.4 Materials

3.4.1 Instrument Development

The current study adopted two types of writing assessment rubric: a scoring rubric and a narrative frame (Table 3.3). The scoring rubric was used for self-assessment and was also adopted in the first and final sessions of peer assessment, while the narrative frame was applied for the second, third, and fourth sessions of peer assessment. The scoring rubric was also used by the four teachers in order to compare teacher assessment and self- and peer assessment. The details are described in Sections 3.3.2 and 3.3.3.

Table 3.3

Assignment of the Assessment Scoring Rubric and the Narrative Frame to the Student Assessment Groups

Session	Self-assessment group	Peer-assessment group
1 Pretest	Scoring rubrics for assessment	Scoring rubrics for assessment
2nd session	Scoring rubrics for assessment	Narrative frame
3rd session	Scoring rubrics for assessment	Narrative frame
4th session	Scoring rubrics for assessment	Narrative frame
5 Post-test	Scoring rubrics for assessment	Scoring rubrics for assessment

3.4.2 The Writing Assessment Rubric

The writing assessment rubric was developed before the study based on the results of a needs analysis for Japanese senior high school teachers (Oi, 2019a). In the needs analysis, a total of 61 senior high school teachers in the Kanto area who had previously taught English

Expression I responded to a survey and recommended what should be assessed in students' English compositions based on writing tasks covered in English Expression I. Oi (2019a) found that most of the Japanese high school English teachers who participated in the needs analysis survey perceived task fulfilment, coherence, and organization as the most important writing assessment criteria as opposed to other components such as language use. This is because those teachers stressed the importance of writing readability. It was felt by the teachers that coherence could help readers to understand written production more easily. However, the teachers also mentioned the importance of language use. A potential explanation for this is that the teachers mostly spent time teaching grammar and vocabulary usage in English classes (Benesse, 2018; Oi, 2019a).

The rubric employed in this study comprised four assessment criteria: (1) Task Fulfilment; (2) Structure and Coherence; (3) Appropriate Usage of Vocabulary; and (4) Grammatical Accuracy. Each criterion was on a four-point scale, with a total of 16 points. The assessment criteria were described in both English and Japanese (Appendix B). These criteria based on the results of Oi (2019) were similar to the writing assessment rubric of EIKEN Grade 3 (EIKEN, 2018). EIKEN Grade 3 is aimed at the English achievement level of junior high school graduation (EIKEN, 2018).

As noted by Wall, Clapham, and Alderson (1994), the descriptors of a scoring rubric should be adjusted according to the audience. However, the same assessment rubric was used by both the self-assessment group and teachers in this study. This is because the assessment rubric of the current study was shared between teachers and students in the classroom. This enabled the teachers to give students feedback using the same rubric so as to ensure that students understood the assessment criteria. It was also efficient to use the same rubric because the current study aimed to examine the reliability of the two types of student assessment against teacher assessment.

3.4.3 Narrative Frame

Instead of the rubric assessment, this study employed a narrative frame (Barkhuizen & Wette, 2008) for the assessment of peers' compositions (Appendix D). This is because peer grading could negatively influence students' motivation towards peer assessment according to some results of previous studies (Cheng & Warren, 2005; Hanrahan & Isaacs, 2001; Kwan & Leung, 1996; Rushton, Ramsey, & Rada, 1993; Sluijsmans, Moerkerke, Dochy, & van Merriënboer, 2001). Peer assessment seems to be overwhelmingly supported in previous studies in terms of its usefulness, reliability, and validity (Cheng & Warren, 2005; Cho, Schunn, & Wilson, 2006; Sluijsmans & Prins, 2006; Zarei & Mahdavi, 2014). While Cheng and Warren (2005) reported the benefits of peer positive feedback, other studies reported that some students might express an aversion to grading their peers, especially in small and thoroughly established groups (Cheng & Warren, 1997; Kwan & Leung, 1996; Nanine, Gennip, Segers, & Tillema, 2010; Saito & Fujita, 2004; Stefani, 1992; Topping, Smith, Swanson, & Elliot, 2000). Other points reported were that students feel more comfortable just giving feedback rather than marking peers (Sluijsmans, et al., 2001), and that teachers are afraid that rating peers may potentially pose "interpersonal problems" (Brown, 1998, p. 55) or instill in students a "fear of hurting other people's feelings" (O'Malley & Valdez Pierce, 1996, p. 156; Saito & Fujita, 2004, p. 18). Therefore, as Liu and Carless (2006) stated, it is considered desirable for students to be involved in peer assessment processes where they provide feedback to one another rather than just marking peers' work and comparing the scores with those of the teacher. Those students' discomfort might be caused by criticizing each other's work or finding it difficult to grade their peers due to their fear of damaging peer relationships (Topping et al., 2000). From a different perspective, however, if the students in the peer assessment group were asked to use a more open-ended rubric, some students might be confused or lost regarding what aspects of their peers they should assess without referring to specific assessment criteria.

Given the difficulties of using a scoring rubric for peer assessment suggested above, a narrative frame was designed for peer assessment based on the same writing assessment rubric used by the teachers and the self-assessment group in this study. A narrative frame is compared to a "skeleton for scaffolded writing" (Warwick & Maloch, 2003, p. 59). It is composed of a template of "starters, connectives, and sentence modifiers, which gives children a structure within which they can concentrate on communicating what they want to say whilst scaffolding them in the use of a particular generic form" (Wray & Lewis, 1997, p. 122). Barkhuizen and Wette (2008) also agreed that those frames could provide students with a supportive and guiding function in assessment and feedback. Narrative frames are sets of templates that "provide guidance and support in terms of both the structure and content of what is to be written" (Barkhuizen & Wette, 2008, p. 376). To be specific, starters, connectives, and sentence modifiers give writers scaffolds to write their experiences and ideas in story form (Barkhuizen & Wette, 2008).

Previous studies suggest that a narrative frame helps participants to broaden their views. At the same time, it could enable participants to disclose information more freely in frames and escape from the face-threatening context with peers. For instance, Barnard and Viet (2010) compared narrative frames to semi-structured interviews, but semi-structured interviews would restrict participants in terms of the depth of their responses. On the other hand, a narrative frame allows participants to reflect more fully on questions than an interview. For instance, Barnard and Viet (2010) used a narrative frame to elicit Vietnamese teachers' beliefs and knowledge about task-based language teaching (TBLT). In this case, the narrative frame encouraged the participants to reflect thoughtfully on TBLT, and it also allowed data to be gathered on teachers' beliefs.

The narrative frame developed for the purpose of this study followed the description of the writing assessment scoring rubric that was used by self-assessment group students and four teachers (Appendix D). It comprised 10 sentences that led writers to freely write assessment including greetings and appreciation. Five of those sentences aimed to guide students to assess their peers' writing in terms of general impression, task fulfilment, structure and coherence, vocabulary, and grammatical accuracy. The students in this study were expected to write narratives in either Japanese or English. The reasons for including both Japanese and English frames are twofold. Firstly, it was deemed that the students would feel neither peer pressure nor tension in writing in Japanese when they had to conduct an unfamiliar task (Barnard &Viet, 2010). Secondly, the option of writing narrative frames in English was also provided because doing so based on their actual experience would encourage students to "use English to express their personal meaning attached to it" (Hiratsuka, 2014, p. 3).

3.4.4 Writing Tasks

Table 3.4 presents the lists of writing prompts employed in this study. As shown in the table, they appeared most frequently in English Expression I (Oi, 2018) and the previous questions from the EIKEN Grade 3 test.

Table 3.4

Lists of Writing Prompts

	Self-assessment group	Peer assessment group
Before the experiment (Practice for assessment)	1) What is your favourite season? 2) Which country do you want to visit?	
1st session	What do you do in your free time?	
2nd session	Where would you like to go in this summer holiday?	
3rd session	Which do you like better, staying at home or spending time out of the house?	
4th session	What is your favourite food?	
5th session	What is your future dream?	

The ones employed in assessment training were adopted from EIKEN Grade 3, while the others were selected from lists of the most frequently occurring writing prompts in English Expression I (Oi, 2018). Specifically, writing tasks were presented to the students as follows:

Example task 1. Write an English composition about your favourite season, titled "My Favorite Season". You should present reasons and examples in your writing in approximately 50 words in 10 minutes without using dictionaries.

Example task 2. Write an English composition about your free time. You should clearly explain in detail experiences or examples in your writing in approximately 50 words in 10 minutes without using dictionaries.

EIKEN Grade 3 was used for assessment training in the current study because it was not considered difficult for senior high school students. As mentioned above, students did not specifically have writing training in class, so relatively simpler tasks were thought to be adequate for students, as it was thought to be necessary to have equal conditions for every student.

EIKEN Grade 3 is aimed at Japanese junior high school graduates. Examinees are expected to be able to understand and use language concerning familiar, everyday topics, such as likes and dislikes, and basic personal and family information. The level of EIKEN Grade 3 corresponds to A1 on the CEFR (EIKEN, 2018). The summary of "can-do" statements for EIKEN Grade 3 states: "Can write simple texts about himself/herself". Grade 3 questions are very similar to writing tasks and questions from the writing and speaking textbook for English Expression I. This course mainly aims to enhance students' ability to evaluate facts and opinions from multiple perspectives, as well as communicate through reasoning and a range of expressions, while fostering a positive attitude toward communication through the English language (MEXT, 2010). Students are asked to write about approachable and personal matters such as their interests, family, and friends (Appendix C).

3.4.5 Questionnaire on Writing Anxiety and Learner Autonomy

The questionnaire was composed of two parts: the questions about writing anxiety and learner autonomy. It comprised 32 questions in total, with 22 questions about writing anxiety adopted from the Second Language Writing Anxiety Inventory (SLWAI) (Cheng, 2004) and 10 questions about learner autonomy adopted from Chang's study (2007).

The 22 items adopted from the SLWAI (Cheng, 2004) were aimed at assessing the level of anxiety that the students felt regarding second language writing. The SLWAI has been devised as a writing skill-specific scale. Chang (2004) reported that a high correlation has been found between the SLWAI and writing achievement, supporting the fact that the SLWAI specialized in writing anxiety in second language learning. Its validity has been verified by several studies. The original SLWAI items were based on a five-point Likert scale ranging from "strongly agree" to "strongly disagree", but the current study adopted a four-point Likert scale for 22 items. This is because most of the questionnaires distributed in school used a four-point scale (e.g., student questionnaires on students about matters relating to their school and personal lives). Therefore, it was decided that senior high school students were more familiar with the four-point rating scale. On the SLWAI, a higher rating obtained indicates a higher level of writing anxiety.

Appendix A presents 32 items included in the final version of the questionnaire employed in this study that comprises three sections: (1) the SLWAI questionnaire; (2) a learner autonomy questionnaire; and (3) an open-ended questionnaire. The SLWAI was composed of three subscales, namely somatic, cognitive, and behavioural anxiety (avoidance behaviour), based on the implementation of written production. Somatic anxiety (Items 2, 6, 8, 11, 13, 15, and 19) refers to one's perception of the physiological effects of the anxiety experience, as reflected in an increase in the arousal of unpleasant feelings, such as nervousness and tension (Cheng, 2004, p. 316). Avoidance behaviour (Items 4, 5, 10, 12, 16, 18, and 22) refers to the mental aspects of anxiety experience, including negative expectations, preoccupation with performance, and concern about others' perception (Cheng, 2004, p. 316). Cognitive anxiety (Items 1, 3, 7, 9, 14, 17, 20, and 21) means the cognitive aspects of the anxiety experience, including negative expectations, preoccupation with performance, and concern about others' perception (Cheng, 2004). Somatic and avoidance anxiety each comprised seven items, while cognitive anxiety was measured through eight items out of the 22 items in total. Among the 32 questionnaire items in total, 15 were worded negatively and 17 were worded positively. The ratings provided by study participants were recoded when the data sets were prepared, so that higher ratings corresponded to the higher writing anxiety level or learner autonomy level and lower scores corresponded to the lower writing anxiety level or learner autonomy level.

In regard to the questionnaire items for learner autonomy (the last 10 items in Appendix A), 10 autonomy-level questions were adopted from the group/autonomy questionnaire developed by Chang (2007). This aimed to analyse the extent to which the students' learner autonomy changed after their assessment in terms of the level of responsibility toward particular actions, the level of self-awareness about one's strengths and weakness, and the level of their personal goal setting. A four-point rating scale was also adopted for this learner autonomy scale for the same reason described above.

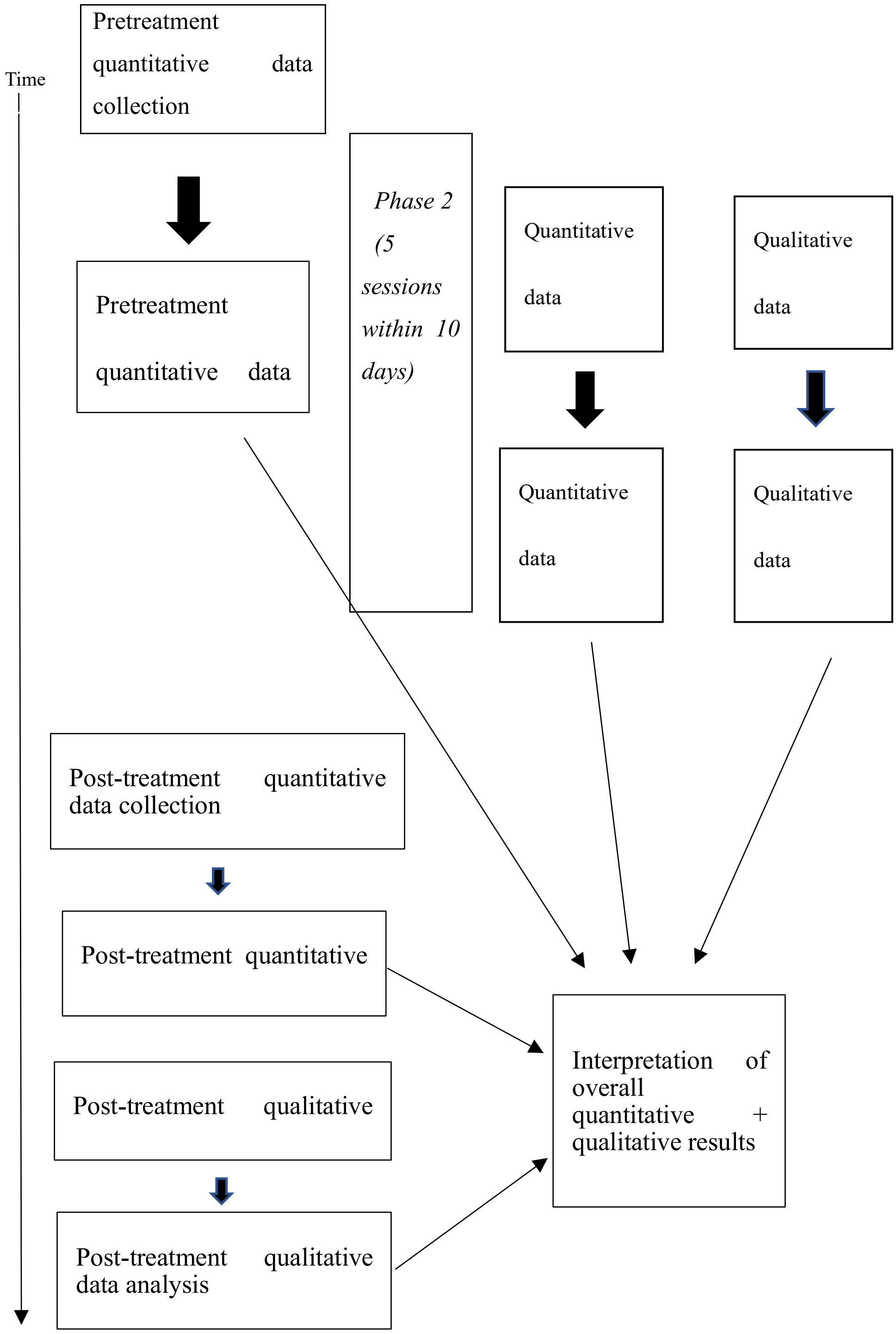
The researcher translated the 22 items from the SLWAI (Cheng, 2004) and the 10 items from learner autonomy (Chang, 2007) into Japanese. After that, one bilingual teacher of English and Japanese, who had been an English teacher in a Japanese senior high school for seven years, provided a back translation from Japanese to English in order to verify the accuracy of the translation. Then, the questionnaire was pilot-tested with 40 senior high school students. The students commented on the Japanese wording to suggest ways of making the translated sentences easier to understand, and then it was revised based on their comments. As a result of this, a few Japanese descriptions were changed into more easily understandable expressions.

Open-ended questions presented at the end of the writing anxiety and learner autonomy questionnaire (Appendix A) were presented independently of other parts to students after the current study in order to ask their opinions about effective points and challenges of student assessment. The students were allowed to answer the questions either in Japanese or English. They were asked to reply to open-ended questions after answering a total of 32 questions about writing anxiety and learner autonomy.

3.5 Procedure

3.5.1 Research Design

As noted in Chapter 1, the present study adopted a mixed-methods study approach. In particular, the convergent mixed-method research design, which aims “to merge the results of the quantitative and qualitative data analysis” (Creswell, 2015, p. 36), was adopted to collect data and interpret the results for the research. Quantitative and qualitative measures were concurrently employed, but the former were somewhat more emphasized than the latter. The mixed-method design adopted in this study is illustrated in Figure 3.1. The long black arrow to the left indicates the study time frame. The boxes present the sequencing of the three phases of the study. The arrows pointing to the box in the bottom-right part of the figure show the integration of the data and interpretation of the results across the different parts.



Time
Pretreatment quantitative data collection
Pretreatment quantitative data
Phase 2 (5 sessions within 10 days)
Quantitative data
Qualitative data
Quantitative data
Qualitative data
Post-treatment quantitative data collection
Post-treatment quantitative
Post-treatment qualitative
Post-treatment qualitative data analysis
Interpretation of overall quantitative + qualitative results

Figure 3. 1

The Convergent mixed-method research design

In the convergent study design depicted in Figure 3.1, quantitative analyses were conducted in order to analyse self- and peer assessment results in terms of the reliability of scores against teacher assessment, the effects on writing ability, and the relationships with writing anxiety and learner autonomy before and after the experiment. The qualitative analysis mainly focused on examining the effect of student assessment on formative assessment. The quantitative and qualitative results were integrated for exploring new perspectives of student assessment in order to improve writing instruction in the Japanese high school setting.

The study was carried out over 10 consecutive days in the spring of 2019 to prevent students from being influenced by other factors such as English study out of school. Table 3.5 summarizes key activities associated with the three phases of the study. In Phase 1, data were collected about the writing anxiety and learner autonomy of study participants. In Phase 2, students were asked to write English compositions in five sessions (see Table 3.4) over the 10 consecutive days. In each session, the self-assessment group evaluated their own writing by themselves, while the peer assessment group assessed their classmates' writing. In each of the five sessions, the four teachers administered the self- and peer assessments and the survey and observed the students. They also assessed the students' English compositions, using the same assessment rubric used by the students. In Phase 3, students were asked to complete the same questionnaire as the one given to them before the experiment as the post-test. Finally, 12 students and two teachers were interviewed about the effect of student assessment on learner writing performance and affect.

Table 3.5

The Study Design

Phase	Aim	Self-assessment group	Peer assessment group
Phase 1 (2 days within a week) 4 classes x 2 sessions = 8 classes	Pretest & assessment training	1. Writing anxiety & learner autonomy questionnaire 2. Pretest to assess writing ability 3. Training twice in how to assess English compositions using a rating rubric	1. Respond to questionnaires about writing anxiety & learner autonomy 2. Pretest to assess writing ability 3. Training twice in how to assess English compositions using a rating rubric & narrative frame
Phase 2 (5 sessions within 10 days) 4 classes x 5 sessions = 20 classes	Consecutive writing assessment	1. Write five 40- to 50-word English compositions without using dictionaries for 10 minutes five consecutive five times (i.e., one composition per day) 2. Self-assessment of English compositions after writing session for 10 minutes, using self-assessment rating sheet 3. Submit self-assessment sheet	1. Write five 40- to 50-word English compositions without using dictionaries for 10 minutes five times (i.e., one composition per day) 2. Peer assessment of classmates' English compositions after writing session for 10 minutes, using a narrative frame sheet and show it to partners 3. Submit peer assessment sheet
Phase 3 (2 days within a week) 4 classes x 2 sessions = 8 classes	Post-test	1. Post-test to assess writing ability 2. Writing anxiety & learner autonomy questionnaire 3. Interview with 12 students & 2 teachers for 15–20 minutes	1. Post-test to assess writing ability 2. Respond to the same questionnaire about writing anxiety & learner autonomy as that of the pre-research 3. Interview with 12 students & 2 teachers for 15–20 minutes

3.5.2 Phase 1

3.5.2.1 Investigation of Learner Writing Ability, Writing Anxiety, and Learner Autonomy before the Study. Before the study, the writing ability of the students was examined as the pretest by using the composition questions as part of the EIKEN Grade 3 test. Students were asked to write one English composition on the topic of how to spend one's free time (Table 3.4), for 10 minutes, consisting of 30 to 40 words, without using dictionaries. The English compositions were evaluated by the same four teachers who participated in the current study. The scores of the English composition were reported to all students after the completion of Phase 3.

3.5.2.2 Assessment Training. As described in detail below, students practised assessing writings in two training sessions (30 min. each), where they learned how to assess English compositions, one week before the inception of Phase 2. Both groups (self-assessment and peer assessment) were shown sample compositions and learned how to assess English compositions. The model composition was selected from a collection of writing samples from past EIKEN Grade 3 writing questions, as model EIKEN compositions had already been published in Japan (EIKEN, 2018).

The self-assessment and peer assessment groups received different assessment training (Table 3.3). While the self-assessment group used the scoring rubric only, the peer assessment group used both the scoring rubric and the narrative frame. The scoring rubric was presented to the peer assessment group in Sessions 1 and 5 of Phase 2 for assessing their peers' writing. It was believed that having the peer assessment group use the scoring rubric in the first session would help the students to understand the description of the assessment criteria on which the narrative frame for peer assessment was based. In addition, the peer assessment group used the same scoring rubric for their assessment training as well as for the first and final sessions of the present study. However, the scores assigned to writings by students in the peer assessment group were not shown to their partners. The scoring rubric was only used to teach the peer assessment group the meaning of the components of assessment and survey the reliability of peer assessment before and after the study. Table 3.6 shows how each group's assessment training was conducted. This process was repeated twice in two regular English class sessions. In the training session, about 30 minutes were allocated for the teacher to explain how to evaluate, write an English composition, and assess their own work by using assessment rubric. In addition, peer assessment group students received an explanation as to how to express their ideas in narrative frames. Furthermore, informed consent was obtained from the students to participate in the study. Students were told how to respond to questionnaires and were encouraged to respond without worrying about evaluation, because it did not have an effect on their school records.

Table 3.6

The Procedure of Assessment Training

Time	Self-assessment group	Peer assessment group
30 min. x 2	1. Learned meaning of the descriptions of the four components of the writing assessment rubric (Appendix B). 2. Studied how to judge the level of each component, using the 4-point Likert scale rubrics written in both English and Japanese. 3. Wrote an English composition from the 3rd Grade EIKEN twice, and also evaluated their composition by themselves. 4. Submitted their composition and assessment rubric, including their scoring, to a teacher. 5. In the next period, the self-assessment rubrics and compositions were returned to the students with written comments from a teacher. Teachers' written comments were related to students' scoring judgement, indicating the difference between the teachers' and students' scoring. This process was repeated twice in two regular English class sessions.	1. Received two types of assessment rubric (the same assessment rubric used by the self-assessment group and teachers & a narrative frame). 2. Learned how to assess English compositions, using the assessment scoring rubric and narrative frame. 3. Received a lecture from a Japanese English teacher about how to check and judge English compositions. 4. Wrote comments following the instructions of the narrative frame. 5. The narrative frame for practice was gathered and read by teachers, but it was not given to students' partners. 6. Provided with a general evaluation of their comments based on the narrative frame. This process was repeated twice in two regular English class sessions.

Similarly, the four teachers, namely the two Japanese English teachers and the two native-speaking English teachers, also practised how to assess students' English compositions, using the same assessment rubric based on four-point scales as the one the self-assessment group used. Before the assessment, the four teachers had a meeting to discuss the scoring rubric, referring to a model writing of the EIKEN Grade 3 test. After that, each teacher evaluated a sample student writing and discussed differences in the scores they had assigned. Since the four teachers presented high consistency among their scores (inter-rater agreement: 98.5%), no further training was provided.

3.5.3 Phase 2

Students in the self-assessment and peer assessment groups were both asked, in each of the five sessions, to write a 40- to 50-word English composition without using dictionaries in 10 minutes. The four teachers who assessed students' English compositions gave the students directions, such as the time limit, the number of words expected, and a brief explanation of the task. In every session, self-assessment group students assessed their own English compositions, using the scoring rubric. On the other hand, the peer assessment group filled in the blanks of the narrative frames (see Appendix D) either in Japanese or English. The completed narrative frames for peer assessment were collected by the teachers before being returned to the writers in order to check whether the descriptions were appropriate from the educational point of view.

3.5.3.1 Student Observation. All of the teachers in the study regularly taught English and were acquainted with students in the same school, so their experience and information were useful for deepening the understanding of the students' writing performance and learning attitudes. Therefore, the current study takes the form of participant observation. According to Flick (2014), "participant observation is the most prominent way of doing observation, because the researchers enter the field and try to become part of the field and an active member of it" (p. 296). The researchers do not have to hesitate to be involved in activities in the field; in other words, they take part in them (Flick, 2014). Schensul, Schensul, and Lecompte (1999) also defined participant observation as "the process of learning through exposure to or involvement in the day-to-day or routine activities of participants in the researcher setting" (p. 91).

Participant observation has advantages and disadvantages. As one of the advantages, participation offers insights into the field and "towards finding better access to relevant processes and practices that the researchers want to observe for their study" (Flick, 2014, p. 296). However, it might be difficult to strike a balance between "participation" and "observation", so it is important to maintain the distinction between these two positions. Therefore, not only the researcher but also other teachers observed the students who participated in the intensive student assessment sessions in order to make the observation fairer and more specific. In the current study, the four teachers who were shown in Table 3.2 were asked to observe students and take notes freely during the experiment. While taking notes, observers were also asked not to show students that they were making notes, because teacher or observer note taking might influence learner affect. Specifically, students' tension and worries could be intensified. Since the current study aims to investigate the effect of student assessment on writing performance and learner affect, attempts were made to minimize those teacher effects on students' consciousness. Subsequently, these teachers underwent semi-structured interviews on what they had observed in the classroom.

3.5.4 Phase 3

In Phase 3, students were asked to complete the same writing anxiety and learner autonomy questionnaire as the one administered in Phase 1. First, students were given an explanation about the purpose of the questionnaire. Then, they completed the questionnaire, including open-ended questions that asked about effective points and challenges regarding student assessment within about 15 minutes during a regular English class.

3.5.4.1 Interviews. Interviews can range from "unstructured" to "highly structured" (Flick, 2014). This categorization indicates the extent to which the interviewer's ideas are implanted in the content of the interviews (Harrell & Bradley, 2009). Each kind of interview has benefits. The current study adopted semi-structured interviews, because they could guide questions and topics that must be covered for the present study. Furthermore, the interviewer could ask interviewees questions in a conversation style to collect detailed information, and the order of questions could be standardized by being adjusted to the interviewees.

According to Creswell (2007), participants' elaboration on the influences of student assessment allows interviewers to develop a more intimate perspective. This is also because participants' views disclose the perspectives that the researchers are studying (Bryman, 2004). In order to enhance the insight into understanding student interviewees, semi-structured interviews were conducted in Japanese by the present author with 12 students. Two teachers (JET B & NET C) were interviewed in English by the present author.

A total of 12 student interviewees were asked about: (1) the change in the students after the experiment in terms of writing anxiety, learner autonomy, and writing ability; and (2) the effectiveness and challenges of self-assessment and peer assessment. Six interviewees from each of the self-assessment and peer assessment groups were selected based on the average scores for the respective groups (self-assessment group: mean = 12.27; SD =1.64; peer assessment group: mean = 12.16; SD = 1.87) across the five compositions that they wrote during the five intensive writing sessions. As a result, two with low scores, two with scores around the average score, and two with high scores were selected (Table 3.7). All 12 students were interviewed in Japanese for 15 to 20 minutes.

Table 3.7

Interviewees: Students

	Means of all	SHS S 1	SHS S 2	SMS S 1	SMS S 2	SLSS 1	SLSS 2	PHS S 1	PHS S 2	PMS S 1	PMS S 2	PLSS 1	PLSS 2
Pre-test	12.22	14	15	12	12	10	9	15	15	12	12	9	8
Post-test	12.43	15	14	12	12	10	9	14	14	12	12	11	11
Age		15	16	18	16	17	16	18	15	16	17	16	16
Gender		M	F	M	F	M	F	M	F	M	F	M	F

Note. SHSS = high-scoring students in the self-assessment group; SMSS = middle-scoring students in the self-assessment group; SLSS = low-scoring students in the self-assessment group; PHSS = high-scoring students in the peer assessment group; PMSS = middle-scoring students in the peer assessment group; PLSS = low-scoring students in the peer assessment group; M stands for male students, while F stands for female students.

Two teacher interviewees, JET B and NET C, who fully observed the experiments in class, were interviewed by the present author (Table 3.2). They were interviewed in English for 20 minutes. These teachers were asked about: (1) students' attitudes toward self-assessment and peer assessment, especially their change from before to after the experiment; and (2) the effect of implementation of self-assessment and peer assessment on students' writing performance and affect in terms of formative assessment.

The interviews started with a list of probing questions, and subsequently received interviewees' responses to unanticipated questions (Bryman, 2004). The interview probing questions were formed based on the research questions in order to explore participants' general opinions on the role of student assessment (see Table 3.8 for an English translation of the interview questions). Student interviewees were asked about any effect on their writing performance and learner affect, while teacher interviewees were asked questions about the benefits and challenges of student assessment, and also the implementation of student assessment in writing classes.

Table 3.8

Semi-Structured Interview Questions

Interviewees	
Students	Teachers
● In your opinion, is student assessment important for developing writing ability? ● If so, in what ways? ● Did you feel a decrease in writing anxiety due to self-assessment/peer assessment? ● Did you feel a development of learner autonomy due to self-assessment/peer assessment? ● How can you utilize self-assessment/peer assessment in writing classes? *Note.* Interviews were conducted in Japanese.	● In your opinion, is student assessment important for developing writing ability? ● If so, in what ways? ● Did you observe that students decreased writing anxiety by using self-assessment/peer assessment? ● Did you observe that students developed learner autonomy by using self-assessment/peer assessment? ● How can you utilize self-assessment/peer assessment in writing classes? ● What are the effects and challenges of student assessment? *Note.* Interviews were conducted in English.

3.6 Analyses

After gathering the data, various types of quantitative and qualitative data analyses were conducted. First, quantitative analyses were conducted to test the three hypotheses associated with: Research Question 1 How do self- and peer assessment compare with each other in terms of: (1) the reliability of student assessment against teacher assessment (Hypothesis 1.1); (2) the effects of student assessment on writing ability (Hypothesis 1.2); and (3) the effects of intensive writing practice with student assessment on writing anxiety and learner autonomy (Hypothesis

1.3)? Subsequently, qualitative analyses were carried out to address Research Question 2, regarding more specifically: (1) the difference in the effect on writing performance and learner affect between self- and peer assessment; and (2) the effect of student assessment on the process of learning in terms of formative assessment. Table 3.9 summarizes the analyses conducted in the current study.

Table 3.9

Summary of the Analysis Conducted in this Study

Phase	Phase 1	Phase 2	Phase 3
Hypotheses	Hypothesis 1.3: the effects on writing anxiety & learner autonomy	Hypothesis 1.1: the reliability of scores against teacher assessment; Hypothesis 1.2: the effects on writing ability; Hypothesis 1.3: the effects on writing anxiety & learner autonomy	Hypothesis 1.1: the reliability of scores against teacher assessment; Hypothesis 1.2: the effects on writing ability; Hypothesis 1.3: the effects on writing anxiety & learner autonomy RQ 2: The relationship between student assessment & formative assessment
Purpose	1. To calibrate the reliability of student assessment before the experiment 2. To examine writing ability before the experiment 3. To measure writing anxiety and learner autonomy before the experiment	1. To analyse the reliability & improvement of writing ability after consecutive student assessment sessions 2. To analyse the effects on writing anxiety & learner autonomy	1. To calibrate the reliability of student assessment after the experiment 2. To examine writing ability after the experiment 3. To analyse features of self-assessment and peer assessment 4. To discuss the relationship between student assessment & formative assessment based on a mixed-method analysis
Analyses	● MANCOVA	● Many-facet Rasch measurement analysis ● MANCOVA	● Opinion words extraction ● Grounded theory

3.6.1 Quantitative Analysis

3.6.1.1 How do Self- and Peer assessment Compare with each other in terms of: a) the Reliability of Scores against Teacher Assessment? The reliability of student assessment scores was tested in order to examine Hypothesis 1.1. To be specific, the many-facet Rasch measurement (Linacre, 1989; Linacre & Wright, 1993; McNamara, 1996) was conducted to analyse the consistency of ratings and rater severity. The many-facet Rasch measurement is an extension of the Rasch model (Rasch, 1980) developed by Linacre (1989). This model makes it possible to place persons, test items, and examiners on an equal-interval scale in terms of their differing abilities (persons), difficulties (items), or leniencies (examiners), as many-facet analysis can "incorporate more variables, or facets, than the two [true score and error] that are included in a classical testing situation" (Eckes, 2015, p. 28).

It is inevitable that many variable factors, such as the severity of the raters, the conditions of candidates, and task differences, influence test takers' scores in performance-based tests. Moreover, the interaction between these factors might cause unfair differences, so these factors should be removed, for instance by conducting rater training. However, previous studies have shown that rater training cannot completely eliminate rater severity and leniency, though it can improve individual raters' consistency (Lumley & McNamara, 1995; Weigle, 1998). Therefore, it was assumed that the many-facet Rasch measurement, which jointly calibrates the effects of various facets on scores assigned by raters, was deemed suitable for analysing the score variability that can be explained by each facet. As another advantage, the results of joint calibration could be shown in a graphic display.

Another well-known analytic approach to examining the effects of various measurement facets on score variability is generalizability theory (G-theory), which is helpful in examining the variability inherent in scores when raters rate examinees' performance in test items or tasks.

G-theory and the many-facet Rasch measurement have common strengths. For example, G-theory allows an investigator to "decide which facets will be of relevance to the assessment context of interest like the many-facet Rasch measurement" (Lynch & McNamara, 1998, p. 159). Moreover, it is possible for us to examine various facets of measurement and examine the differences in their effects, through the estimated variance components, on the dependability of decisions or interpretations provided by the test scores.

However, there are notable differences between these two analytic approaches as well. Because of the two differences below, the current study chose the many-facet Rasch measurement. First, the generalizability coefficient (a reliability coefficient) generalizes to very similar conditions because all the variance components, such as main effects, interactions, and random error, must keep their values (Linacre, 1993). On the other hand, the many-facet Rasch measurement is a more useful device for predicting for each examinee a measure that is as free as possible from different sources of measurement error that affect the raw score (Linacre, 1993). Furthermore, the many-facet Rasch measurement enables us to acquire information about how well the performance of each person, rater, or task fits the expected values estimated from the model generated in the analysis. For instance, it allows us to identify questionable (misfitting) components within facets (Lynch & McNamara, 1998). Second, G-theory does not adjust any examinee's raw score for the effects of particular raters, tasks, or items that the examinee encountered. On the other hand, the many-facet Rasch measurement calculates a measure for each examinee by excluding the effects of score variability, adapted for the particular items and raters satisfied by that examinee; in other words, it is fairer than the raw score (Linacre, 1993). Hence, the many-facet Rasch measurement makes it possible to analyse individual score variability more fairly.

This study employed the many-facet Rasch measurement in order to analyse the difficulty of five English writing tasks (English compositions), the reliability of assessors, and the effect on writing ability in terms of its development of writing ability. This approach was particularly useful for addressing Hypothesis 1.1, which focuses on the investigation into the difference between self-assessment and peer assessment. It also focused on the individual students as assessors, so the many-facet Rasch measurement was more powerful in finding the differences among individual students and other related information.

The current study also adopted a repeated-measures analysis within a many-facet Rasch measurement (MFRM) analysis (Linacre, 1989). An analysis of repeated measures is useful for analysing the effects of consecutive writing practice and student assessment type (self- and peer assessment) on writing performance (Hypotheses 1.1 and 1.2). This is because it presents information about the shifts in item difficulty that reflects the individual students' true score from which errors are eliminated. According to Chang and Chan (1995), there are four approaches to applying Rasch analysis to repeated measures: "(1) control the occasion facet and perform separate Rasch analyses of the data obtained on different occasions; (2) control the item facet and perform a single Rasch analysis on an expanded number of 'subjects' by regarding the subjects assessed on different occasions as distinct ones; (3) control the subject facet and perform a single Rasch analysis on an expanded number of 'items' by regarding items used on different occasions as distinct ones; and (4) perform a three-facet analysis by adding occasion as the third facet" (p. 934). The current study adopted the third approach because it is effective for comparing the consistency and stability of item parameter estimations that are supposed to be invariant across different occasions and groups of subjects when examinee ability is held constant.

3.6.1.2 Many-facet Rasch Analysis Procedure. The data consisted of 2,930 valid ratings assigned by four teachers and 293 students on 586 compositions written by 293 students on two writing tasks (the pretest and post-test). However, some of the data for the self-assessment group and peer assessment group were missing because some students could not participate in all of the sessions. Owing to those missing data, a FACETS analysis was conducted on the final data set of 2,430 ratings obtained from 243 test takers, six raters, and two tasks (Table 3.10).

Table 3.9

The Number of Valid Ratings Included in the FACETS Analysis

	No. of partici-pants		No. of ratings	The sum of absentees	No. of valid ratings		No. of ratings
Self-assessment	147	147 students x 2 tests	294	25	122	122 students x 2 tests	244
Peer assessment	146	146 students x 2 tests	292	25	121	121 students x 2 tests	242
Teachers	4	4 teachers x 2 tests x 293 students	2,344	0	4	4 teachers x 2 tests x 243 students	1,944
Frequency of tests	2				2		
Total	2,930				2,430		

The analysis was conducted using FACETS 3.81.2 (Linacre, 1998). Model parameter estimates were obtained for examinee ability, rater severity, and task difficulty on a common log-linear metric or the logit scale. Using FACETS, the relative spread of these estimates within each facet was also obtained. That is, the FACETS analysis provided information about (a) differences in severity among raters and (b) varying abilities among the examinees. FACETS additionally yielded fit statistics for each student, each rater, and each task, which indicated the degree of predictability of each student's ability, each rater's severity, and each task's difficulty. According to Eckes (2015, p. 74), "the rater fit statistics compare model-based expectations with empirical data”. In detail, rater fit statistics present the degree to which ratings yielded by a given rater match the expected ratings that are estimated by the MFRM model (Bond & Fox, 2015; Smith, Rush, Fallowfield, Velikova, & Sharpe, 2008). With regard to raters, the present study employed the mean-square (MS) fit statistic as the rater fit statistics. This statistic provides information as to how well any set of verifiable data satisfies the requirements of a given model. Rasch analysis provides fit statistics as two chi-square ratios: infit and outfit mean-square statistics (Wright, 1984; Wright & Masters, 1981). The present study focused on infit mean-square statistics, because Bond and Fox (2015) stated that infit statistics is “an information-weighted indicator of misfit” (p. 67). On the other hand, outfit statistics is more susceptible to the influence of outlying scores. The present study discussed the writing ability of students in terms of the items’ difficulty, so infit statistics was considered to be more relevant than outfit statistics.

Using FACETS, separate analyses were performed (a) for the composite score, or the total score across all analytic rating scales (4–16), and (b) for each analytic rating scale. First, an analysis of the composite score was conducted to test how raters, examinees, the rubric, and writing tasks work together in order to evaluate examinees’ writing ability. The composite score (total score) provides useful and practical information for both teachers and students in a school, so it is beneficial to analyse the composite scores and find a useful implementation in class. In this analysis of the composite score, the extent to which teachers were equally severe in their evaluation was examined, and whether the rating scales worked well to separate the examinees (students) was also examined. Second, a separate FACETS analysis was conducted by modelling the individual analytic ratings. This is because this analysis would provide information about commonalities and differences between scales. Additionally, such analysis of the relationship between each analytic rating scale provides meaningful information for teachers and students about how to instruct and learn. In other words, the present study aims to explore an effective usage of student assessment in the classroom, so it is necessary to investigate the characteristics of four analytic rating scales using FACETS.

In the analysis of the analytic scores, this study adopted the rating scale model (RSM), which "adds a threshold parameter to indicate the comparative difficulty of a transition from one category of the rating scale to the next scale" (Eckes, 2015, p. 27; Wright, 1998). This approach was preferable to the partial credit model (PCM), which allows "the number of ordered response categories and/or their threshold values to vary from item to item" (Bond & Fox, 2015, p. 368; Masters, 1982), because the current study defined the rubrics as criterion-referenced scales in which the score levels were defined as representing the same level of performance.

In each, three facets were modelled: Students (i.e., examinees, *n* = 243), Raters (i.e., four teachers and self-assessment group students, and peer assessment group students), and Occasion (pretest and post-test). It should be noted that the "Occasion" facet was in fact an "Occasion by Task" facet because the pretest and post-test prompts were different from each other. However, these tasks were found to be similar in difficulty based on a previous FACETS analysis conducted on a separate sample obtained from the same student population (Oi, 2018). For this reason, this facet was called the "Occasion" facet. Initially, the composite score across the four assessment criteria was analysed (see Figure 4.1.1); following that, another FACETS analysis was conducted for the analytic scores separately (see Figure 4.1.2). In the Rasch analysis conducted in the present study, a combined analysis procedure was carried out based on a three-facet model in order to acquire the estimates of item, subjects, and occasion parameters. The Rasch analysis made it possible to calibrate simultaneously and independently the impact of different facets, i.e., students' ability, the severity and consistency of raters, and the difficulty of tasks, into one logit scale, so it could provide a more objective estimate of each facet.

3.6.1.3 Unidimensionality and Global Model Fit. As regards assumption checks for conducting the many-facet Rasch measurement analysis, unidimensionality and global model fit were tested. First, unidimensionality was checked by examining whether items in a test or a set of assessment criteria could work together to form a single underlying pattern of verifiable observation in the framework of a many-facet Rasch analysis (Eckes, 2015, p. 124). This is because Rasch models are used to model a single latent variable or dimension; that is, they are unidimensional models (Brentani & Golia, 2007; Eckes, 2015). In order to test the unidimensionality of the data for the four analytic rating scales, a principal component analysis (PCA) of standardized residuals (Linacre, 1998) was conducted. The residual PCA approach enables a single latent dimension to explain most of the non-random variance in the data, if the data closely fit the Rasch model (Eckes, 2015, p. 125).

After the satisfaction of the unidimensionality assumption was confirmed, the global model fit was tested. In this analysis, the differences between responses that were observed and responses that were expected on the basis of the model was examined. The differences between observed and expected responses are described as standardized residuals (Eckes, 2015). Linacre (2014) stated that satisfactory model fit is indicated when about 5% or less of (absolute) standardized residuals are ≧ 2, and about 1% or less of (absolute) standardized residuals are ≧ 3.

Figure 3.2 shows the procedure of the FACETS analysis. A total of 2,430 ratings were analysed in terms of: (1) the consistency of ratings; and (2) raters' severity and examinees' proficiency.

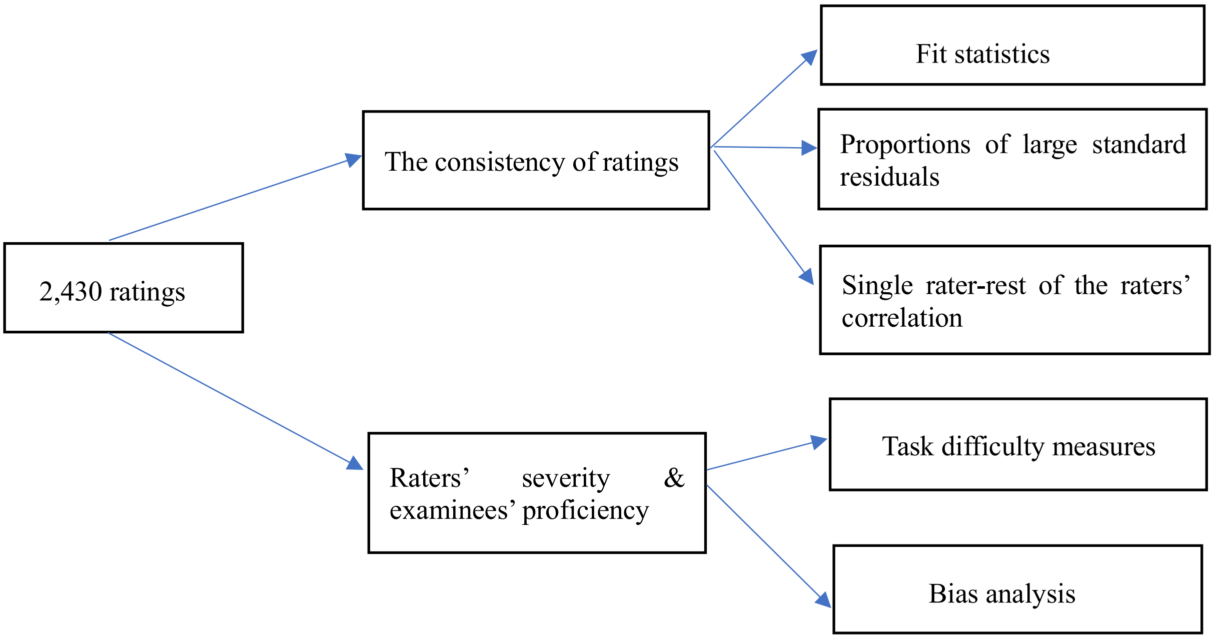

2,430 ratings
The consistency of ratings
Raters' severity & examinees' proficiency
Fit statistics
Proportions of large standard residuals
Single rater-rest of the raters' correlation
Task difficulty measures
Bias analysis

Figure 3. 2 Flow chart of the FACETS analysis procedure

As Figure 3.2 shows, the consistency of ratings was analysed using: (1) rater fit statistics; (2) proportions of large standard residuals of raters' rating consistency; and (3) single rater-rest of the raters' correlation. First, rater fit statistics were obtained through the FACETS program to examine which actual empirical responses differ from the Rasch-modelled theoretical expectations (Bond & Fox, 2015; Eckes, 2015). In other words, those rater fit statistics provided information concerning the degree to which a total of 2,430 ratings correspond to the expected ratings that are produced by the MFRM model. It identifies individual raters as potential sources of the lack of data model fit. Second, the proportions of large standard residuals were calculated to analyse the differences between observed and expected responses predicted by the models tested in order to analyse the consistency of ratings. Third, single rater-rest of the raters' correlation coefficients was obtained to quantify the extent to which the assessments of each rater (four teachers, self-assessors, and peer assessors) are consistent with the assessments of the other raters.

As regards raters' severity and examinees' proficiency, task difficulty measures and bias analysis results were examined to analyse the severity differences among the four teachers, the students in the self-assessment group, and the students in the peer assessment group (Figure 3.2). First, the analysis of task difficulty measures was carried out to examine raters' severity and examinees' proficiency; this is because an MFRM analysis enables the specified facets to be calculated on a single linear scale, that is, the logit scale. The joint calibration of facets estimates rater severity on the same scale as examinees' proficiency, task difficulty, and criterion difficulty. Therefore, task difficulty is gauged on a common scale and makes it possible to simultaneously compare tasks to each other. Second, bias analysis was conducted to measure the bias between the expected score and the observed score in order to analyse the difficulty of the composite score as well as the four analytic rating scales in the pretest and the post-test.

3.6.1.4 How do Self- and Peer Assessment Compare with each other in terms of: b) the Effects on the Improvement of Writing Ability? As for the analysis of Hypothesis 1.2, the effects of student assessment on the improvement of writing ability were examined. It is hypothesized that self- and peer assessment would influence the improvement of writing ability differently, because the severity of rating is assumed to be connected with writers' objectivity and metacognition. It is also assumed that the development of objectivity and metacognition have a positive effect on the improvement of writing ability (Black, Harrison, Lee, Marshall, and William, 2004; Harlen & Winter, 2004). However, the discussion on how differently self-assessment and peer assessment may affect the development of writing ability has been scarce in the literature. Therefore, MFRM analysis was employed in order to analyse the differences in writing ability between the self-assessment and peer assessment groups.

The results of the MFRM analysis were used to analyse the improvement in the writing ability of students, by focusing on the analytic rating scales and estimating the impact of the task difficulty and rater severity on examinees' (students') writing ability. In other words, the task difficulty, rater severity, and their impact were factored into the ability estimates for all examinees (students). In order to examine the effect of student assessment type on writing ability, the MRFM analysis was employed to analyse: (1) the composite scores across four analytic rating scales; (2) four analytic rating scales; (3) four student analytic rating scales for self- and peer assessment groups. The same data of the analysis of four analytic rating scales were partly shared in order to analyse the reliability of assessment and task difficulty. The rating scores assigned by teachers were used for the MFRM analysis, because teacher assessment is often considered to be the standard assessment in the classroom.

Additionally, the Wald statistics were obtained to test the difference in proficiency estimates of different examinee pairs. Following the context of examining data model fit (Fischer & Scheiblechner, 1970), the following formula, namely the Wald statistics, was obtained in order to statistically test the difference in proficiency estimates between two examinees *n* and *m*:

$$t_{n,m} = \frac{\hat{\theta}_n - \hat{\theta}_m}{(SE_n^2 + SE_m^2)^{1/2}}$$

where

$\hat{\theta}_n$ = proficiency estimate of examinee *n*

$\hat{\theta}_m$ = proficiency estimate of examinee *m*

SE_n = standard error of examinee *n*

SE_m = standard error of examinee *m.*

3.6.1.5 How do Self- and Peer Assessment Compare with each other in terms of: c) the Effects on Writing Anxiety and Learner Autonomy? Hypothesis 1.3 was examined by investigating study participants' responses to the writing anxiety and learner autonomy questionnaire, which was administered twice (in Phases 1 and 3). It examined the effects of student assessment type on the level of writing anxiety and learner autonomy.

The analysis of participants' responses to the questionnaire items was conducted in two steps. Firstly, simple descriptive statistics (means, minimum, maximum, standard deviation) were obtained to determine the level of writing anxiety and learner autonomy of both groups before and after the student assessment sessions. Secondly, multivariate analysis of covariance (MANCOVA) were performed to examine the effect of student assessment type on writing anxiety and learner autonomy as observed at the end of the intensive student assessment method. MANCOVA was conducted separately to analyse the effects of the student assessment method on writing anxiety and learner autonomy. As for the reason for the implementation of MANCOVA, it is a useful method for comparing two or three groups when there are covariates and two or more dependent and independent variables. The total score in the pretest questionnaires' scores was specified as a covariate. In addition, MANCOVA makes it possible to overview the relationship between subconstructs. The post-questionnaires' scores were used as dependent variables. The dependent variables were somatic anxiety, avoidance, and cognitive anxiety for writing anxiety, and learner autonomy. Student assessment type (i.e., self-assessment and peer assessment) was specified as the fixed factor (independent variable). A preliminary assumption check was conducted to test normality, linearity, univariate and multivariate outliers, the homogeneity of variance-covariance matrices, and multicollinearity before conducting the MANCOVA.

3.6.1.6 Qualitative Analyses. In the current study, two types of qualitative data were collected to examine Research question 2: (1) protocols of one-on-one semi-structured interviews with 12 students and two teachers; and (2) self- and peer assessment group students' responses to the open-ended questions collected in regular English classes just before and after the experiment in Phase 2. Both types of data were analysed thematically by focusing on writing ability, writing anxiety, and learner autonomy. Finally, these qualitative analyses were consolidated to address the second main research question: How can student assessment work as a formative assessment in the classroom?

3.6.1.6.1 Semi-structured Interviews. Semi-structured interviews were recorded and transcribed by the researcher, and then analysed using grounded theory (Glaser & Strauss, 1967), which aims to construct theory based on the perspectives of study participants. Grounded theory was developed in the 1960s by Barney Glaser and Anselm Strauss. This approach aims to extract categories from data, connect with categories, and develop the structure and process of phenomena. The grounded theory approach comprises four kinds of categories that vary in the level of abstraction: property, dimension, label, and category (Saiki, 2018). Property and dimension have lower abstraction levels than the other two categories, but they play a core role in this approach. Property presents the viewpoint of researchers, and dimension refers to the positioning of intercept data in terms of property. Both property and dimension are expected to have five roles: (1) the basis of extracted categories from the data; (2) presentation of hints to understand categories; (3) connection between categories; (4) presentation of patterns of change process in a phenomenon; and (5) explanation of the reason why the researcher interprets the data. Other qualitative approaches, such as narrative research, phenomenology, ethnography, and case study, have been used in previous research to explore and analyse more cultural, individual, or phenomenological data (Creswell, 2015). However, those qualitative approaches have not proved to be approachable as methodology (Nadamitsu, Asai, & Koyanagi, 2014). For instance, narrative studies have been carried out to explain abstract concepts and deepen understanding of others' ideas and opinions (Nadamitsu et al., 2014). Yet Fisher (1985) stated that there are no established methods in narrative research except for critical reading of a text. In contrast, grounded theory has presented a distinct data analysis method. Furthermore, grounded theory has the advantage of investigating the inner, psychological world of a person through an interview, because grounded theory stresses how a subject interprets an event. In addition, grounded theory enables a common theory to be found in a series of events by analysing the data generated from interviews (Glaser & Strauss, 1967). Therefore, grounded theory is considered to be the best approach adapted in the current study as the current study's focus is on the assessment of senior high school students in the classroom, i.e., the study is aimed at the development of a theory from data obtained from students in the classroom.

In the current study, five steps were taken in the analysis of the semi-structured interview data in order to increase the level of extraction: (1) open coding; (2) organization of the codes into categories or axial coding; (3) integration of the categories into theoretical codes (emergent themes); (4) development of emergent themes (major categories); and (5) development of a core category (Charmaz, 2006; Corbin & Strauss, 2008). Two Japanese English teachers were involved in the analysis: the present author and another Japanese English teacher (JET A), who also served as a coder for the grounded theory analysis. The present author developed coding guidelines with examples, and JET A and the present author analysed the data referring to the model coding. After the coding, the present author and JET A compared and discussed when they found discrepancies between their coding results.

3.6.1.6.2 Student Responses to Open-ended Questions. Students' responses to the open-ended questions collected in Phase 3 were analysed by manually counting the frequently appearing terms and phrases based on term extraction (Hu & Liu, 2004; Qiu, Liu, Bu, & Chen, 2011). Term extraction aims to identify the aspect of expressions that refer to the product or service properties (or attributes) from student responses. As Figure 3.3 shows, term extraction was generally employed in two steps.

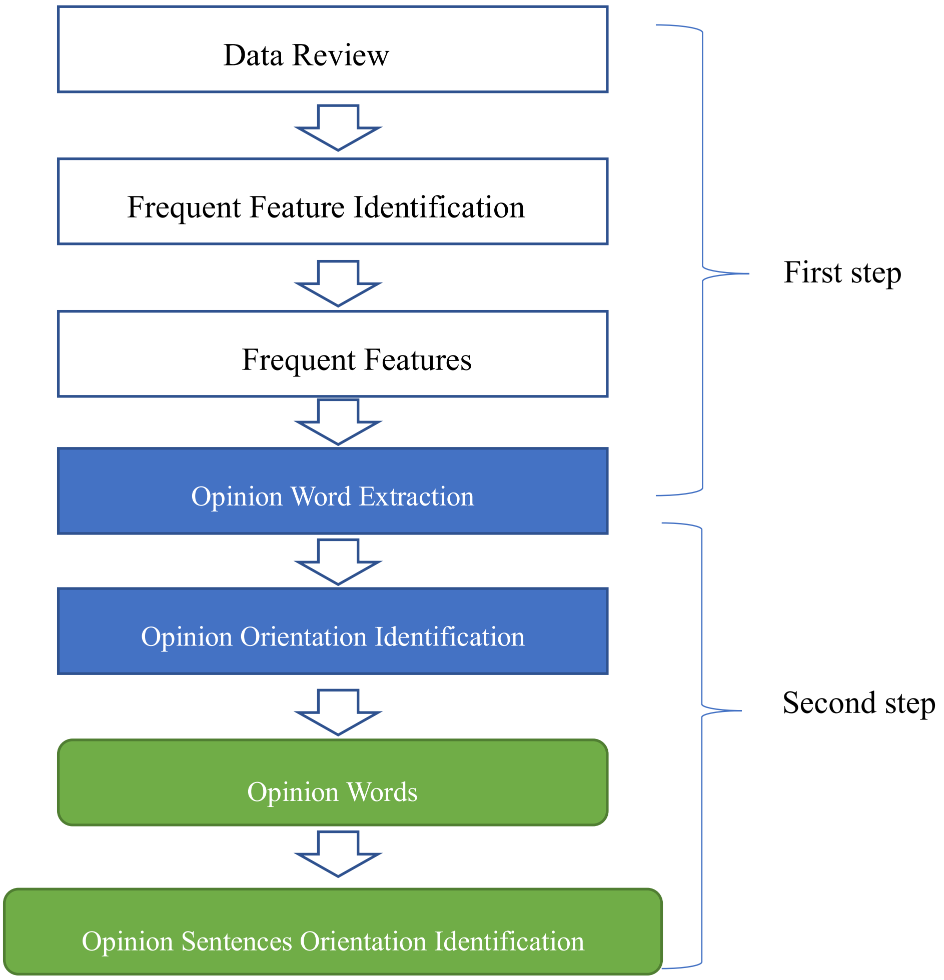

Data Review
Frequent Feature Identification
Frequent Features
Opinion Word Extraction
Opinion Orientation Identification
Opinion Words
Opinion Sentences Orientation Identification
First step
Second step

Figure 3. 3 Procedure of opinion word extraction

Firstly, the author extracted participants' opinions about student assessment from open-ended questions, employing the opinion mining or sentiment analysis method (Qiu et al., 2011), and then JET A read the extractions and checked for incoherence in the results. Secondly, opinion words were listed, for instance *good*, *excellent*, *poor*, and *bad*, which are used to indicate positive or negative sentiment in order to carry out sentiment classification and feature-based opinion summarization (Hu & Liu, 2004). The opinion words are also identified through the syntactic relations in order to analyse the meaning of selected words by examining the context surrounding them (Figure 3.3).

3.7 Synthesis of Quantitative and Qualitative Results

The goal of the current study was to explore the contribution of student assessment to formative assessment, so all of the results from the above data were inductively integrated and interpreted to uncover the role of student assessment for formative purposes. As Figure 3.4 shows, Hypothesis 1, about the difference between self- and peer assessment, was discussed from both quantitative and qualitative perspectives. Research question 2, on the relationship between student assessment and formative assessment, was considered based on the integration of qualitative and quantitative analysis.

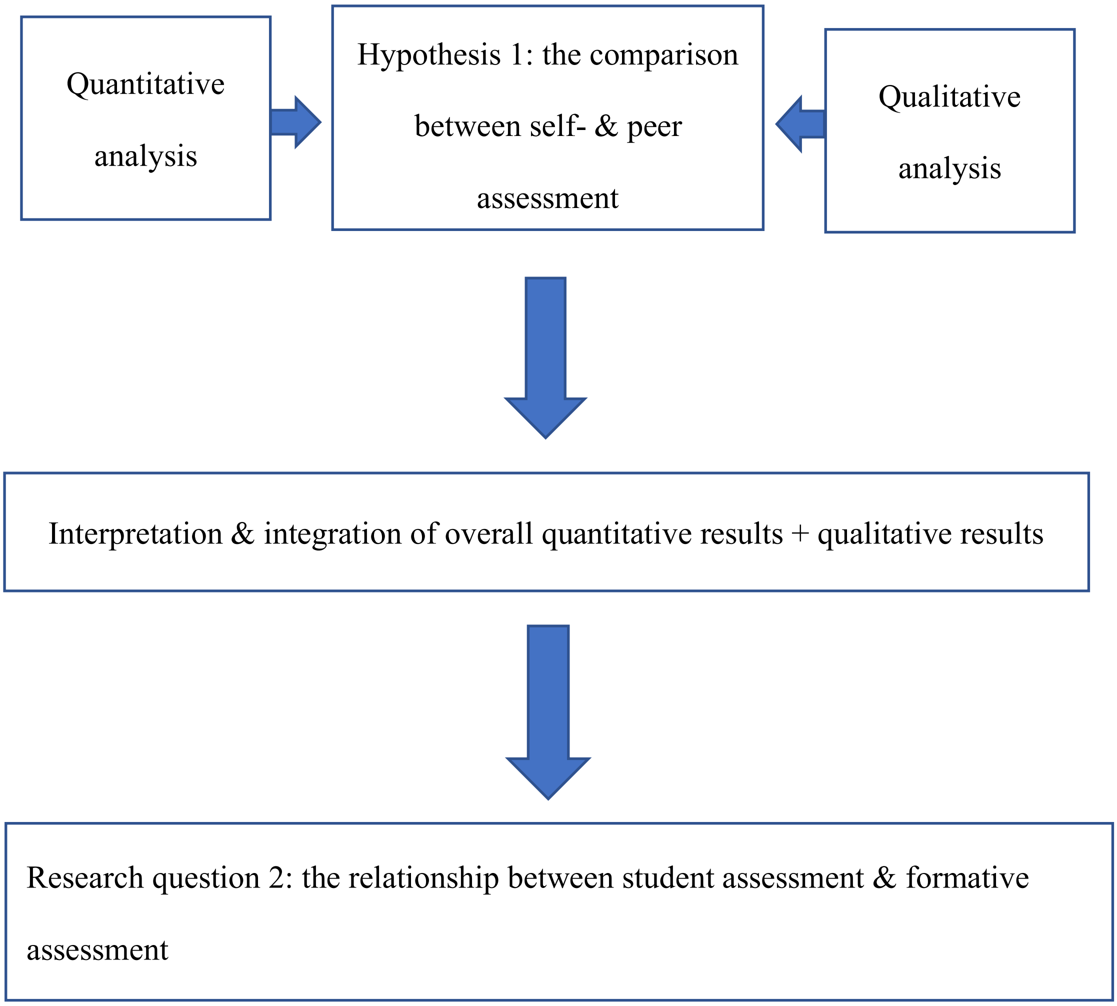

Quantitative analysis
Hypothesis 1: the comparison between self- & peer assessment
Qualitative analysis
Interpretation & integration of overall quantitative results + qualitative results
Research question 2: the relationship between student assessment & formative assessment

Figure 3. 4 Synthesis of quantitative and qualitative results

Chapter 4

QUANTITATIVE STUDY

This chapter presents the results of the quantitative analyses conducted to examine the effects of student assessment type (self-assessment and peer assessment) on the reliability, improvement of writing ability, and learner affect of senior high school students. The results presented in this chapter specifically address Research Question 1.

4.1 The Reliability of Student Assessment

This section addresses Hypothesis 1.1 on the reliability of scores against teacher assessment. In order to compare the reliability of the assessments between the two student assessment types, parameter estimates of rater differences were analysed in terms of severity and consistency as well as task difficulty by means of the MFRM analysis (Linacre, 1989) conducted using the FACETS program (Linacre, 1996). The reliability of student assessment against teacher assessment was also analysed at three levels: (1) the composite score across the four analytic rating scores; (2) the four analytic rating scales of all students; and (3) the four analytic rating scales of self- and peer assessment groups.

4.1.1 Many-facet Rasch Analysis Results on the Composite Score across the Four Analytic Rating Scales

4.1.1.1 Assumption Checks for the Many-facet Rasch Analysis. Preliminary assumption checks for the many-facet Rasch analysis showed that the unidimensionality of data and global model fit were acceptable. First, with regard to the unidimensionality, the variance explained by Rasch measures of the present study was 50.58 %. To be specific, the data derived from the PCA of the present study showed that the raw-score variance of observations was 3.12 (100.0 %), the variance explained by Rasch measures was 1.58 (50.58 %), and the variance of residuals was 1.54 (49.42 %). According to Engelhard (2013, p. 185), unidimensionality is satisfied if the variance explained by Rasch measures is ≧ 20 %. Accordingly, the unidimensionality of the present study was accepted.

Second, with respect to the global model fit, a total of 2,920 responses were analysed for estimation of (non-extreme) parameter values. Of these, 100 responses (or 3.4 %) were connected with absolute standardized residuals ≧ 2, and 42 responses (or 1.4 %) were associated with absolute standardized residuals ≧3. Satisfactory model fit is indicated when about 5 % or less of (absolute) standardized residuals are ≧ 2, and about 1 % or less of (absolute) standardized residuals are ≧ 3 (Linacre, 2014). Thus, these results would present satisfactory model fit. The results of log-likelihood chi-square also indicated that the data log-likelihood chi-square was 9077.69 with an approximate degree of freedom of 2,609 (p < .00). According to Eckes (2015, p. 69), if the results of log-likelihood chi-square are statistically significant, it cannot be said that the data fit the Rasch analysis. However, as Eckes (2015) mentioned, it can occur for any set of empirical observations (p. 69). Therefore, it is interpreted that the data of the present study fit the Rasch analysis since the other preliminary assumption checks were satisfactory.

4.1.1.2 The Wright Map for the Composite Score across the Four Analytic Rating Scales. The composite score of the four analytic rating scales was analysed, using FACETS analysis. Figure 4.1.1 presents the Wright map for the analysis of the composite score. In the figure, the vertical scale along the left presents the logit scale, which is the same for all facets. *Students* represents the examinee ability level; they are ordered with the most able examinees at the top, and the least able at the bottom. Each asterisk (*) symbolizes three students, and a dot (.) represents one or two students. The other facets are ordered so that the most difficult element of each is towards the top and the least difficult towards the bottom. In this figure, raters are represented in the *Raters* column as *Self* for self-assessment group students, *Peers* for peer assessment group students, and *Teachers 1 to 4* for the four teachers. The most severe rater is the uppermost rater in the figure, while the most lenient rater is located towards the bottom. The *Occasion* column displays the difficulty of the pretest and the post-test. The most difficult task is located toward the upper end, while the easiest task is located toward the bottom. In this case, however, the pretest and post-tests were parallel to each other. This means the task difficulty of the pretest and post-test is similar. The scale in the right column shows the most likely scale score for each ability level on the logic scale. Hence, the figure graphically presents the differences across the different facets and allows the facets to be compared against each other simultaneously.

According to the figure, examinee ability estimates ranged from a high of about 2.5 logits to a low of close to -1 logits. The results for the task facet suggest that the difficulty of the post-task and pre-task was almost the same. In contrast to the level of students, the difficulty of tasks was relatively low because there are many more asterisks indicating that students were located higher than the locations of both the pre-test and post-test tasks. As shown in the *Raters* column, the peer assessors were by far the most lenient, whereas the other raters were at similar levels of severity. Thus, the self-assessors assessed their own compositions as severely as the teachers did. On the other hand, the peer assessment group assessed peers' compositions leniently to a large degree.

```
+---------------------------------------------------------------------------------------------------+
|Measr |  Students   |  -Rating methods                        |-Occasion              |Scale|
|-----  +-----------  +--------------------------------------+---------------------- -+-----  |
|    3 +              +                                        +                       + (16) |
|      |              |                                        |                       |      |
|      |              |                                        |                       |  15  |
|      |              |                                        |                       |      |
|      | .            |                                        |                       |      |
|      |              |                                        |                       |      |
|      | .            |                                        |                       |  --- |
|      | .            |                                        |                       |      |
|      | .            |                                        |                       |      |
|    2 + *            +                                        +                       +      |
|      | *.           |                                        |                       |  14  |
|      | .            |                                        |                       |      |
|      | ****.        |                                        |                       |      |
|      | ***.         |                                        |                       |  --- |
|      | *            |                                        |                       |      |
|      | *.           |                                        |                       |      |
|      | *******      |                                        |                       |      |
|      | *.           |                                        |                       |  13  |
|      | *.           |                                        |                       |      |
|    1 + ***.         +                                        +                       +      |
|      | ***.         |                                        |                       |      |
|      | *.           |                                        |                       |      |
|      | ***.         |                                        |                       |  --- |
|      | **.          |                                        |                       |      |
|      | *            |                                        |                       |      |
|      | ***.         |                                        |                       |      |
|      | .            |                                        |                       |  12  |
|      | ***.         | Self    Teacher1  Teacher2  Teacher3  Teacher4               |      |
|      | .            |                                        |                       |  --- |
       0   * *. * * * *                                          Post-test  Pretest         |
|      | **           |                                        |                       |  11  |
|      | .            |                                        |                       |      |
|      | .            |                                        |                       |  --- |
|      | *            |                                        |                       |  10  |
|      |              |                                        |                       |      |
|      |              |                                        |                       |  --- |
|      | .            |                                        |                       |   9  |
|      |              |                                        |                       |  --- |
|      | .            |                                        |                       |   8  |
|   -1 + .            + Peers                                  +                       +  --- |
|      |              |                                        |                       |   7  |
|      |              |                                        |                       |  --- |
|      |              |                                        |                       |      |
|      |              |                                        |                       |   6  |
|      |              |                                        |                       |      |
|      |              |                                        |                       |  --- |
|      |              |                                        |                       |      |
|      |              |                                        |                       |      |
|      |              |                                        |                       |   5  |
|   -2 +              +                                        +                       +  (4) |
|-----  +-----------  +--------------------------------------+----------------------- +-----|
|Measr| * = 5         |-Rater                                  |-Task                  |Scale|
+---------------------------------------------------------------------------------------------------+
```

Figure 4.1.1 Wright map for the composite scores of four analytic rating scales

Notes. Self: self-assessment group; Peers: peer assessment group

.

4.1.1.3 Summary of Rasch Statistics. The many-facet Rasch analysis of the composite score also yielded various statistics including the separation ratio (G), separation (strata) index (H), and separation reliability (R) of the composite score (see Appendix E for the full table). Those results added more specific information on the variability of examinees, severity of raters, and difficulty of tasks to the information on the Wright map (Figure 4.1.1), as depicted in the subsequent sections.

4.1.1.3.1 Examinees (students). According to the summary of Rasch statistics (Appendix E), the examinee separation value (G) showed that the population of the examinees is separable into 2.26 levels of ability. The number of measurably different levels of examinee (student) proficiency was indicated by the examinee separation index (H), or the number of examinee strata (H). The value of this index was 3.36, suggesting that among 243 examinees (students) included in the analysis, there were about three and a half statistically distinct classes of examinee proficiency levels (Appendix E). This value almost corresponds to the structure of the scale comprising four levels. This suggests that the measurement system functioned adequately to identify at least as many reliably different levels of examinee proficiency as intended. The examinee separation reliability index was .84, showing that the examinees were reliably distinguished in terms of writing ability based on the composite score (Appendix E). To be specific, the separation reliability (R) can be interpreted differently depending on the facet considered. The separation reliability (R) provides information concerning how well examinees can be distinguished in terms of their levels of ability. In other words, the examinee (students) separation reliability indicates how varied the examinee proficiency measures were. Generally speaking, performance assessments aim to differentiate among examinees in terms of their proficiency as clearly as possible, so high examinee separation reliability is preferred.

4.1.1.3.2 Raters (Self-assessors, Peer assessors, and Four Teachers). The *G* value of the rater facet showed that the variability of the severity measures was more than 11 and a half times larger than that of precision. The obtained separation index (*H*) was 15.78, indicating that, overall, the self-assessors, peer assessors, and four teachers in the current study represented statistically different classes of rater severity.

The results of rater severity in Figure 4.1.1 showed that the peer assessment group's ratings were by far the most lenient, whereas the other assessment methods, namely the four teachers' assessment and the self-assessment group, were at similar levels of severity. The Wald statistics also supported that the severity of self-assessors was not different from that of teacher assessment. In short, the self-assessment group could assess their own compositions similarly to the teachers in terms of severity. On the other hand, the peer assessment group assessed peers' compositions leniently to a large degree. The rater separation reliability was as high as .99, attesting to a marked heterogeneity of severity measures. Furthermore, the separation (strata) index indicated 15.78 (Appendix E). This suggests that all of the raters (the four teacher assessors, self-assessors, and peer assessors) formed 15.78 heterogeneous classes (Appendix E). One explanation for this result is that such high heterogeneity was caused by the leniency of the peer assessment group as opposed to the similarity in the severity of the scores assigned by the four teachers and the self-assessment group.

With regard to the *R* statistic for raters, rater separation reliability could be differently interpreted. To be specific, if rater separation values are close to 0, it means that raters have a similar degree of severity in rating. In terms of rater variability, low rater separation reliability would be advisable because this would mean that raters could be interchangeable, suggesting the equality of rating. On the other hand, when raters do not present similar degrees of severity, rater separation reliability will be close to 1. In short, rater separation reliability indicates how different severity measures are. The *R* statistic for raters obtained for the analysis of the composite score was .99, so the raters' severity varied greatly (Appendix E).

4.1.1.3.3 Occasion. The reliability of criterion separation refers to how different the criteria are in terms of their levels of difficulty. Specifically, a low separation value (G) is desired when all criteria within a scoring rubric are planned to be similarly difficult. On the other hand, a high separation value (G) is expected when the set of criteria range widely over the underlying difficulty trait to evaluate performance features. Results on the occasion facet suggested that the difficulty of the pre-task (pretest) and post-task (post-test) is almost the same (the pretest and post-test had different task content). This is because the separation ratio (G) presented for this facet was .98 (Appendix E). This indicates that the difficulty level did not differ much between the two tasks. This was also consistent with the low task (test) separation reliability (.49) and the separation strata index (1.64). Therefore, in general, these results suggest that the two tasks were very similar in terms of difficulty.

In sum, Rasch statistics of the composite scores of pre- and post-tests indicated that: (1) the ability of examinees (students) was classified into about three and a half levels; (2) the severity of raters varied; and (3) the difficulty of the pre- and post-test was similar.

4.1.1.4 The infit Statistics for the Rater Types. In order to examine the consistency of rating of rater types, the rater fit statistics for the individual rater types were examined (Table 4.1.1). The rater fit statistics have two perspectives, i.e., infit and outfit, which indicate the degree of rating consistency. According to Linacre (2002), the rater infit statistic (information-weighted mean squares) responds to unexpected inlying ratings, while the rater outfit statistic (outlier-sensitive mean-square fit statistics) responds to outlying unexpected ratings such as outliers. In other words, the infit statistic could provide the information about the ability of examinees comparable to the item's difficulty. On the other hand, the outfit statistic tends to be sensitive to the influence of outlying scores. The current study focused on the infit statistic because it is considered that infit is weighted with more information and higher estimation precision than outfit provides (Linacre, 2002; Myford & Wolfe, 2003).

Table 4.1.1

Raters' Measurement Report about the Composite Scores

Raters	Observed Average	Measure logit	Model *SE*	Infit *M* Sq	*Z* Std	Outfit *M* Sq	*Z* Std
Self	12.42	.21	.05	2.91	9.0	2.94	9.0
Teacher 3	12.32	.20	.03	.60	-6.4	.61	-6.5
Teacher 1	12.33	.20	.03	.60	-6.4	.61	-6.4
Teacher 2	12.33	.20	.03	.60	-6.4	.61	-6.4
Teacher 4	12.33	.20	.03	.60	-6.3	.61	-6.4
Peers	13.97	-1.01	.05	2.44	9.0	2.57	9.0
Mean	12.76	.00	.04	1.29	-1.3	1.33	-1.3
SD (pop.)	.61	.45	.01	.99	7.3	1.02	7.3
SD (sample)	.66	.50	.01	1.08	8.0	1.11	8.0

As regards rater consistency, infit statistics for the self-assessment and peer assessment group presented in Table 4.1.1 were high (≧1.4). It could be said that both student assessment types indicated misfit (or underfit), suggesting that there was too much variation as well as unpredictability in the scores provided by the students for their own writings or those written by their peers. On the other hand, the infit value of all teacher assessors was .60, so the consistency of teacher assessment was reasonable for measurement (Table 4.1.1).

4.1.1.5 Inter-rater Agreement. The reliability of the composite of four analytic rating scales of student assessment was analysed in terms of severity, consistency, and agreement with teacher assessment. With regard to the agreement with teacher assessment in terms of the composite scores, teacher assessment indicated reliable consistent severity (Table 4.1.2).

Table 4.1.2

Inter-Rater Agreement of Composite Scores Between Raters

	Teachers	Self-assessment group	Peer assessment group
Teachers	98.5 % (3334/3382)	81.7 % (3814/4666)	77.8 % (3626/4662)

4.1.1.6 Summary of the Many-facet Rasch Analysis of the Composite Score. This section summarizes the MFRM analysis of the composite score in terms of reliability. First, with respect to severity, teacher assessment indicated reliable consistent severity. The severity of self-assessment can be closer to that of teacher assessment, but the analysis of the composite of four analytic rating scales showed that self-assessment raters were more lenient than teachers. However, it was not found that there was agreement between teacher assessment and self-assessment (Table 4.1.1). Therefore, it is difficult to state that self-assessment is reliable. On the other hand, peer assessment students were found to be fairly lenient; further, their evaluations did not agree with teacher evaluations (Table 4.1.1). Moreover, the reliability statistics provided the desirable reliability of teacher assessment, but neither student assessment methods showed comparable severity levels among raters (Table 4.1.1).

Second, the results of rater consistency showed that both student assessment methods indicated misfit (or underfit), so their methods were variable and unpredictable (Table 4.1.1).

In contrast to the consistency of both student assessment methods, the infit statistics of teacher assessment were acceptable and productive for measurement. In short, neither student assessment types presented similar levels of consistency among teacher raters.

According to the bias analysis of the infit statistics (the full bias analysis report is shown in Appendix F), the *t*-values of the four teachers' ratings ranged from -.03 to .04, so it is interpreted that the four teachers' ratings were fairly consistent and stable. On the other hand, the ratings of self- and peer assessment students were considered to be unpredictable. To be precise, it is estimated whether expected scores were higher than observed scores. If the *t*-value is higher than +2, and lower than -2, it is considered that there was significant bias in the scoring pattern. As Table 4.1.3 shows, the four teachers' ratings were consistent, but self- and peer assessors were not consistent. The difference between the rating of teachers and the rating of self-assessors was smaller than that of peer assessors, as shown in Table 4.1.4. Therefore, it is considered that the consistency of self-assessors was closer to the consistency of teachers than that of peer assessors.

Table 4.1.3

The Frequency of t-Values Greater than 1.5 of Composite Scores

	Pretest	Post-test
Self-assessment	1	1
Peer assessment	1	1
Teacher 1		
Teacher 2		
Teacher 3		
Teacher 4		

Table 4.1.4

t-Values between Raters of Composite Scores

	Self vs T1	Self vs T2	Self vs T3	Self vs T4	Peer vs T1	Peer vs T2	Peer vs T3	Peer vs T4	Self vs Peer
Pretest	.25	.25	.25	.25	11.18	11.18	11.18	11.18	10.41
Post-test	4.51	4.51	6.06	4.51	-11.54	-11.54	-9.99	-11.54	16.05

Note. T stands for teacher.

The results of rater consistency showed that both student assessment and peer assessment methods indicated misfit (or underfit), so these methods were variable and unpredictable (Table 4.1.2), while the infit statistics of teacher assessment were acceptable and productive for measurement. In short, neither of the methods of student assessment presented the same level of consistency as raters.

In sum, the reliability statistics provided the desirable reliability of teacher assessment, but neither of the student assessment methods showed the same severity and consistency as raters (Table 4.1.1). The severity of self-assessment can be closer to that of teacher assessment, but self-assessment raters were more lenient than teachers. On the other hand, peer assessment students were found to be fairly lenient. With respect to consistency, teacher assessment indicated reliable consistency in rating, but neither of the assessment types presented consistency in rating. Accordingly, neither self-assessment nor peer assessment was reliable in terms of severity and consistency, but self-assessment had the possibility of being comparable to teacher assessment from the perspective of its severity. On the other hand, it would be difficult to use peer assessment as a substitute for teacher assessment, because the peer assessment ratings were too generous compared to those obtained from teacher assessment and self-assessment. Therefore, the reliability of student assessment in terms of the composite score depended on student assessment types (Hypothesis 1.1).

4.1.2 Many-facet Rasch Analysis Results on the Analytic Rating Scales

4.1.2.1 Assumption Checks for Many-facet Rasch Analysis of Analytic Rating Scales. For an assumption check of many-facet Rasch analysis of analytic rating scales, unidimensionality and global model fit were tested. With respect to unidimensionality, the data from the PCA of four analytic rating scales indicated that the raw-score variance of observation was 0.66 (100.0 %), the variance explained by Rasch measures was 0.22 (34.60 %), and the variance of residuals was 0.43 (65.40 %). As described in Chapter 3, as the variance explained by Rasch measures exceeded the threshold value of 20 %, the result suggested that the unidimensionality of these data was supported.

As for the global model fit, a total of 14,016 responses were analysed for estimating parameter values for the student, rater, and task facets. Of these, 100 responses (or 7.13 %) were connected with absolute standardized residuals ≧ 3. Because this exceeded the threshold for an acceptable model fit (< 5 %), it did not present satisfactory model fit. As an alternative approach, the results of the log-likelihood chi-square were also checked. The obtained data log-likelihood chi-square value was 26,237.32 (approximate model *df* = 13,709; *p* > .001). Hence, the present data set could be said to satisfy the assumptions for fitting the Rasch model sufficiently.

4.1.2.2 The Wright Map for Four Analytic Rating Scales. Figure 4.1.2 shows a graphic ruler (the Wright map, or variable map) summarizing results of the FACETS analysis on the four analytic rating scales. The notations in the figure are the same as those for Figure 4.1.1. The only difference from Figure 4.1.1 is that the fourth column displays occasion by rating scale.

```
+---------------------------------------------------------------------------------------------+----
|Measr|+Ss          Rating methods        | Occasion * Rating scale                        |Scale|
|-----+------------+------------------------+------------------------------------------------+----
|  4  +            +                        +                                                + (4)|
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     | .          |                        |                                                |    |
|  3  + *          +                        +                                                +    |
|     | .          |                        |                                                |    |
|     | **         |                        |                                                |    |
|     | *          |                        |                                                |    |
|     | ***        |                        |                                                |    |
|     | ********.  |                        |                                                |    |
|     | *******    |                        |                                                |--- |
|  2  + ****.      +                        +                                                +    |
|       *********. |                        |                                                |    |
|     | *********. |                        |                                                |    |
|     | *******    |                        |                                                |    |
|     | *********  |                        |                                                |    |
|     | ***********|                        | Pre-Grammatical Accuracy                       |    |
|     | ******.    |                        |                                                | 3  |
|  1  + *****      +                        +                                                +    |
|     | ***.       |                        |                                                |    |
|     | ****       |                        | Post-Grammatical Accuracy                      |    |
|     | *.         |                        |                                                |    |
|     | *.         |                        | Post-Appropriate Usage of Vocabulary           |    |
|     | *.         | Teacher1  Teacher2  Teacher3 | Pre-Appropriate Usage of Vocabulary      |    |
|     | .          | Teacher4               |                                                |---|
*  0    *** * .*                             Post-Structure & Coherence                     |    |
|     |            | Self                   |                                                |    |
|     | .          |                        |                                                |    |
|     |            |                        | Pre-Structure & Coherence                      |    |
|     |            |                        |                                                |    |
|     | .          | Peers                  |                                                |    |
|     | .          |                        | Pre-Task Fulfilment                            |    |
| -1  +            +                        +                                                + 2 |
|     |            |                        | Post-Task Fulfilment                           |    |
|     | .          |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
| -2  + .          +                        +                                                +    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |---|
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
|     |            |                        |                                                |    |
| -3  +            +                        +                                                +(1) |
|-----+------------+------------------------+------------------------------------------------+-----|
|Measr| * = 3      |-Rater                  |-Task                                           |Scale|
+---------------------------------------------------------------------------------------------+----
```

Figure 4.1.2 Wright map for four analytic rating scales

Examinees' ability estimates in the second column ranged from a high of about 3.1 logits to a low of close to -2 logits. However, the majority of examinees were located above 0 on the logit scale. Therefore, examinees generally received high evaluations, suggesting that the tasks were not too difficult. The results on the task facet in the third column display the order of task difficulty from the most difficult to the least difficult as follows: *Grammatical Accuracy*, *Appropriate Usage of Vocabulary*, *Structure & Coherence*, and *Task Fulfilment*. Figure 4.1.2 also shows that the *Grammatical Accuracy* of the pretest was located distantly from the other three analytic scales. The order of difficulty of the rating scales for the post-test was the same order as that of the pretest in terms of the order of logits; in other words, *Grammatical Accuracy* was the most difficult component of all both before and after the intensive writing sessions. However, the distance of the *Grammatical Accuracy* of the post-test from the other three components of the post-test became closer than that of the pretest. Compared to the distances among the criteria on the pretest, the distances among them became smaller on the post-test (Figure 4.1.2). In general, of the four analytic rating scales, *Grammatical Accuracy* was the most difficult, while *Task Fulfilment* was the easiest analytic rating scale on both the pre- and post-tests.

In Figure 4.1.2, the severity of raters is presented in the third column from the left. The harshest raters were three teachers, namely Teacher 1, Teacher 2, and Teacher 3, while Teacher 4 followed closely after that. As for the student assessors, the self-assessment group students were more lenient than the four teachers. Peer assessment group students were the most lenient raters. The results of Wald statistics supported that the severity of teacher assessment was different from that of self- and peer assessment.

Given the changes and differences in difficulty estimates found among the individual rating scales between the pre- and post-tests, the Wald statistics were calculated for all pairs of the rating scales by occasion to test whether those differences were statistically significant, as shown in Tables 4.1.5 and 4.1.6.

Table 4.1.5

The Wald Statistics among the Four Analytical Rating Scales by Occasion

	Pretest			Post-test		
	Grammatical accuracy	Appropriate usage of Vocabulary	Structure & Coherence	Grammatical accuracy	Appropriate usage of Vocabulary	Structure & Coherence
Grammatical accuracy						
Appropriate usage of Vocabulary	-.54*			-2.82		
Structure & Coherence	-16.40*	-6.40*		-6.01	-3.18	
Task fulfilment	-19.20*	-.51*	-.19	-12.49*	-9.99*	-7.18*

Note. *$p < .05$

Table 4.1.6

The Wald Statistics for the Four Analytic Rating Scales between the Pretest and Post-test

Pre-Grammar & Post-Grammar	Pre-Vocabulary & Post-Vocabulary	Pre-Structure / Coherence & Post-Structure / Coherence	Pre-Task fulfilment & Post-Task fulfilment
-5.00*	-1.80	4.06*	-1.40

Note. *$p < .05$

First, when the analytic scales were compared against one another, applying a level of statistical significance of .05 and a critical value of 1.96 as the criteria, all of the results of the Wald statistics, except for the difference between *Structure & Coherence* and *Task Fulfilment* (the two easiest rating scales), showed statistically significant differences in difficulty in the pretest (Table 4.1.5). On the other hand, the three Wald statistics for the post-test were statistically significant, showing that *Task Fulfilment* was significantly easier than the other three analytical rating scales, namely *Grammatical Accuracy*, *Appropriate Usage of Vocabulary*, and *Structure & Coherence* (Table 4.1.5). Second, Table 4.1.6 presents the results of the Wald tests for four analytic rating scales between the pretest and post-test. The difficulty estimates for the pretest of *Grammatical Accuracy* and post-test of *Grammatical Accuracy* differed by -0.25 logits; this difference was significant, $t_{\text{pre-Grammar, post-Grammar}}$ (484) = -5.00, $p < .05$. Similarly, the difficulty estimates for the pretest of *Structure and Coherence* and the post-test of *Structure and Coherence* also differed by 0.23 logits, $t_{\text{pre-structure, post-structure}}$ (484) = 4.06, $p < .05$. On the other hand, the difficulty estimates for the pretest of *Appropriate Usage of Vocabulary* and the post-test of *Appropriate Usage of Vocabulary* differed by 0.09 logits, but this difference was not significant, $t_{\text{pre-vocabulary, post-vocabulary}}$ (484) = 1.80, *ns*.

4.1.2.3 Summary of Rasch Analysis Analytic Rating Scales of Pre- & Post-tests. In this section, the results of Rasch statistics are summarized in terms of separation ratio (*G*) and the separation (strata) index (*H*), and separation reliability (*R*). The results are specifically analysed from the perspective of examinees (students), raters, and criteria. The analysis of the present three-facet sample data is fully shown in Appendix G.

4.1.2.3.1 Examinees (students). The separation ratio (*G*) of examinees (students) indicated a value of 2.92, so the variability of the examinee proficiency measures was about three times larger than the precision of those measures. The separation (strata) index (*H*) shows the number of measurably different levels of examinee proficiency that was obtained by the examinee separation, or number of examinee strata, index. The value of this index was 4.23, suggesting that among the 243 examinees included in the analysis, there were about four statistically distinct classes of examinee proficiency. This corresponded to the four levels of writing evaluation; that is, the measurement system employed to produce at least as many reliably different levels of examinee proficiency as the writing test was supposed to differentiate.

The last separation statistic, i.e., the separation reliability (*R*), was .90, so examinees were reliably distinguished.

4.1.2.3.2 Raters (Self-assessors, Peer Assessors, and Four Teachers). The *G* value of the rater facet indicated that the variability of the severity measures was more than 11 times larger than that of precision. The value of the separation index (*H*) also indicated 15.16, which suggests that among the four teachers, self-assessors, and peer assessors included in the analysis, there were nearly 15 statistically distinct classes of rater severity – far more than would be expected in the adoption of the standard view with its implied objective of applying to raters from a homogeneous group. The separation reliability (*R*) was .99, therefore the severity of raters was distinctively varied.

4.1.2.3.3 Occasion by Rating Scale. The separation ratio (*G*) of the criteria of four components reached a value higher than the examinee or rater separation ratio: 20.43. The separation index for the criteria of four components also indicated a value greater than 27. As such, a value much greater than the number of criteria was actually included in the analysis. This means that the spread of the criterion difficulty measures was greater than the precision of those measures. Generally, when a large number of observations are available for each element in a given facet, standard deviation of the measure leads to high separation of criteria. As the separation reliability (*R*) was 1.00, the criterion reliability separation was close to its theoretical maximum. Therefore, it is considered that the levels of difficulty of the analytic rating scales, namely the four analytic rating scales of pre-and post-tests, were different from one another.

Summary Rasch statistics of analytic rating scales of pre- and post-tests show that: 1) the abilities of examinees (students) were almost divided into four different-level groups, corresponding to a four-point scale; 2) the strictness of the raters varied; and 3) the difficulties of the rating scales in the two occasions varied widely from one another. Hence, the results of

Rasch statistics of analytic rating scales did not suggest a specific difference between teachers and students, or between pre- and post-tests, so the next section reports the details of rater behaviour anlaysis in terms of raters' severity and consistency.

4.1.2.4 Rater Severity of Four Analytic Rating Scales. Table 4.1.7 presents the output from each rater measurement report obtained by Rasch statistics. Raters are arranged in descending order of severity. In detail, the severest rater was Teacher 2, while the most lenient rater was the peer assessment group. In terms of observed average, the ratings of the four teachers were close to each other. To be specific, the observed average of two of the teachers was 3.06 and that of the other two teachers was 3.07. On the other hand, the observed average of student assessors was higher than that of teachers: 3.23 for the self-assessment group; 3.45 for the peer assessment group. In short, the observed average of peer assessors was the highest of all the assessors, and that of teachers was the lowest of all of the raters. The mean of all the raters was 3.21.

Table 4.1.7

Raters' Measurement Report about Four Analytic Rating Scales

Raters	Observed Average	Measure logit	Model SE	Infit M Sq	Z Std	Outfit M Sq	Z Std
Teacher 2	3.06	.25	.03	.87	-5.2	.86	-5.1
Teacher 3	3.06	.23	.03	.86	-5.5	.85	-5.4
Teacher 1	3.07	.22	.03	.85	-7.4	.83	-7.7
Teacher 4	3.07	.21	.03	.84	-6.4	.83	-6.3
Self	3.23	-.15	.04	1.50	9.0	1.48	9.0
Peers	3.45	-.77	.04	1.57	9.0	1.62	9.0
Mean	3.21	.00	.03	1.08	-1.1	1.08	-1.1
SD (population)	.17	.37	.01	.32	7.2	.34	7.2

The reliability is .99 for all raters, so raters were reliably separated into different levels of severity. The fixed chi-square test also rejected the null hypothesis that all the raters are equal, with the chi-square value of 541.1 (*df* = 5; *p* = .00).

With regard to the four teachers' severity, Table 4.1.7 shows that the observed average scores assigned by teachers were 3.06 or 3.07. The reliability statistics of the four analytic rating scales confirmed that the four teachers were consistently severe in their evaluation of student writings on the four analytic rating scales with a separation reliability estimate of .00. Moreover, the fixed chi-square tests the null hypothesis that all the elements of the facet were equal. The chi-square was 1748.9 (*df* = 3, *p* = .65). This means that the null hypothesis is not rejected. That is to say, the four teachers' ratings were equally severe.

In contrast, both the self-assessment group and the peer assessment group indicated high reliability, though a low reliability is desirable because ideally the different raters would be equally severe: .75 for the self-assessment group, and .82 for the peer assessment group. Furthermore, both groups presented a variability of severity in the Separation Ratio (G): 1.75 for the self-assessment group, and 2.16 for the peer assessment group. In other words, the student assessment groups did not evaluate the four analytic rating scales in the same way in terms of severity.

Looking at all the raters' variance again, as the Rater Separation Ratio for the entire sample of raters indicated 12.19, the variance among all raters was about 12 times the error estimates. In other words, the rater severity across all raters was not equal in this analysis, so it can be said that the variety among all raters was caused by the variance of self- and peer assessment.

4.1.2.5 Rater Consistency of Four Analytic Rating Scales. The rater consistency of the four analytic rating scales was analysed in terms of the inter-rater agreement with teacher assessment, fit statistics, and bias analysis. First, the inter-rater agreement of the four analytic rating scales between raters with teacher assessment is shown in Table 4.1.8. Similarly to the inter-rater agreement of the composite score, neither student assessment type presented the same level of consistency as the teachers. In contrast, 97.6 % (17,093/17,520) agreement was obtained among the four teachers on the analytic rating scales, so the assessment assigned by the teachers was highly consistent.

Table 4.1.8

Inter-Rater Agreement of Four Analytic Rating Scales Between Raters

	Teachers	Self-assessment group	Peer assessment group
Teachers	97.6 % (17093/17520)	76.7 % (20256/26421)	79.0 % (19714/24966)

Second, the infit statistics were analysed in terms of bias analysis (the full report of bias analysis is shown in Appendix H). The bias analysis of the four analytic rating scales shows that the ratings of the self-assessment group and peer assessment group were unpredictable, i.e., inconsistent, as shown in the infit statistics that were both higher than 1.5 (Table 4.1.9). In contrast, the infit statistics for the four teachers' ratings ranged from 0.84 to 0.87, so it is considered that the four teachers' ratings were predictable and consistent.

Table 4.1.9

The Frequency of t-values Greater than 1.5

Assessors	Grammatical Accuracy		Appropriate Usage of Vocabulary		Structure & Coherence		Task Fulfilment	
	Pretest	Post-test	Pretest	Post-test	Pretest	Post-test	Pretest	Post-test
Self-Assessment	4		1	1	2		1	
Peer assessment	1	1	1	1	1			
Teacher 1							1	1
Teacher 2							1	1
Teacher 3							1	
Teacher 4							1	1
Sum	6/21		4/21		3/21		8/21	

Third, the bias analysis of the four analytic rating scales was reported in order to analyse the interaction between raters and the four analytic rating scales. Here, the *t*-value between raters of the four analytic rating scales and the frequency of *t*-values are focused on. As Table 4.1.9 and Table 4.1.10 show, it was also found that 21 *t*-values of 47, that is, about less than half of all *t*-values, were higher than +2 or lower than -2. In other words, these *t*-values suggested inconsistency among raters. In contrast to the bias analysis of the total sum, even teachers' *t*-values indicated rating inconsistency. With the frequency of *t*-values of the four components greater than 1.5, inconsistency in ratings were observed across analytical rating scales: *Grammatical Accuracy* accounted for 6 out of 21; *Appropriate Usage of Vocabulary* accounted for 4 out of 21; *Task Fulfilment* accounted for 8 out of 21; *Structure & Coherence* accounted for 3 out of 21 (Table 4.1.7). As regards the difference between the pretest and post-test, the frequency of components of the pretest was greater than that of the post-test: the frequency of the pretest was 15; the frequency of the post-test was 6. In sum, the rating of the four components indicated inconsistency of evaluation in both the student assessment and the teacher assessment.

However, all of the teachers showed consistency in ratings of the four analytic scales, except for the *Task Fulfilment* rating scale. On the other hand, students showed inconsistency. In all of the analytic rating scales of the pretest rating scores, the self-assessment method showed inconsistency in the pre-*Task Fulfilment* scale. As regards the post-test of *Task Fulfilment*, both self- and peer assessment methods showed consistency overall. However, Teacher 1 still presented rating inconsistency. As for the gap in understanding of Teacher 1, it should be analysed in the qualitative study, especially through an interview with Teacher 1 about how Teacher 1 understood and evaluated the *Task Fulfilment*.

As Table 4.1.10 shows, negative bias was frequently found in the difference in *t*-values between teachers and student assessors in terms of the four analytic rating scales. In particular, the largest difference was between peer assessors and teachers in all analytic rating scales. The difference in *t*-values between self-assessors and peer assessors was also noticeable, but the *t*-values of self-assessors were lower than those of peer assessors. These results are similar to the results of bias analysis of the composite score.

Table 4.1.10

t-Value between Raters of Four Analytic Rating Scales

	Self vs T1	*Self vs T2*	*Self vs T3*	*Self vs T4*	*Peer vs T1*	*Peer vs T2*	*Peer vs T3*	*Peer vs T4*	Self vs Peer
Pre-Task	-4.04	-3.50	-3.11	-2.94	-4.88	-4.75	-4.65	-4.56	3.06
Post-Task	-2.27	-0.61	-2.33	-2.21	-5.09	-5.25	-5.15	-5.04	2.42
Pre-Structure	-4.01	-3.69	-3.51	-3.32	-8.68	-8.01	-7.85	-7.69	5.73
Post-Structure	-2.42	-0.62	-2.35	-2.23	-7.37	-7.07	-6.93	-6.79	3.36
Pre-Vocabulary	-1.39	-3.32	-3.15	2.98	-8.22	-8.47	-8.30	-8.13	4.87
Post-Vocabulary	-2.56	-0.69	-2.63	-2.49	-5.53	-5.70	-5.59	-5.47	2.42
Pre-Grammar	-2.90	-3.14	-2.98	-2.82	-7.77	-8.01	-7.85	-7.69	4.87
Post-Grammar	-2.18	-0.69	-2.63	-2.63	-5.15	-6.00	-5.88	-5.76	-6.90

Note. T stands for teacher; underlined marker indicates greater than an absolute 2.

In sum, it was found that self- and peer assessment showed both similarities and differences in terms of reliability across the four analytic rating scales. First, with respect to the severity of the four analytic rating scales, teacher assessment was found to be the strictest among all rater types. Self- and peer assessors were not as strict in their rating as teachers. Peer assessors were the most lenient in terms of rating. Second, with regard to the consistency of the four analytic rating scales, teacher assessment presented consistency in all analytic rating scales except *Task Fulfilment*. On the other hand, self- and peer assessors did not present consistency in the four analytic rating scales, but the difference in *t*-values between teachers and self-assessors was less than that of peer assessors. Therefore, self- and peer assessment were not as reliable as teacher assessment in terms of the four analytic scales, although even teacher assessment partly showed inconsistent rating in *Task Fulfilment*.

4.1.3 Conclusion of RQ1: How do Self- and Peer Assessment Compare with each other in terms of: a) the Reliability of Scores against Teacher Assessment?

The purpose of this section was to investigate the reliability of self-assessment and peer assessment against teacher assessment, applying a many-facet Rasch measurement analysis. Based on the procedure described in Chapter 3, a many-facet Rasch analysis was conducted on pretest and post-test scores to address Research Question 1: How do self- and peer assessment compare with each other in terms of: a) the reliability of scores against teacher assessment? The reliability was analysed in the extent to which each assessment method had reliable severity and consistency in their assessment against teacher assessment. The results demonstrated that self- and peer assessment had different qualities in terms of reliability.

With respect to the composite score analysis, neither self-assessment nor peer assessment was reliable in terms of agreement with teacher assessment, but self-assessment was comparable to teacher assessment in terms of rater severity. On the other hand, it would be difficult to use peer assessment as a substitute for teacher assessment, because peer assessment was found to be different in terms of severity. The peer assessment rating was too generous compared to teacher assessment and self-assessment. As for rater consistency, teacher assessment was consistent, while self- and peer assessment were not.

In regard to rater consistency of the four analytic rating scales, teachers presented consistent ratings, but neither student method indicated rater consistency. It was found that the difficulty of the four analytic rating scales varied in consistency. Among them, *Task Fulfilment* was the easiest component of all of the analytic rating scales, while *Grammatical Accuracy* was the most difficult. This finding was in line with the results of a previous study (Oi, 2018). It was also found that teachers presented fairly reliable severity in assessment of the four analytic rating scales, while neither self- nor peer assessment methods presented reliable severity.

In conclusion, neither student assessment types presented the same level of consistency as that of teacher assessment (Hypothesis 1.1). However, self-assessment was similar to teacher assessment in terms of rater severity of the composite score.

4.2 The Effects of Student Assessment on the Improvement of Writing Ability

4.2.1 Introduction

This section mainly discusses Hypothesis 1.2, regarding the effects of student assessment on the improvement of writing ability. The results presented in Section 4.1 showed that both types of student assessment were not as consistent as teacher assessment, although self-assessment was comparable to teacher assessment in terms of rating severity. Yet, they did not address the issue of how self-assessment and peer assessment would influence the improvement of writing ability. Therefore, a many-facet Rasch measurement (MFRM) analysis was carried out to examine whether student involvement in self-assessment or peer assessment resulted in the improvement of writing ability in the present study. The scores assigned by teachers were used to examine the improvement of writing ability. With respect to the MFRM analysis, the effects of each student assessment type on the improvement of writing ability was analysed from two perspectives: the composite scores across four rating scales and analytic rating scores.

4.2.2 The Effects on the Improvement of Writing Ability

4.2.2.1 Descriptive Statistics. Table 4.2.1 presents descriptive statistics for the composite scores, including means and standard deviations of the pretest and post-test scores for the self-assessment group and the peer assessment group. As can be seen in the table, the means for the self-assessment group were higher than those for the peer assessment group in both the pre- and post-tests. As for the difference in standard deviation (SD) between the two types of student assessment, the SDs for the peer assessment group were larger than those for the self-assessment group on both the pre- and post-tests. In other words, the scores of the peer assessment group were spread out compared to those of the self-assessment group in both pre- and post-tests. With respect to the difference in means, the self-assessment group had higher scores than the peer assessment group in both pre- and post-tests, yet the means of the self- and peer assessment groups were very close, especially for the means of the pre-test: 12.27 for the self-assessment group and 12.16 for the peer assessment group.

Table 4.2.1

Descriptive Statistics for the Composite Scores of Pretest and Post-test of Self-assessment and Peer assessment Groups

		N	Min	Max	Mean	SD
Pretest	SA	147	4	16	12.27	1.64
	PA	146	4	16	12.16	1.87
	Total	293	4	16	12.21	1.76
Post-test	Post-test of SA	147	6	16	12.52	1.44
	Post-test of PA	146	6	16	12.33	1.62
	Total	293	6	16	12.43	1.54

Note. SA = self-assessment group; PA = peer assessment group

4.2.2.2 Composite Scores of Analytic Rating Scales in terms of the MFRM Analysis. The composite scores of the analytic rating scales assigned by teachers were analysed to investigate the effects of each student assessment type on the improvement of writing ability. Here, the data of the MFRM analysis obtained in Section 4.1 were analysed for the improvement of writing ability in terms of task difficulty based on the results of the four teachers' assessment.

According to the results of the composite scores assigned by teachers exhibited in Section 4.1, the task difficulty between the pretest and post-test indicated similarity to each other. Therefore, the improvement of writing ability was analysed by focusing on the change in task difficulty in terms of the composite scores assigned by teachers.

Table 4.2.2 presents the task measurement reports on the composite scores of the four analytic rating scales assigned by teachers. Measure logits of the pretest were .33 for self-assessment group students and -.33 for peer assessment group students. This means that the pretest was more difficult for the self-assessment group than for the peer assessment group students. The post-test also showed the same results: the measure logits of the post-test was .05 for the self-assessment group and -.05 for the peer assessment group. Meanwhile, the entire group showed an improvement of writing ability: measure logits of .12 for the pretest and -.12 for the post-test. Therefore, it is considered that, regardless of the assessment methods, student assessment lead to the development of writing ability. This also indicates that self-assessment affected the improvement of writing ability more than peer assessment.

Table 4.2.2

Task Measurement Report about the Composite Scores of Four Teachers

Statistic	SA		PA		All students	
	Pretest	Post-test	Pretest	Post-test	Pretest	Post-test
Observed Average	12.27	12.17	12.71	12.25	12.23	12.43
Measure logit	.33	.05	-.33	-.05	.12	-.12
Model *SE*	.05	.04	.06	.04	.03	.03
Infit *M* Sq	.95	1.03	1.04	.93	1.00	.97
Z Std	-.8	.5	.6	-1.2	.0	-.7
Outfit *M* Sq	.92	1.03	1.04	.93	.99	.96
Z Std	-1.3	.5	.6	-1.3	-.1	-.8
Estm. Discrm	1.06	1.00	.95	1.02	1.02	.99

Note. SA indicates self-assessment group students; PA stands for peer assessment group students.

4.2.2.3 Analytic Rating Scores. In this section, the levels of difficulty estimates of the four analytic rating scales are analysed, borrowing the data that were generated based on the results of the MFRM analysis in Section 4.1. Figure 4.2.1 presents the difference in logit of the four analytic rating scales between pre- and post-tests.

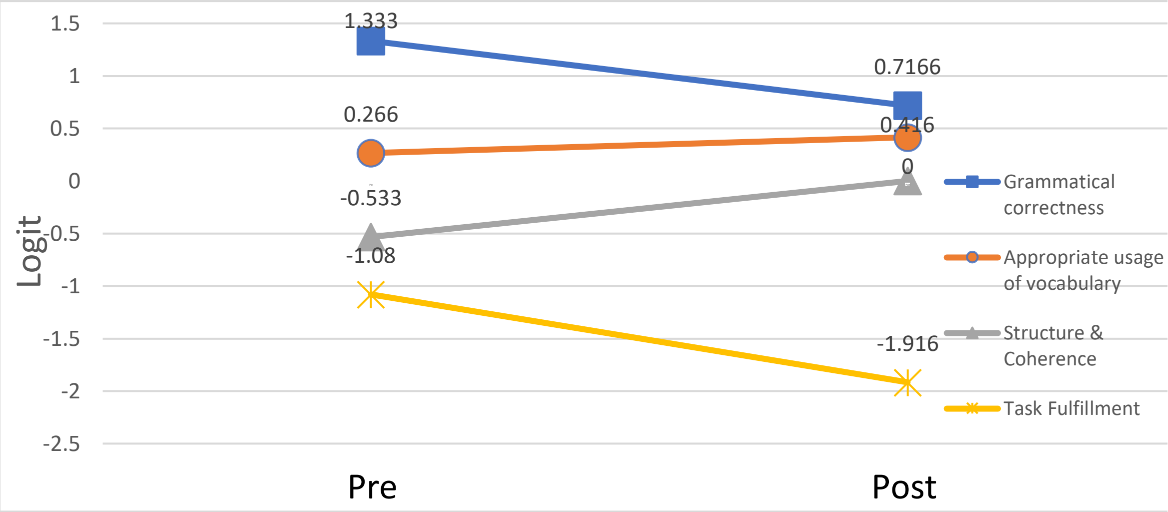


Figure 4.2.1 Pre- and post- four analytic rating scales

As Figure 4.2.1 shows, the pre- and post-tests were rank-ordered the same in terms of difficulty, from the most difficult to the least difficult, *Grammatical Accuracy, Appropriate Usage of Vocabulary, Structure & Coherence,* and *Task Fulfilment*. In terms of the change of logits between pre- and post-test, the logits of *Grammatical Accuracy* and *Task Fulfilment* decreased in the post-test. This means that *Grammatical Accuracy* and *Task Fulfilment* were easier than those of the pretest in terms of task difficulty. On the other hand, the logits for the rating scales of *Appropriate Usage of Vocabulary* and *Structure & Coherence* somewhat increased; in other words, the tasks became more difficult in the posttest than in the pretest.

Table 4.2.3 presents a task measurement report about the four analytic rating scales for both self- and peer assessment groups. The logit values decreased from 1.26 for the pretest to .69 for the posttest on the *Grammatical Accuracy* scale. This is the largest difference among all the analytical rating scales between pre- and post-tests.

Table 4.2.3

Task Measurement Report for the Four Analytical Rating Scales

Analytical Rating Scales	Observed Average	Measure logit	Model *SE*	Infit *M* Sq	*Z* Std	Outfit *M* Sq	*Z* Std
Pre-Grammar	2.56	1.26	.03	.98	-.6	.98	-.5
Post-Grammar	2.81	.69	.04	.77	-7.6	.79	-6.8
Post-Vocabulary	2.97	.37	. 04	.70	-9.0	.71	-9.0
Pre-Vocabulary	3.06	.25	.03	.70	-9.0	.72	-9.0
Post-Structure & Coherence	3.15	.00	.04	1.20	5.9	1.20	5.7
Pre-Structure & Coherence	3.38	-.48	.04	1.13	4.2	1.13	3.9
Pre-Task Fulfilment	3.52	-.89	.04	1.46	9.0	1.26	6.7
Post-Task Fulfilment	3.61	-1.20	.05	1.32	6.9	1.35	6.5
Mean	3.13	.00	.04	1.03	.0	1.02	-.3
SD (population)	.34	.77	.00	.27	7.1	.24	6.6
SD (sample)	.36	.82	.01	.29	7.6	.26	7.0

4.2.2.4 Analytic Rating Scales of Self- and Peer Assessment Groups in terms of the MFRM Analysis. In this section, it is analysed what analytic rating scales specifically influenced the improvement of the composite scores in self- and peer assessment groups. Before conducting the MFRM analysis, assumption checks were carried out.

4.2.2.4.1 Assumption Checks for Many-facet Rasch Analysis of Analytic Rating Scales of Self- and Peer Assessment Groups. Before employing the MFRM analyses, the unidimensionality and global model fit were tested in order to check satisfactory model fit. With respect to unidimensionality, Engelhard (2013) stated that unidimensionality is sufficiently accepted in the case that the variance explained by Rasch measures is ≧ 20 %. The value for the present dataset was 39.3 % based on a PCA, which could be interpreted that the unidimensionality was accepted.

With regard to the global model fit, in the current study, a total of 9,344 responses were employed for estimation of (non-extreme) parameter values. Of these, 100 responses (or 11.5 %) were related to absolute standardized residuals ≧ 3. Because of the exceeding of the threshold for an acceptable model fit (< 5 %), satisfactory model fit was not suggested. Therefore, log-likelihood chi-square was examined. The obtained data log-likelihood chi-square value of 16,898 (approximate model *df* =16,864.73; *p* > .001), which was not statistically significant, suggested the data fit for the Rasch analysis. Thus, it was decided to proceed with the Rasch analysis.

4.2.2.4.2 Analytic Rating Scores of Self- and Peer Assessment Groups. Based on the results of the MFRM analysis (the task measurement report is shown in Appendix I), Figure 4.2.2 was generated to present the difference in logits of each assessment type between pre- and post-tests. The rank-ordering of the analytic rating scale in terms of difficulty did not change except for *Structure and Coherence* of the self-assessment group and that of the peer assessment group between pre- and post-tests. In the post-test, the task difficulty of analytic rating scales tended to be varied slightly less compared to that of the pre-test. The most difficult task was *Grammatical Accuracy* of the self-assessment group, while the easiest task was *Task Fulfilment* of the peer assessment group. As shown in Figure 4.2.2, the logit of *Grammatical Accuracy* decreased in both self- and peer assessment groups. In other words, *Grammatical Accuracy* became easier for both assessment groups. The other analytic rating scales did not present a common increase or decrease of logits between the pre- and post-tests.

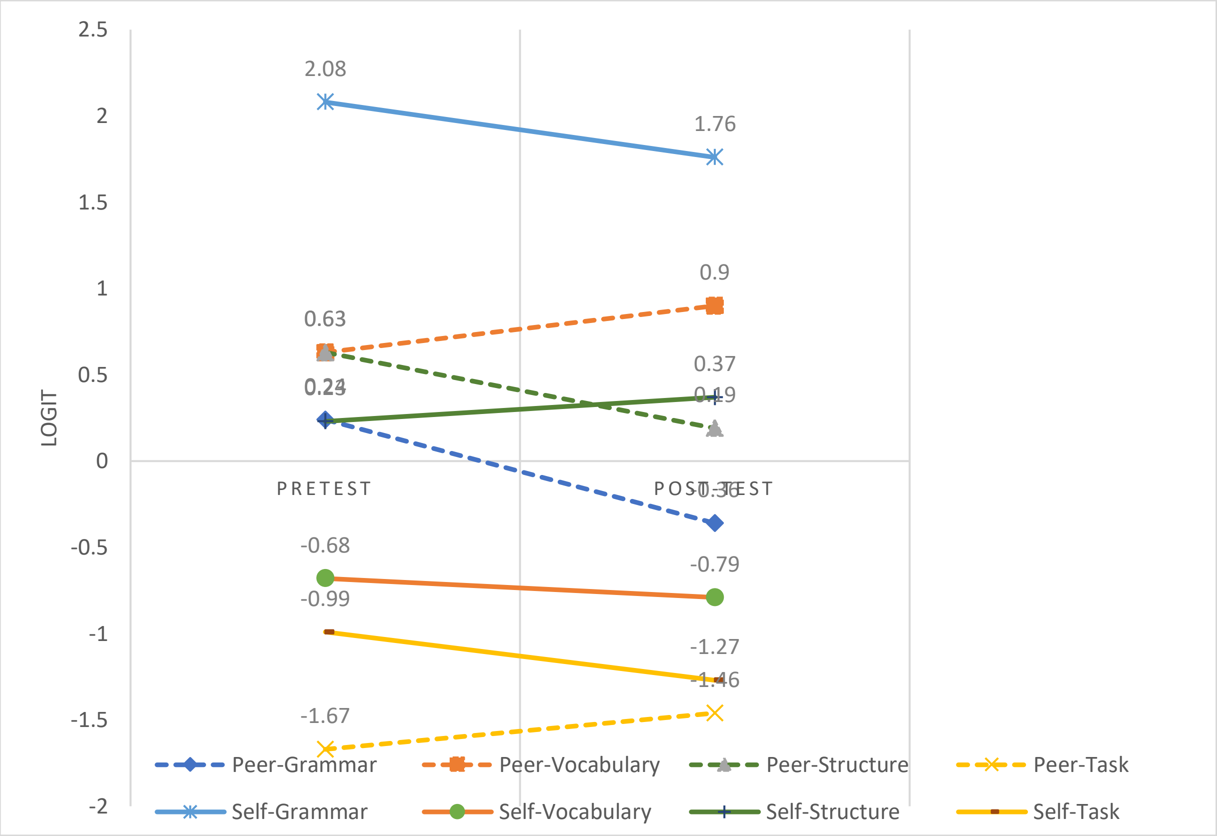


Figure 4.2.2 Pre and post four analytic rating scales of self- & peer assessment groups

Table 4.2.4 shows the report of the measure logits of task measurement and Wald statistics about self- and peer assessment group students. Here, the effect of each assessment type on the improvement of writing ability is analysed in terms of two aspects: (1) measure logits of task measurement; and (2) Wald statistics. The full report of task measurement is shown in Appendix I.

Table 4.2.4

The Measure Logits and Wald Statistics Report about Self- & Peer Assessment Group Students

	Self-assessment group					Peer assessment group				
	Observed Average		Logit		Wald (pretest & post-test)	Observed Average		Logit		Wald (pretest & post-test)
	Pretest	Post-test	Pretest	Post-test		Pretest	Post-test	Pretest	Post-test	
Composite	12.27	12.71	.33	-.33	-5.63*	12.17	12.25	.05	-.05	-1.02
Grammar accuracy	2.33	2.64	2.08	1.76	-2.74*	2.22	2.74	.24	-.36	-5.25*
Vocabulary	2.98	3.07	-.68	-.79	-1.06	2.80	2.96	.63	.9	-1.61
Structure & Coherence	3.34	3.42	.23	.37	-.94	2.96	3.39	.63	.19	-.43
Task fulfilment	3.58	3.62	-.99	-1.27	-.47	3.52	3.68	-1.67	-1.46	-1.61

Note. **p* < .05

Firstly, the results of measure logits are explained. According to Table 4.2.4, measure logits of self-assessment group students were 2.08 for Pre-*Grammatical Accuracy*, and 1.76 for Post-*Grammatical Accuracy*. Similarly, the measure logits for the peer assessment group were .24 for Pre-*Grammatical Accuracy* and -.36 for Post-*Grammatical Accuracy*. The Peer assessment group showed a greater difference between the pre- and post-tests on *Grammatical Accuracy*: .32 for the self-assessment group and .60 for the peer assessment group.

The measure logits of *Task Fulfilment* showed the different results of self- and peer assessment group. The logit estimate for the self-assessment group was -.99 for Pre-*Task Fulfilment*, while that for Post-*Task Fulfilment* was -1.27. On the other hand, the logit estimates for the peer assessment group for the same rating scale were -1.67 for the pre-test, and -1.46 for the post-test. In short, the development of post-task performance was observed in *Grammatical Accuracy* for both groups and in *Task Fulfillment* for the self-assessment group only. Hence, it is considered that *Grammatical Accuracy* for both assessment types were similarly affected by intensive student assessment methods.

Secondly, Wald statistics (Table 4.2.4) were applied to analyse whether there was any difference in writing ability between pretest and post-test for each assessment method, i.e., self-assessment and peer assessment. Wald statistics were examined to test the null hypothesis that there is no difference with the critical value of 1.96 at the .05 significance level as the criterion. According to Table 4.2.4, as for the self-assessment group, statistical significance was found in composite scores and *Grammatical Accuracy* between pre- and post-test. As regards the composite scores of the self-assessment group students, there was the difference of 1.44 logits between pretest and post-test in terms of task difficulty. Wald statistics validated that it was statistically significant, $t_{\text{pre-composite, post-composite}}$ (283)

= 5.63, $p < .05$, so the null hypothesis was rejected. The difference in the task difficulty between Pre- and Post-*Grammatical Accuracy* was also rejected: $t_{\text{pre-grammar, post-grammar}}$ (113) = -2.74, $p < .05$. In short, it is considered that self-assessment group students showed different task difficulty in the composite scores and *Grammatical Accuracy* between the pre- and post-test. In other words, the composite scores and *Grammatical Accuracy* in the post-test became easier tasks for the students in the self-assessment group than those of the pretest.

With respect to the peer assessment group, only *Grammatical Accuracy* presented a difference between pretest and post-test: $t_{\text{pre-grammar, post-grammar}}$ (181) = -5.25, $p < .05$. Unlike for the self-assessment group, however, the results of Wald statistics did not indicate any difference in the task difficulty of the composite scores between pretest and post-test for the peer assessment group.

In summary, the self-assessment group showed improvement between pretest and post-test in terms of the composite scores and *Grammatical Accuracy* after the intensive student assessment sessions. On the other hand, the composite scores of the peer assessment group did not present any difference between pre- and post-tests even after the intensive sessions, but it is considered that the *Grammatical Accuracy* of students in the peer assessment group was improved.

4.2.3 Conclusion of RQ 1: How do Self- and Peer Assessment Compare with each other in terms of: b) the Effects on Writing Ability?

The results of the MFRM analysis and Wald statistics indicated that self-assessment influenced the improvement of writing ability more strongly than peer assessment after the intensive student assessment sessions. This is because self-assessment had positive effects on the improvement of composite scores and *Grammatical Accuracy*. On the other hand, the peer assessment group showed a positive effect only on *Grammatical Accuracy*. In sum, each assessment type influenced the improvement of writing ability differently (Hypothesis 1.2). To be specific, both assessment types positively affected the improvement of *Grammatical Accuracy*. The reasons for the similarities and differences between each student assessment type are explored in the qualitative analysis in Chapter 6.

4.3 The Effects of Student Assessment on Writing Anxiety and Learner Autonomy

4.3.1 Introduction

This section discusses Hypothesis 1.3: the effects of intensive writing practice with student assessment on writing anxiety and learner autonomy. In order to address this hypothesis, descriptive statistics were firstly examined to determine the level of writing anxiety and learner autonomy of the student participants. Secondly, a multivariate analysis of covariance (MANCOVA) was performed to determine the effects of student assessment type on three subscales of writing anxiety and the learner autonomy scale at the end of the intensive student assessment period.

4.3.2 Preliminary analyses

4.3.2.1 Descriptive Statistics. Table 4.3.1 summarizes the descriptive statistics (mean scores and standard deviations) for the subscales of the questionnaires on writing anxiety and learner autonomy conducted before and after the intensive student assessment sessions. In terms of the reliability estimates of the questionnaire, the Cronbach's alpha coefficient was 0.70 for writing anxiety (22 items across the three subscales), and .68 for learner autonomy (10 items).

Table 4.3.1

Descriptive Statistics for Writing Anxiety and Learner Autonomy

Assessment method	Total scores & subscales	NO of Items	Pretest				Post-test			
			Min	Max	Mean	SD	Min	Max	Mean	SD
Self-assessment group (*N* = 147)	Overall writing anxiety	22	1.00	3.45	2.12	.32	1.00	4.00	2.16	.31
	Somatic anxiety	7	1.00	4.00	2.09	.64	1.00	4.00	1.98	.70
	Avoidance behaviour	7	1.00	3.00	2.08	.47	1.00	4.00	2.18	.46
	Cognitive anxiety	8	1.00	3.62	2.17	.49	1.00	4.00	2.30	.43
	Learner autonomy	10	1.00	3.40	1.93	.60	1.00	3.30	2.23	.63
Peer assessment group (*N* = 146)	Overall writing anxiety	22	1.09	3.09	2.22	.39	1.00	3.13	2.21	.29
	Somatic anxiety	7	0.57	3.85	2.26	.74	1.00	3.71	2.18	.66
	Avoidance behaviour	7	1.14	3.14	2.11	.58	1.00	3.00	2.14	.45
	Cognitive anxiety	8	1.12	3.00	2.29	.55	1.00	3.25	2.31	.43
	Learner autonomy	10	1.00	3.70	1.96	.60	1.20	4.00	2.33	.69
All students (293)	Overall writing anxiety	22	1.00	3.45	2.14	.36	1.00	4.00	2.18	.30
	Somatic anxiety	7	0.57	4.00	2.18	.70	1.00	4.00	2.08	.68
	Avoidance behaviour	7	1.00	3.85	2.09	.53	1.00	4.00	2.16	.45
	Cognitive anxiety	8	1.00	3.62	2.24	.52	1.00	4.00	2.30	.43
	Learner autonomy	10	1.00	3.70	1.95	.60	1.00	4.00	2.26	.66

Note. The figures in this table are based on the mean rating across items in each scale for each student.

As for the mean scores of overall writing anxiety on the questionnaire conducted as the pretest, the mean across students in the self-assessment group was *M* = 2.12 (SD = .32), while that for the peer assessment group was *M* = 2.22 (SD = 39). These results suggest that, before the intensive writing sessions, the students in the peer assessment group indicated somewhat greater writing anxiety than those in the self-assessment group, although this difference between the two groups was not statistically significant with a small effect size (*t* (291) = .61, *p* = .53, *r* = .03). Because the ratings of 2 and 3 on the Likert scale corresponded to "Disagree" and "Agree" to statements about writing anxiety respectively, it could be said that both groups showed low levels of writing anxiety initially. In addition, the standard deviations for both assessment types were also similar to each other, so the self- and peer assessment group were similar in terms of the variability of the writing anxiety level before the experiment as well.

With respect to learner autonomy, both student assessment type groups had similar mean scores on the pre-questionnaire (*M* = 1.93, SD = .60 for the self-assessment group; *M* = 1.95, SD = .60 for the peer assessment group). The independent *t*-test also indicated that this difference was not significant (*t* (291) = -0.38, *p* > .05) with a small sized effect (*r* = .02). Accordingly, both groups were considered to have similar and low levels of learner autonomy before the intensive writing sessions. Additionally, the standard deviations for both assessment types were identical to each other, therefore, the variability of the learner autonomy level was the same between two groups before the intensive sessions. To sum up, it is considered that the levels of writing anxiety and learner autonomy were similar between the self- and peer assessment groups before the intensive writing sessions.

With regard to the results of the post-questionnaire compared to the pre-questionnaire, the mean overall anxiety score across students in the self-assessment group increased slightly (*M* = 2.12 for the pre-questionnaire; *M* = 2.16 for the post-questionnaire). On the other hand, the mean overall anxiety across students in the peer assessment group decreased, although the difference was subtle (*M* = 2.22 for the pre-questionnaire; *M* = 2.21 for the post-questionnaire). The level of writing anxiety of both assessment types were also categorized into slightly lower level of writing anxiety in post-tests. As for the standard deviations, the discrepancies existed in the data obtained from the post-questionnaire. The variability in writing anxiety became larger for the self-assessment group than for the peer assessment group. In short, the level of writing anxiety did not change between the pretest and the post-test in either group, but the degree of score variability differed slightly between the self- and peer assessment groups on the post-test. With respect to the writing anxiety subscale scores, the patterns of score increase and decrease were very similar between the two groups. With the mean scores around 2 on the Likert scale, of the writing anxiety levels of both groups were considered as slightly lower in both questionnaires conducted as the pre- and post-tests. With respect to the standard deviations, the self-assessment group showed the slightly larger variability than the peer assessment group did on the posttest (SD = .31 for the self-assessment group; *M* = .29 for the peer assessment group). Thus, both self- and peer assessment groups were similar in the writing anxiety level and its variability in both occasions.

With regard to learner autonomy presented in Table 4.3.1, both student assessment type groups increased the means from the pretest to the posttest (from *M* = 1.93 to *M* = 2.23 for the self-assessment group; from *M* = 1.96 to *M* = 2.33 for the peer assessment group). Therefore, it is considered that both student assessment types slightly increased the students' learner autonomy levels after the intensive writing sessions. As for SD, while the values were identical for both groups in the pretest (SD = .60) as noted above, the score variability increased slightly and with a somewhat larger value for the peer assessment group than for the self-assessment group on the questionnaire conducted as the post-test (SD = .63 for the self-assessment group; SD = .69 for the peer assessment group). In sum, both groups were similar in having low levels of learner autonomy before the intensive writing sessions. After that, however, both assessment types slightly increased the learner autonomy levels.

4.3.2.2 Preliminary Assumption Checks for MANCOVA. Before the employment of MANCOVA, preliminary assumption checks were conducted in terms of (1) the identification of outliers; (2) checks for normality, absence of multicollinearity, and homogeneity of variance in regard to the dependent variables and (3) additional assumption checks including the covariates.

First, as for univariate outliers, following Tabachnick and Fidell's (2014) criterion, an absolute value of *z* score greater than 3.29 $p < .01$, no outliers were identified on the four dependent variables (the mean writing anxiety subscale and the learner autonomy scale scores on the post-test). Regarding the check for multivariate outliers for each of the dependent variables, 14 Mahalanobis outliers were identified for each group. After excluding these outliers, 133 cases for the self-assessment group and 132 cases for the peer assessment group were retained in the final dataset used for subsequent analyses. Therefore, the remaining assumption checks described below were conducted on this final dataset.

With respect to normality, skewness and kurtosis values obtained for each of the dependent variables by group suggested univariate normality of the score distributions of the dependent variables. Box's test of equality of variance-covariance matrices was non-significant, suggesting that the matrices were equal between the two groups. Univariate F tests were also conducted in order to examine homogeneity of variance between the two student assessment groups on the four dependent variables by using Levene's test. Given that the results were statistically nonsignificant (*p* > .005), the assumption of homogeneity of variance was met. Next, multicollinearity was tested by examining Pearson correlations among the dependent variables. According to Tabachnick and Fidell (2014), it is advisable for conducting a MANOVA that the correlations among dependent variables are not high (p. 310). As Table 4.3.2 shows, none of the dependent variables were highly correlated with one another. Thus, multicollinearity among the dependent variables was not a concern.

Next, a series of assumptions involving the covariates (the mean writing anxiety subscale and learner autonomy subscale score) were examined. Linearity was checked by examining scatterplots among the dependent variables and covariates. There was no tendency of curvilinearity. The test of homogeneity of regression slopes showed that there was no significant interaction between the covariates and the grouping variables (assessment types), so the homogeneity of regression assumption was confirmed. The significance of the regression slopes was also checked. There was no interaction effect between assessment types and the covariates (subscales of pre-questionnaire), so the regression slopes were parallel between the groups. Finally, independence of covariates was checked. The *t*-test results comparing the writing anxiety and learner autonomy subscales between the two student assessment groups were not statistically significant, confirming the independence of covariates from the independent variables. In sum, it was interpreted that all assumptions for conducting the MANCOVA were met.

Table 4.3.2

Observed Pearson Correlation Matrix of Observed Scores between Subscales of Writing Anxiety & Learner Autonomy

		Pre-somatic	Pre-avoidance	Pre-cognitive	Pre-autonomy	Post-somatic	Post-avoidance	Post-cognitive
Self-assessment group	Pre-somatic							
	Pre-avoidance	.39*						
	Pre-cognitive	.62*	.47*					
	Pre-autonomy	.04	-.06	-.09				
	Post-somatic	.12	.04	.16	.10			
	Post-avoidance	.02	.15	.08	.05	.68*		
	Post-cognitive	.04	.07	.16	.16	.58*	.57*	
	Post-autonomy	-.01	.04	.04	.58*	.03	.03	.12
Peer assessment group	Pre-somatic							
	Pre-avoidance	.62*						
	Pre-cognitive	.66*	.66*					
	Pre-autonomy	-.04	-.01	-.04				
	Post-somatic	.11	.09	.09	.04			
	Post-avoidance	-.06	.02	-.02	-.06	.49*		
	Post-cognitive	.07	.08	.10	.02	.60*	.48*	
	Post-autonomy	-.09	-.13	-.16	.57*	.01	-.01	-.01

Note. **p* < .05

4.3.3 MANCOVA results

4.3.3.1 The Results of MANCOVA. A one-way MANCOVA was conducted to investigate the difference in the students' writing anxiety and learner autonomy between the questionnaires conducted as the pre- and post-tests for both assessment types. There were four dependent variables (the mean post-somatic, post-avoidance, and post-cognitive writing anxiety scores, and the mean post-learner autonomy score) and one independent variable (student assessment type: self-assessment and peer assessment). The covariates were the mean pre-somatic, pre-avoidance, and pre-cognitive writing anxiety scores, and the mean pre-learner autonomy scores.

According to the results of Pillai's trace, there was no main effect of assessment type (the between-subject variable) on the multivariate distribution of the four dependent variables: $F(1, 263) = .665, p = .617, \eta_p^2 = .00$). Next, the univariate tests results were checked to examine the effects of the within-subject variables on the four dependent variables. There was no interaction effect between assessment type and occasion on any of the dependent variables. According to the results presented in Table 4.3.3, there was a main effect of occasion (pre- and post-questionnaires) on learner autonomy with a large effect size (F (1, 263) = 31.82, $p < .001$, $\eta_p^2 = .306$). According to Mizumoto and Takeuchi (2008), the effect size of η_p^2 is influenced by the number of dependent variables (King & Minium, 2003), so η^2 was also calculated: $\eta^2 = .24$ (Field, 2009). It is concluded that both self- and peer assessment groups slightly increased learner autonomy with a large effect size after the intensive student assessment sessions.

Table 4.3.3

The Univariate Analysis Results for the Subscale Scores (MANCOVA)

Source	*DV*	*SS*	*df*	*MS*	*F*	*P*	η_p^2
			Between Subjects				
Assessment types	Somatic	.66	1	.66	.02	.88	.00
	Avoidance	28.58	1	28.58	1.07	.30	.004
	Cognitive	4.70	1	4.70	.18	.67	.00
	Autonomy	45.37	1	45.37	1.52	.21	.00
Error	Somatic	9131.89	263	31.38			
	Avoidance	7769.25	263	26.69			
	Cognitive	7532.89	263	25.88			
	Autonomy	8680.63	263	29.83			
			Within Subjects				
Pre-questionnaire & Post-questionnaire	Somatic	64.04	1	64.04	9.28	.82	.03
	Avoidance	366.55	1	366.55	13.15	.16	.04
	Cognitive	74.27	1	74.27	9.77	.66	.03
	Autonomy	2642.50	1	2642.50	93.57	.00*	.24
Assessment types x Pre-/Post questionnaire	Somatic	77.90	1	77.90	11.29	.09	.03
	Avoidance	41.75	1	41.75	1.49	.22	.00
	Cognitive	20.75	1	20.75	2.73	.09	.00
	Autonomy	23.06	1	23.06	.81	.36	.00
Error	Somatic	2007.03	263	6.89			
	Avoidance	8110.84	263	27.87			
	Cognitive	2211.20	263	7.59			
	Autonomy	8218.09	263	28.24			

*$p < .05$

4.3.3.2 The Analysis of the Correlations among Writing Anxiety and Learner Autonomy Scales. Table 4.3.4 presents the adjusted correlations among the three writing anxiety subscales and the learner autonomy scale obtained from the adjusted data from the MANCOVA for each student assessment group. The results of this analysis suggest four tendencies. In general, there were similar patterns of correlations observed between the two assessment type groups and between the two occasions, that is, the pre- and post-questionnaires. First, learner autonomy was not correlated with any of the three types of writing anxiety. To be specific, for the self-assessment group, the correlations among learner autonomy, somatic anxiety, avoidance, and cognitive anxiety were not significant. Also, for the peer assessment group, the correlations among learner autonomy, somatic anxiety, avoidance, and cognitive were not significant. Second, there were positive and moderate correlations among the three measures of writing anxiety for each group in both occasions ($.30 \leqq r \leqq .68$ for self-assessment group; $.48 \leqq r \leqq .69$ for peer assessment group). Finally, the learner autonomy level on the pretest was positively correlated with that on the posttest ($r = .56$ for self-assessment group; $r = .55$ for peer assessment group).

In sum, both assessment types had similar tendencies in both writing anxiety and learner autonomy across the questionnaires conducted as pre- and post-tests. Learner autonomy was not correlated with any of the writing anxiety subscales, while the subscales of writing anxiety were moderately correlated with one another. In addition, those students who had higher learner autonomy before the intensive student assessment sessions tended to have higher learner autonomy after the intensive student assessment sessions as well.

Table 4.3.4

Correlation Matrix between Subscales of Writing Anxiety & Learner Autonomy

		Pre-somatic	Pre-avoidance	Pre-cognitive	Pre-autonomy	Post-somatic	Post-avoidance	Post-cognitive
Self-assessment group	Pre-somatic							
	Pre-avoidance	.30*						
	Pre-cognitive	.56*	.38*					
	Pre-autonomy	.03	-.12	-.13				
	Post-somatic	.09	.05	.15	.05			
	Post-avoidance	.02	.17	.08	.02	.68*		
	Post-cognitive	.04	.08	.17	.13	.58*	.58*	
	Post-autonomy	-.03	-.00	.02	.56**	.00	.02	.09
Peer assessment group	Pre-somatic							
	Pre-avoidance	.52*						
	Pre-cognitive	.58*	.69*					
	Pre-autonomy	.01	.06	.01				
	Post-somatic	.02	-.01	.00	.11			
	Post-avoidance	-.15	-.07	-.11	-.04	.48*		
	Post-cognitive	.04	.06	.08	.07	.59*	.48*	
	Post-autonomy	-.04	-.07	-.12	.55*	.06	.00	.01

Note. *$p < .05$

4.3.4 Conclusion of RQ 1: How do Self- and Peer Assessment Compare with each other in terms of: c) the Effects on Writing Anxiety and Learner Autonomy?

The Research Question 1.3 asked: How do self- and peer assessment compare with each other in terms of: c) the effects on writing anxiety and learner autonomy? Neither self-assessment nor peer assessment affected writing anxiety, while the intensive student assessment seemed to have some effects on learner autonomy.

Before the experiment, it was hypothesized that both types of student assessment would have positive effects on the development of learner autonomy but that writing anxiety would be affected only by self-assessment (Hypothesis 1.3). This is because both assessment types were considered to encourage students to autonomously think and write in English. Another reason was that self-assessment would provide students with opportunities to reflect on themselves, and such a reflection might decrease writing anxiety. Previous studies suggest that self-assessment would reduce anxiety about writing performance (MacIntyre, Noels, & Clement, 1997; McDonald & Boud, 2003; Tsui & Ng, 2000). On the other hand, very few studies on peer assessment have focused on its effects on learner affect such as writing anxiety.

In the current study, the MANCOVA results did not suggest the effects of either student assessment type on learner writing anxiety, while a positive effect of both assessment types on learner autonomy was observed for both assessment types. The correlations between the scores of subscales showed similar patterns between the two assessment types as well. Thus, this finding supports the view that self-assessment and peer assessment are similar to each other in their effects on writing anxiety. Regarding learner autonomy, the correlation coefficients between pre- and post-learner autonomy levels obtained from the MANCOVA were positive and moderate in both self- and peer assessment groups. Therefore, it is also considered that both assessment types had positive effects on learner autonomy with a large effect size, and that those students who had higher learner autonomy tended to have higher learner autonomy after the intensive student assessment sessions.

4.4 Summary of the Quantitative Analyses

Chapter 4 presented the quantitative data analysis results. In relation to Research question 1, it was found that self- and peer assessment had similarities and differences in terms of reliability as well as on the effects on student writing performance and learner affect.

First, the reliability of student assessment was examined by employing a many-facet Rasch analysis (Section 4.1) for the composite score across four analytic ratings (Task Fulfilment, Structure and Coherence, Appropriate Usage of Vocabulary, and Grammatical Accuracy) and for the individual analytic ratings from the perspective of severity and consistency (Hypothesis 1.1). With respect to the composite score, the severity of ratings assigned by self-assessors was close to that of teacher assessment, but peer assessors were more lenient than teachers and self-assessors. Regarding the consistency, ratings assigned by neither student assessment group were as consistent as teachers' ratings. The many-facet Rasch analysis results for the individual analytic scores were similar to the results of the composite score. In other words, neither self- nor peer assessors were different from the consistency of teachers. In sum, as for the composite score, both assessment types could not present the comparable consistency to teacher assessment, while only self-assessment group had comparable severity to teacher assessment. On the other hand, with respect to the analytic rating scales, neither assessment type could not present similar levels of severity nor consistency to those of teacher assessment.

Second, the effects of student assessment on the improvement of writing ability is reported (Hypothesis 1.2). Through a many-facet Rasch analysis, it was found that both student assessment methods had positive effects on the improvement of grammatical accuracy, while only self-assessment resulted in the improvement of composite score across the four analytic rating scales (Section 4.2).

Finally, the effects of student assessment on learner affect were examined (Hypothesis 1.3). The results of MANCOVA suggested that neither student assessment type had positive effects on writing anxiety, while both assessment types affected learner autonomy of students positively with a large effect size. The adjusted correlations among writing anxiety and learner autonomy subscales obtained from the MANCOVA, showed similar patterns for both assessment types and also had similar patterns in both occasions (before and after the intensive writing sessions). Furthermore, it was found that there were positive correlations among the three measures of writing anxiety, while learner autonomy was not correlated with those measures of writing anxiety. Additionally, the students who had higher learner autonomy before the intensive student assessment sessions tended to have higher learner autonomy after the intensive student assessment sessions.

Chapter 5 qualitatively explores the reason why self- and peer assessment are comparable or different in terms of the effects of student assessment on writing performance and learner affect.

Chapter 5

QUALITATIVE STUDY

5.1 Introduction

This chapter aims to explore the effects of self-assessment and peer assessment on the development of writing ability and learner affect through qualitative analysis. The qualitative analysis comprises analyses of students' responses to an open-ended questionnaire item and students' and teachers' responses in semi-structured interviews that were employed to explore the efficacy and challenges of self-assessment and peer assessment.

The students in both groups of assessment types were asked to respond to the open-ended questionnaire items after intensive assessment sessions. After the sessions, 12 students and two teachers were interviewed: six students from each of the self-assessment and peer assessment groups, and JET B and NET C, who fully participated in the study sessions in class. Two students were selected from each of the three score levels on the writing tasks: low-scoring, average-scoring, and high-scoring groups. This is because involving students at different levels in this investigation could shed light on the effects of student assessment and reflect their unique viewpoints to complement the open-ended questionnaires. Furthermore, the interviews with the two teachers uncovered hidden perspectives on the effect of student assessment on the development of writing ability and learner affect.

In the first section of this chapter, learners' responses to the open-ended questionnaire items were analysed to survey the similarities and differences between self- and peer assessment using opinion mining (Qiu et al., 2011). In the second section, the transcribed semi-structured interview data were analysed to provide specific information, employing grounded theory. These results were then combined to examine the similarities and differences between self-assessment and peer assessment. The key results from these analyses are summarized in the third section.

5.2 Analysis of Open-Ended Questionnaire Responses about the Benefits and Challenges of Student Assessment

Open-ended questionnaire items asked students about the effective points and challenges of student assessment. Students' open-ended questionnaire responses were analysed quantitatively to determine their perception of the benefits and challenges of student assessment based on the frequency of words that appeared in their responses, as shown in Table 5.1. The total number of words in students' open-ended questionnaire responses, in Japanese, was 20,029 characters for the self-assessment group and 18,019 characters for the peer assessment group.

Table 5.1

The Frequency of Words that Appeared in the Open-Ended Responses

Ranking	Self-assessment group			Peer assessment group		
	Translation of open-ended responses	Frequency of occurrence	Number of participants involved	Translation of open-ended responses	Frequency of occurrence	Number of participants involved
1	I reflected (hansei) on what I wrote.	50	45	I could learn from peers' writing.	66	65
2	I found I had a limited vocabulary size and knowledge of grammar.	46	37	Peer assessment helped me to improve my ability in English writing.	34	17
3	I don't have the confidence to do self-assessment.	42	42	I admired and praised my peers' writing.	26	26
4	I am satisfied with the content of my writing.	35	20	I enjoyed reading my peers' writing/peer assessment was fun.	25	24
5	I should make efforts to study vocabulary and grammar much more.	31	24	I realized the importance of understanding others' opinions and ideas.	25	24
6	I could accomplish the task following the rubric of self-assessment.	30	25	I was very serious about reading and understanding peers' writing.	24	24
7	Self-assessment motivated me to study English.	20	10	It is very difficult to evaluate peers' writing because of a lack of English proficiency.	20	20
8	I tried to write a composition using adequate words.	18	10	My awareness of my weak points in English writing was not changed even after peer assessment.	19	10
9	Self-assessment taught me the importance of self-reflection.	18	18	I made efforts to write a composition because my composition was evaluated by my peers.	19	19
10	I reflected on the lack of coherence in my composition.	14	13	The time restriction was too severe to assess peers' writing.	10	5
SUM		304	244		268	234

Note. Translated from Japanese into English by the author.

As can be seen in Table 5.1, the three highest-ranking items in the questionnaires of the self-assessment group students were generally related to self-reflection, a lack of knowledge about vocabulary and grammar, and a lack of confidence in self-assessment. These three categories accounted for 45.4 % (138 out of 304) of all the comments made by the students in the self-assessment group (n = 244). On the other hand, the results of the peer assessment group students' responses, learning from peers, the positive effect on the improvement of English writing, and admiration for peers' writing, accounted for 47.0 % (126 out of 268) of the responses on peer assessment made by the peer assessment group students (n = 234).

These results suggest that about half (50.8 %) of the students in the self-assessment group referred to self-reflection, self-awareness of having limited vocabulary and grammar, and a lack of confidence. This might be interpreted as indicating that about half of the students in the self-assessment group paid attention to the extent to which they could attain their goals but became aware of the gap between the present and the ideal level. As another salient feature for the self-assessment group, conflicting views were found to coexist in the data. In detail, the item that was ranked third, lack of confidence in self-assessment, accounted for 13.8 % (42 out of 304) of all the responses made by the self-assessment group. The fourth-ranked item, self-satisfaction with achievements in English writing, accounted for 11.5 % (35 out of 304) of all the responses made by the 244 students in the self-assessment group. These two inconsistent comments were also sometimes found to be reported by the same student. This could be interpreted as meaning that such conflicting items, that is, high self-esteem as opposed to a sense of inferiority, might coexist in the emotions of the students in the self-assessment group. To sum up, it is apparent that the students in the self-assessment group reviewed and judged their writing following the assessment criteria, in particular language use and writing content. Self-assessment seemed to make the students aware of their present attainment in the study process. At the same time, this self-reflection might have prompted some students to feel two inconsistent emotions: self-esteem and a sense of inferiority.

On the other hand, as Table 5.1 shows, the three highest-ranking items in the questionnaires of the peer assessment group students were generally related to the good effects of the students' assessment experience on their English writing and the respect for peers' writing. About half of the students in the peer assessment group mentioned positive attitudes towards peer assessment, suggesting that peer assessment could give students new knowledge and help them to develop English writing.

Additionally, three tendencies in Table 5.1 differed between the self-assessment and the peer assessment group. First, the students in the peer assessment group expressed favourable attitudes towards peer assessment, as the fourth-ranking item indicates their enjoyment of reading their peers' writing (25 out of 268 responses, or 9.3 %; 24 out of 234 students in the peer assessment group, or 10.3 %). On the other hand, the students in the self-assessment group did not report such enjoyment. Second, it is noteworthy that the item related to the effect of peer assessment on the development of writing ability ranked second in the peer assessment group, accounting for 12.9 % (34 out of 268) of all the responses on self-assessment made by 7.3 % of the 234 students in the peer assessment group. In the self-assessment group, those responses were not found among the 10 items that were ranked most highly in Table 5.1. Third, the item related to the difficulty of assessment ranked third in the self-assessment group, accounting for 17.2 % (42 out of 244). However, only seventh in the peer assessment group, accounting for 7.5 % (20 out of 268).

To sum up, the students in the peer assessment group apparently had positive attitudes towards peer assessment and believed in its potential for improving their writing ability. In contrast, the students in the self-assessment group tended to pay special attention to features such as vocabulary and grammar when reviewing their writing. Furthermore, the open-ended questionnaire responses identified differences in students' perspectives between self-assessment and peer assessment. However, it was not evident why a common item related to a lack of confidence showed such contrastive rankings between the groups. Therefore, it is necessary to seek viable reasons behind those commonalities and differences emerging from each assessment type. Semi-structured interviews with teachers and students, to be presented below, were conducted to clarify the relationship between each assessment type and the effect on students' perception of student assessment.

5.3 Semi-structured Interviews with Students and Teachers

The semi-structured interviews mainly focused on the effects of the student assessment type on the change in writing ability and learner affect. A total of 12 students, two each from the high scoring, average-scoring, and low-scoring groups from both assessment types, and two teachers, JET B and NET C, participated in the interviews. As detailed below, in general, the participants supported the positive effects of the two types of student assessments on the development of their writing ability, yet each student assessment group presented different perspectives on learner affect. These different perspectives were supported by the two teacher interviewees.

5.3.1 Theoretical Codes for the Self-assessment Group

The initial coding for the self-assessment group resulted in 52 open codes, which were organized into 16 axial codes. They were merged into 10 categories, which were then grouped into four theoretical codes (emergent themes), conceptualizing the relationship between categories and codes (Table 5.2) and enabling the researchers to develop and organize them into concepts involving theoretical sampling. According to Charmaz (2006), theoretical sampling means the process of fitting emergent theories.

An overview of the 10 initial categories for the self-assessment group is presented in Table 5.2. To extract theoretical codes (emergent themes), salient meaning units were identified and then themes that shared similar meanings were combined. Lastly, these themes were inductively coded, considering their implications and the relationships among the 10 categories. This resulted in four theoretical codes that emerged from the initial 10 categories: (A) consciousness of assessment criteria and content; (B) external presence; (C) learners' perception of English writing; and (D) affect caused by English writing and self-assessment. Table 5.2 presents the relationship between the four theoretical codes (the left-most column) and the 10 categories (the second column from the left). These are presented sequentially and followed by a conceptual code that describes self-assessment in this study (Figure 5.1). The findings from the student interviews concerning each of the four theoretical codes will be discussed below, in Sections 5.3.1.1 to 5.3.1.4, along with corresponding comments obtained from the teacher interviews, followed by a synthesis in Section 5.3.1.5.

Table 5.2

Theoretical Codes (Emergent Themes) and Major Categories from the Self-assessment Group Students

Theoretical codes	Categories	Descriptions	Select code members	Example quotes
A.Consciousness of criteria and content	A1. Assessment of language use	●Evaluate the size and variety, accuracy of vocabulary, and accuracy of grammar and language use through self-assessment	●Vocabulary ●Grammar ●Accuracy of language use and punctuation	●I could use new vocabularies which I have just learned, so I am happy. (L) ●I could use grammar within the scope of my knowledge. (M) ●If I could not find a good word fit, I could write it in another way. (M) ●I tried to be careful about the spelling of words and accuracy of grammar. (H) ●I should have studied grammar and vocabulary to describe things more freely. (L)
	A2. Assessment of structure and coherence	●Interest in the structure of composition ●Whether writing is coherent	●Lack of coherence ●Well organized or not ●Linkage using conjunctions	●I worried about whether I can write coherent composition. (H) ●I was nervous about whether sentences are linked in a natural way to make myself understood. (M) ●I am wondering if my composition is disorganized. (L)
	A3. What I wrote	●Reflect on the writing content or attitude towards writing in English	●Reflect on personal experiences/memories ●The extent to which I could describe the topic ●Writing content is interesting or not	●I could experience what I enjoyed before a second time by writing a composition. (M) ●I feel like I am travelling because I wrote about France in my composition. (M) ●I can think well about my future by writing a composition. (H) ●I am interested in whether I could express what I want to convey. (H) ●I wrote a large amount for the first time. (L)
B.External presence (teachers and readers)	B1. Aware of readers' presence	●Becoming conscious of whether their writing is intelligible	●Interest in developing objectivity ●Interest in whether my writing is easy to read	●I should have written compositions more objectively to make my writing easier to understand. (M) ●I am not sure whether my writing is reader friendly. (L)
	B2. Necessity	●Absence of teachers	●Teacher assessment is needed	●The teacher should give us an evaluation and feedback instead of

	of teacher assessment	making students feel the necessity of a teacher	for reliable assessment	self-assessment because students cannot evaluate accurately. (M) ●Self-assessment is not reliable because of students' English ability. (H)
C.Learners' perception of English writing	C1. Attitudes towards English writing	●Changed attitudes towards English writing	●Becoming reflective and objective ●Habits to revise ●Strategy to develop writing skills	●I could develop a habit to revise my writing. (H) ●I am trying to write more objectively. For instance, I give examples. (M) ●I thought a lot about how to improve my writing skills after the session. (H)
	C2. Effect on mindset	●Intensive writing and self-assessment sessions had some effect on students' mindset	●English writing is a valuable experience ●Enjoy writing in English ●Do not want to write in English ●Nervous about writing in English	●It is a good experience to write in English, express my feelings, and evaluate my composition myself. (H) ●Self-assessment is a meaningful activity because self-assessment helped me to improve my English ability. (H) ●I enjoyed writing in English. (H) ●I was nervous about writing in English because I was confused about what I should do. (M)
D.Affect caused by English writing and self-assessment	D1. High self-assessment	●High self-evaluation because they assessed their effort as completing the English writing task	●Self-satisfaction/accomplishment about their own efforts ●High evaluation especially about task fulfilment	●I could complete the task, such as achieving a minimum number of words, writing about a topic, so I felt good. (M) ●I could write about the task to the end. (L) ●I could use my knowledge about English, so I wanted to evaluate my writing highly. (H) ●I could express what I wanted to say, so it is good. (M)
	D2. Motivation	●Develop motivation to study English and change attitudes towards English study	●Motivated to study English ●Aware of shortcomings ●Change attitudes towards English study and English writing ●How to study English ●Useful for entrance examination ●Negative to write in English	●I am highly motivated to study English. (H) ●I am aware of my weak points, such as grammar. (L) ●I rethought how to study English. (H)
	D3. Lack of confidence	●Finding that I am not a good writer ●Difficult to assess my	●Lack of confidence to carry out assessment because of limited English ability	●I am not confident in evaluating my composition because I do not have a strong ability in English. (M) ●Even if I assess my composition myself, my assessment is not

	composition because of lack of confidence	●Doubtful of the reliability of self-assessment ●Being nervous while writing	reliable. (L) ●I cannot complete English composition alone. (M) ●Even if I write or assess compositions, I feel negative about studying English because I am not good at English. (L)

Notes. H = high; M =average, L = low

5.3.1.1 Consciousness of Assessment Criteria (Theoretical Code A)

One of the codes that appeared most frequently in the interviews with the students in the self-assessment group was Theoretical Code A (consciousness of assessment criteria). To be specific, the students in the self-assessment group became more conscious of vocabulary, grammar, structure, and coherence (Category A1, assessment of language use). In particular, all of the six student interviewees from the self-assessment group indicated the importance of accurate usage of vocabulary and grammar. As an example quote displayed in Table 5.2 shows, students "tried to be careful about the spelling of words and accuracy of grammar" (SHSS1, high-scoring group). As another example in Table 5.2, when they could not produce or remember an adequate word, they "tried to paraphrase what they wanted to write in other words" (SMSS1, average-scoring group). Moreover, all of the six student interviewees reported that they had made efforts to achieve satisfactory scores for the vocabulary and grammar assessment criteria. To be specific, as displayed in Table 5.2, a low-scoring student in the self-assessment group (SLSS2) mentioned, "I could use new vocabularies which I have just learned, so I was happy". They also consciously evaluated the extent to which they could use grammar that they had learned in class. Similarly, an average-scoring student in the self-assessment group (SMSS1) commented, "I could use grammar within the scope of my knowledge". All these students commented that the assessment criteria made them conscious of the usage of vocabulary and grammar. It was a salient common feature among the six student interviewees in the self-assessment group.

Among the six student interviewees, five students commented that self-assessment helped them to develop a habit of revising or checking for errors. Self-assessment can be considered to have encouraged the students to form the habit of reflecting on their writing.

For instance, a high-scoring student in the self-assessment group (SHSS1) described self-assessment as having motivated him to proofreading what he wrote following the assessment criteria:

> I acquired a habit to look at my writing once more to check the accuracy before finishing. Even after handing in my writing, I tried to check the accuracy of the way I used vocabulary and grammar using a dictionary or by asking my friend. (SHSS1)

Hence, the assessment criteria functioned as a trigger prompting students to review their grammatical errors and vocabulary usage.

As for Category A2 (structure and coherence of writing), three of the six student interviewees stated that they had read their work again after writing and considered whether their writing was consistent and well organized. For instance, SHSS2, who was a high-scoring student in the self-assessment group, described self-assessment as having made her think of the relationship between sentences, their logical connection in particular:

> I tried to link a sentence with the next sentence to be natural. I also tried to follow the basic structure which I learned in class before, for example, how to state reasons and conclude my writing. Self-assessment criteria made me reflect whether my writing should be organized. Also, it helped me follow the assessment criteria, but I felt difficulty in revising alone. (SHSS2)

Thus, SHSS2 recognized that the assessment criteria had prompted her to revise the structure and coherence of her writing. In addition, SHSS2 tried to organize her writing structure using conjunctions and discourse markers that she had learned in class. Hence, it was found that the student interviewees tended to use the assessment criteria as a tool to review or revise vocabulary, grammar, structure, and coherence. In other words, the assessment criteria drew the attention of the self-assessment group students to vocabulary, grammar, structure, and coherence.

In terms of Category A3 (what I wrote), all six of the student interviewees showed an interest in the uniqueness of their writing in terms of the content. As one of the typical responses, a low-scoring student in the self-assessment group (SLSS1) commented on the purpose of writing, in other words, what writers should convey through writing:

> I think that content is the most important, that is to say, how interesting or entertaining it is for readers. Even if it is not grammatically accurate, if the text is boring, writing is meaningless. (SLSS1)

This comment might symbolize the intrinsic purpose of writing. Similarly, a high-scoring student in the self-assessment group (SHSS2) supported the importance of what writers should express and how from a unique perspective:

> What I wrote must be interesting or create an impression for readers. After writing, I always think about whether it is unique or interesting. (SHSS2)

Not only these two students but also the four other students supported the importance of writing content in terms of uniqueness, but the item of writing content was not included in the assessment criteria. Therefore, the students can be considered to have become conscious of vocabulary, grammar, structure, and coherence by carrying out self-assessment, yet they were also interested in the uniqueness of writing content.

Hence, the students in the self-assessment group stressed that the assessment criteria worked as a device to make them think about their writing in terms of language use, structure, and coherence. In addition, writing content was emphasized by the students in the self-assessment group, though the uniqueness of writing content was not included in the assessment criteria.

In relation to Theoretical Code A (consciousness of criteria and content), the two Japanese and native English teacher interviewees provided two types of comments corresponding to the students' ideas about self-assessment in terms of consciousness of assessment criteria. First, as mentioned previously, all of the six student interviewees mentioned that they became more careful about the usage of vocabulary, grammar, structure, and coherence of writing following the assessment criteria (Category A1). Both JET B and NET C affirmed that students tended to review their language use, particularly after the intensive sessions of self-assessment. The two teachers also explained that the students in the self-assessment group paid more attention to language use than previously, suggesting that the assessment criteria encouraged them to reflect on their writing, especially their language use. These two teachers also found that students mostly made the best use of their knowledge to write well while heeding the assessment criteria. For instance, JET B personally captured the phrase from one middle-scoring student in the self-assessment group after school, "I could use new words which I have just learned in the previous class, so I marked the best score in the assessment sheet". JET B mentioned that the assessment criteria encouraged students to link the instruction in class with

English writing. JET B also commented that some students tended to evaluate their effort highly or show self-satisfaction rather than evaluating their present ability. Therefore, JET B doubted the reliability of self-assessment because it might be influenced by individual students' self-esteem. In contrast, JET B agreed on the positive effect of self-assessment on the development of students in their consciousness of assessment criteria.

Second, for the writing content (Category A3), both teacher interviewees reported that, to students, writing meant a creative activity to express their individual ideas and experiences. Therefore, both teachers thought that English writing was a self-expression activity for students as well as an English learning activity. As a typical comment, JET B described how students understood English writing as an opportunity to express themselves and expected the readers' enjoyment of the written content and their response to the author as follows:

> Learning how to write in English is important in English classes, but it sometimes involves a self-expression activity. In this case, writing becomes just a tool to express their experiences and ideas. Some students regard what they wrote as important. In other words, those students want to entertain or impress others. I understand that students want to receive comments from someone in terms of writing content as well as language use.

Thus, JET B stated that students expected to receive reactions to their writing content from someone, even though they were informed that, in the session, they would receive no feedback except self-assessment. Additionally, JET B stated that some students liked to express their ideas or experiences to enjoy the activity by themselves because writing gave them a chance to look back on their experiences. As Table 5.2 displays, an average-scoring student in the self-assessment group (SMSS2) commented, “I could experience what I enjoyed before a second time by writing a composition”. Another average-scoring student in the self-assessment group (SMSS1) mentioned, “I feel like I was travelling because I wrote about France in my composition”. This suggests that writing also gave them a new experience. Therefore, JET B concluded that some students recognized writing as an opportunity for communication or self-expression, so a writing class might not be just a language learning class for some students.

In summary, all of the six self-assessment interviewees indicated the effect on the consciousness of the assessment criteria, especially language use. Furthermore, all of the student interviewees mentioned the importance of writing content even if it was not included in the assessment criteria. Therefore, the focus on language use and the writing content were the most prominent issues commented on by the student interviewees. Concerning proofreading, in particular, five of the six student interviewees stated that they had developed a habit of rereading to check the accuracy of language use, coherence, and structure. The students’ views on the habit of proofreading and the importance of writing content above were supported by both of the teacher interviewees as well. The two teachers stated that self-assessment worked as a device to alter students’ state of consciousness, especially concerning vocabulary and grammar. Regarding writing content, the two teachers commented that students tried to perform a writing task to express their ideas and experiences and that students considered uniqueness and individuality to be important, though the uniqueness of writing content was not assessed in the criteria.

5.3.1.2 External Presence (Theoretical Code B). Theoretical Code B (external presence) emerged as one of the theoretical codes for the self-assessment group. Half of the student interviewees in the self-assessment group commented on two points related to external presence: (1) readers' presence (Category B1); and (2) teacher assessment (Category B2). With respect to Category B1 (readers' presence), as Table 5.2 shows, an average-scoring student in the self-assessment group (SMSS1) commented on intelligibility, "I should have written compositions more objectively to make my writing easier to understand". In addition, a low-scoring student in the self-assessment group (SLSS2) mentioned her worries, "I am not sure whether my writing is reader friendly". Hence, the students appeared to be aware of objectivity and reader consciousness. As noted above, the self-assessment group students' writing was not expected to be read by peers or teachers, so the absence of readers seemed to illuminate the presence of imaginary readers for self-assessment group students.

In addition, it was found that the teachers' presence (Category B2) seemed to be essential for the mentality of the students. In short, self-assessment made the importance of the external reader evident by its absence because some students felt doubtful about the reliability of self-assessment without a reader other than the writer and the teacher. It was also found that students usually depended on teachers. The importance of teachers' presence was indicated by three student interviewees, who reported that they felt helpless when they faced difficulty in solving a problem. An average-scoring student in the self-assessment group (SMSS2) noted her anxiety because of the absence of the teacher, who plays the role of the assessor, and a lack of confidence in assessment because of her lack of English ability:

> Assessment should be conducted by teachers, because I do not have confidence in my English ability, so it is very difficult for me to judge whether my English is correct. (SMSS2)

Similarly, three of the six student interviewees indicated the absence of readers under the condition of self-assessment. They worried about whether the accomplishment of their writing pleased themselves but not others. They recognized the effectiveness of self-assessment, but the absence of readers made them feel stressed. One of those students, a low-scoring student in the self-assessment group (SLSS2), insisted on the importance of readers because she considered that readers' specific comments could elucidate the intelligibility of her writing:

> I think that my composition should be reader friendly, but I am not sure whether my writing is intelligible for others. I cannot imagine the intelligibility alone. (SLSS2)

Thus, half of the student interviewees indicated the effect of readers' absence on the students' mindset as writers.

Concerning the external presence discussed above from the perspective of the student participants, one teacher also observed students' anxiety about the absence of external readers or assessors. JET B thought that it was natural for students to hope to be given feedback from others. She also commented that students would be motivated to study and improve their objectivity with external readers or assessors because the presence of readers would encourage them to become more reader-friendly writers.

In summary, Theoretical Code B (external presence), referring to readers (Category

B1) and teachers (Category B2), became prominent for some students because of the lack of a third party since the intensive self-assessment session did not include teacher assessment or peer readers. In particular, three of the six student interviewees' responses signified the necessity of teacher assessment to give them reliable feedback and enable them to gain new knowledge. Two teacher interviewees also indicated the effectiveness of the presence of readers because it helped students to develop their objectivity as writers and created a feeling of motivation. Therefore, external presence was illuminated by the lack of readers in the self-assessment group.

5.3.1.3 Learners' Perception of English Writing (Theoretical Code C).

Theoretical Code C (learners' perception of English writing) became apparent from the comments made by student interviewees in the self-assessment group. Theoretical Code C (learners' perception of English writing) comprised two categories: (1) attitudes towards English writing (Category C1); and (2) effect on mindset (Category C2). All of the six student interviewees in the self-assessment group recognized the positive effect of their self-assessment experience on their affect and attitudes towards English writing (Category C1), most of the codes showing students' positive sentiment towards writing and self-assessment. For instance, a high-scoring student in the self-assessment group (SHSS1) indicated that self-assessment stimulated his motivation to study English because it helped him to understand the difference between the study goal and his present ability:

> I was highly motivated to learn English after the session because I found my weak points, so I try to fill the gap between my final goal and the present level. I could

> reflect myself while assessing my composition by myself. It encouraged me to study much more. (SHSS1)

On the other hand, two of the six student interviewees reported that they were almost lost in the direction of the study goal (Category C2). For instance, an average-scoring student in the self-assessment group (SMSS2) confessed her disorientation due to the lack of guidance or leadership from others.

> I felt nervous about self-assessment because I was lost about what I should do. With someone's help, I could have written and evaluated my writing confidently. (SMSS2)

SMSS2 indicated the uncertainty of self-assessment and insisted on the necessity of reliable support for her study to develop her English ability. Because of such a lack of certainty in self-assessment, she did not have confidence in the development of her English ability.

Hence, all of the six student interviewees agreed on the positive effects of self-assessment on English writing, while two of them indicated negative effects on learners' perception of English writing. Meanwhile, the teacher interviews recognized the positive and negative effects of self-assessment on learners' perception of English writing. Referring to the positive effects on learners' perception, the two teachers similarly noted that students acquired reflective attitudes towards writing (Category C1). JET B reported that such self-reflection led her students to develop a habit of self-revision and identification of an appropriate writing strategy. NET C found that most students enjoyed writing their ideas and experiences in English and that some students considered English writing as a valuable experience (Category C2).

At the same time, regarding the negative effects on learners' perception, both teachers reported some students' unmotivated attitudes towards English writing (Category C1). As the sessions were intensively repeated, such negativity seemed to be amplified for those students, without any solutions being found (Category C2). Compared with the six student interviewees, the two teachers supported the self-assessment procedure's positive effects on learners' perception, but they found that the negative effects on learners' perception were intensified by engaging in the assessment activity alone, without any external help. To prevent such a negative situation, they insisted on the necessity of teachers' help or guidance to satisfy students' needs.

In summary, the six students in the self-assessment group as well as the two teacher interviewees supported the positive effects of self-assessment on Theoretical Code C (learners' perception), but some students felt anxious about what they should have done. Those students preferred to be instructed directly by peers or teachers, but this session did not allow them to depend on others. The teachers stated that those students' anxiety would be connected with demotivation towards English writing and that teachers need to help those unmotivated students. In short, the student and teacher interviewees acknowledged that self-assessment had positive effects on learners' perception, yet negative effects on learners' perception were also indicated in the views of student and teacher interviewees.

5.3.1.4 Affect Caused by English Writing and Self-assessment (Theoretical Code D). As for Theoretical Code D (affect caused by English writing and self-assessment), both positive and negative affects became apparent from the interviews with the students in the self-assessment group. Two types of positive affect (Categories D1 and D2) and one type of negative affect (Category D3) were reported. With respect to Category D1 (high self-assessment), positive feelings towards English writing and self-assessment were reported by four of the six student interviewees in the self-assessment group (Category D1). In particular, they presented self-attainment or self-satisfaction after completing the session and explained that they had tried their best to describe what they thought in spite of their limited knowledge about language use. An average-scoring student in the self-assessment group (SMSS1) described a sense of self-attainment. He made an effort even though he considered English writing to be a weak point. To be specific, self-assessment allowed SMSS1 to feel a sense of accomplishment and adopt a positive attitude towards English writing as follows:

> It was very difficult to write my idea in English, because I am not so good at English. However, I know how seriously I set about my task. The rubric did not have a component to evaluate to what extent I made efforts to write. However, I completed a task, so I felt good. Therefore, I could feel positive towards English writing now. (SMSS1)

Four of the six student interviewees in the self-assessment group indicated the development of motivation to study English (Category D2). For instance, SHSS2 commented on her heightened motivation because of self-assessment. In addition, she “reconsidered how to study English”.

However, one of the six student interviewees, a low-scoring student in the self-assessment group (SLSS2), mentioned that self-assessment could not fill a gap between her current level and the target level (Category D3), so she reported irritation after the self-assessment experience. To be specific, self-assessment made her confused about the way to study English. She described a negative effect of self-assessment experience on her attitude towards English writing. Self-assessment might function to illuminate a lack of English ability and demotivate her to write in English, as shown in the excerpt below:

> I tried to write in English and also assessed my writing; however, my poorness in English did not change at all. Anyhow I found myself still to be poor at English. So, I do not feel good towards English writing. (SLSS2)

Meanwhile, the two teachers also reported that such positive and negative affect was caused by English writing and self-assessment. For instance, NET C mentioned the different effects of self-assessment on individual students. According to her observation, high-scoring students tended to accept self-assessment and English writing positively (Categories D1 and D2), while low-scoring students tended to feel inferior to others by engaging in self-assessment (Category D3). JET B agreed with this observation. JET B explained that low-scoring students tended to have negative feelings towards self-assessment (Category D3) because it encouraged them to reflect on themselves. According to JET B, this sometimes revealed a lack of knowledge and effort, which in turn made low-scoring students feel inferior to other students or notice a gap between their present level and a study goal (Category D3).

In summary, four of the six student interviewees showed self-satisfaction with or self-attainment in their English writing and heightened motivation (Categories D1 and D2), but one student presented stress between the gap and the study goal after the self-assessment experience (Category D3). Those results were confirmed by the two teacher interviewees. The teachers also supported students' awareness of a difference between the study goal and their current level.

5.3.1.5 Synthesis: Development of Self-assessment Group Students as Writers.

Following Corbin and Strauss (2015), during the coding process, notes were taken to depict statements that indicate how the codes (themes) were related to one another. Diagramming was used to depict schematic representations of the emerging ideas and linkages between codes (themes).

As shown in Figure 5.1, the relationships between the four theoretical codes (emergent themes), (A) consciousness of assessment criteria and content, (B) external presence, (C) learners' perception of English writing, and (D) affect caused by English writing and self-assessment, were analysed in terms of norms, status, and hierarchy and then coalesced into one core category. As a result, one overarching theme emerged from these four theoretical codes (emergent themes): development as a writer.

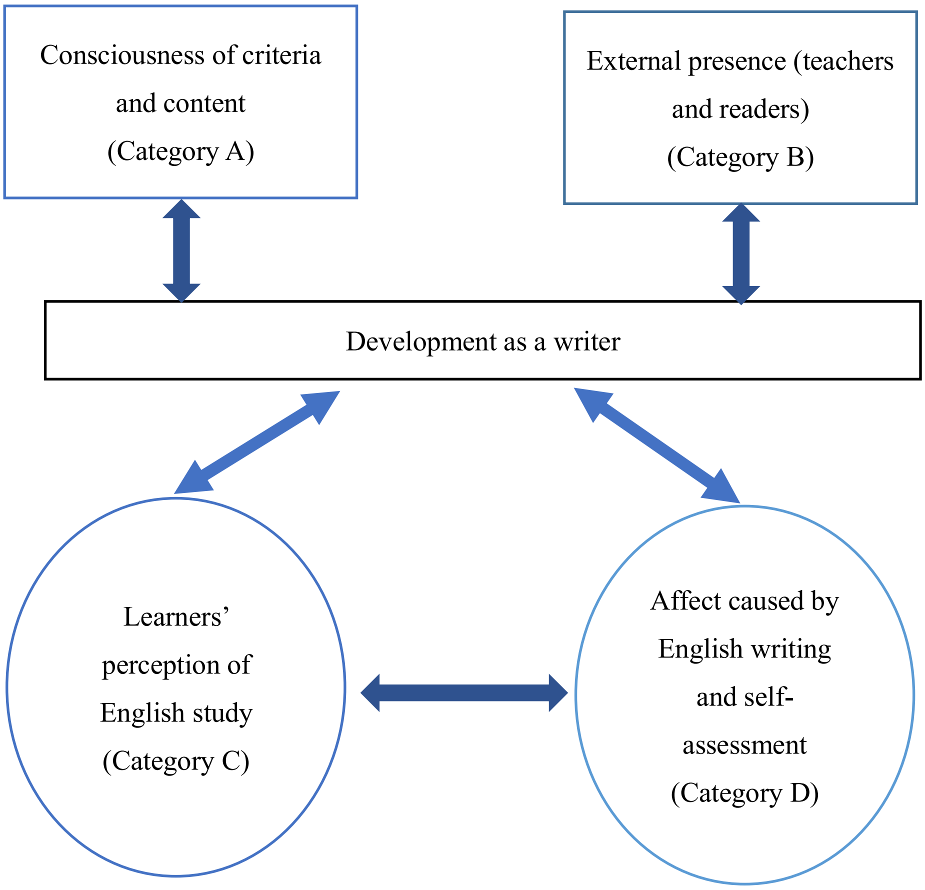

Consciousness of criteria and content (Category A)
External presence (teachers and readers) (Category B)
Development as a writer
Learners' perception of English study (Category C)
Affect caused by English writing and self-assessment (Category D)

Figure 5. 1 A schematic representation of Learners' Development as a Writer through self-assessment

Underlying all the comments that the students and teachers made about how students could improve their writing ability using self-assessment is the theme of their development as writers. As previously noted, students were aware of the necessity to increase their vocabulary and improve their knowledge of grammar to describe their ideas following the criteria of the writing assessment. Moreover, the consciousness of an external presence, such as teachers and readers, seemed to grow during the intensive self-assessment sessions because students wanted to receive feedback as writers on whether their writing was easy to understand or how their writing would be accepted by readers. Those matters encouraged the students to develop as writers. Both criteria, namely external presence and consciousness of criteria and content, will also continue to be influential for all of the writers even until they become professional writers. Therefore, the arrows in Figure 5.1 point in both directions because students' continual development as writers is supposed to interact with being aware of an external presence and the criteria and content. Two kinds of emotional factors, affective factors caused by English writing and self-assessment and effects on learners' perception of English writing, also had an effect on the learners' development as writers. These emotional factors and the development as a writer bidirectionally influenced each other in terms of both positive and negative states of mind regarding writing.

5.3.2 Theoretical Codes of the Peer Assessment Group

Regarding the peer assessment group, 38 open codes were obtained through initial coding, and they were grouped into 22 axial codes. These were organized into eight theoretical codes (emergent themes). An overview of the eight categories for the peer assessment group is presented in Table 5.3. To analyse the relationship between the eight categories, the conceptual clarity and the key elements of each category were coded. Subsequently, theoretical codes (emergent themes) were determined, checking the relationships between them. The relationships between the codes are presented in Table 5.3.

Three theoretical codes emerged from the eight categories as follows: (E) learning effect from peers, (F) development of metacognition, and (G) effects on learners' perception of English study. The theoretical codes of the peer assessment group reflected the presence of peers in terms of effects on learning and perception of study. Compared with the theoretical codes of the self-assessment group, the pattern of theoretical codes seemed to coalesce more simply into one core category – development as a writer – which is the same as for the self-assessment group. As another contrastive feature, the student interviewees in the peer assessment group generally reported that they enjoyed reading and assessing peers' writing. Both teacher interviewees also commented on the cheerful atmosphere produced by peer assessment, while the students in the self-assessment group rarely mentioned words such as "fun" and "enjoy". The major categories are presented sequentially, followed by a schematic representation of learners' development as a writer through peer assessment in this study (Figure 5.2). The three theoretical codes that emerged from the analysis of student and interview data will be discussed in Sections 5.3.2.1 to 5.3.2.3. In these sections, two teachers' comments extracted from the teacher interviews will also be discussed in a complementary manner. In Section 5.3.2.4, the findings from the student and teacher interviews will be synthesized.

Table 5. 3

Theoretical Codes (Emergent Themes) and Major Categories from the Peer Assessment Group Students

Theoretical codes (Emergent themes)	Categories	Descriptions	Select code members	Example quotes
E.Learning effect from peers	E1. Learning from peers	●Positively try to learn from peers’ English composition ●Positively stimulated by peers’ attitude and composition	●Admiration for peers ●Respect for peers ●Surprise at peers’ work ●Compliment peers’ work ●Try to imitate peers’ attitude towards English	●Peers’ composition gave me hints on how to write in English. (H) ●I could copy peers’ skills and expressions from peers’ writing. (M) ●I could find my level in class from peer assessment. (M) ●I found that my friends did a good job. (M) ●I could compare my writing with my peers’ writing, and I could learn language use. (L)
	E2. Communication with peers	●Enjoy communicating with peers and discovering peers’ opinions and experiences through English writing	●Enjoy communication with peers ●Fun to read peers’ composition	●I found I valued my friend’s ideas and opinions, so I want to become good friends with him/her. (L) ●Peer assessment influenced my way of thinking. (H) ●I could find a new aspect from a peer’s work. (H) ●I felt sympathy for a peer’s composition and also discovered different ideas from others. (M) ●I thought that it is important to know others’ ideas. (M)
F.Development of metacognition	F1. Awareness of readers’ presence	●Peer assessment helps to improve objectivity as a writer and nurture the attitude as a reader	●Become conscious of objectivity ●Motivated to know others’ ideas	●By assessing others’ work, I found that my work was also judged, so I must be persuasive and objective as a writer. Peer assessment made me complete the task. (M)
	F2. Become reader friendly	●Tried to be more reader friendly	●Become aware of the presence of readers	●I found the necessity to make my writing easy to understand for readers. (H)
G.Effects on learners’ perception of English study	G1. Positivity about peer assessment	●Motivated to write in English and exchange English compositions with peers	●Want to continue to read peers’ composition ●Want to write in English with peers	●Thanks to peer assessment, I found my weaknesses and was motivated to study. (M) ●I am happy to receive assessment from peers and also feel good to give peers my feedback. (H)

	●Effective in improving writing skills		●Peers' opinions are very important because they encouraged me to study and find new aspects. (M) ●I felt happy because friends gave me compliments about my composition. (M) ●Peer assessment helped me to discover that everyone made similar mistakes. (H)
G2. Motivation to study English	●Heighten students' motivation to learn English	●Motivated to develop writing ability ●Made efforts to write better because my writing was evaluated by peers	●I felt motivated to learn English language use and other skills from classmates' compositions. (M)
G3. Difficulty of peer assessment	●Feeling unable to assess peers' composition because of limited English ability and relationship with peers	●Reliability of peer assessment ●Lack of English ability ●Feeling of inferiority ●Relationship between peers ●Shy to evaluate peers' work	●I do not have the ability to assess others because I am not good at English. (L) ●I do not want my composition to be read by my peers because I am shy. (M) ●It was very difficult to give peers feedback because I am not sure whether my assessment was good or reliable. (M)
G4. Negativity about peer assessment	●Peer assessment is not particularly reliable and does not work well	●Demotivated peers ●Doubtful about the reliability of peer assessment	●Some people did not give me a good assessment or feedback. (M) ●I was discouraged because some peers did not complete the task. In this case, I cannot give them a good evaluation. (M) ●Peer assessment does not help to improve English ability. (L)

Notes. H = high; M =average, L = low

5.3.2.1 Learning Effect from Peers (Theoretical Code E). Theoretical Code E (learning effect from peers) emerged most prominently from the interviews with the students in the peer assessment group. All six student interviewees in this group stated that they could learn from peers' writing, and five of the six student interviewees praised their peers' writing from the perspective of language use, structure, task fulfilment, and uniqueness of content (Category E1). All six student interviewees in this group commented that they greatly enjoyed reading their peers' compositions and felt pleased to know their peers' ideas and experiences through having an opportunity to become closer friends with their peers by reading their compositions (Category E2). All of the interviewees reported that they enjoyed reading and assessing peers' compositions (Category E2). It was the most prominent feature of the comments made by all the interviewees. A low-scoring student in the peer assessment group (PLSS1) described his delight in reading his peers' compositions because he could discover unknown aspects of them and feel familiarity with them. Hence, peer assessment might allow students to build rapport with peers rather than assessing their writing ability. PLSS1 commented on the joy of discovering his peers' idea by reading their writing as follows:

> I enjoyed reading a classmate's writing because I discovered his/her ideas and interests. Those people did not have any opportunities to have a chat with me, but I wanted to talk with them after the reading and peer assessment. (PLSS1)

Moreover, four of the six student interviewees mentioned the importance of sharing ideas with others because they found different opinions in others' writing (Category E2).

Both sharing others' ideas (Category E2) and learning from peers (Category E1) were mentioned by all six students. One student (PHSS2) indicated the effect of peer assessment on learning (Category E1) because they could learn how to write in English through peer reading. Consistent with PHSS2, another average-scoring student in the peer assessment group (PMSS1) mentioned that peer assessment enabled him to obtain new knowledge (Category E1). To be specific, he explained that he could learn from peers' writing how to use vocabulary and idioms to express ideas and those experiences motivated him to study all the more to catch up with his peers:

> Thanks to my peers' writing, I learned new words and usage. I learned how to organize writing. So, I wanted to mimic my peers' writing when I write the next composition. I was stimulated to study much more because my friend showed me a good job. (PMSS1)

This student's comment presented multiple roles of a student in the peer assessment group because a student in this group was expected to take on the roles of a friend, reader, assessor, and writer. In short, students read and assessed peers' compositions from many perspectives, so peer assessment had an effect on their development as writers.

The two teachers also agreed on the effect of peers on writing ability. In particular, JET B described her positive impression of peer assessment as a learning tool in class. To be specific, peer assessment could make the atmosphere of the writing class active because peer assessment encouraged students to interact with each other and enabled students to learn from their peers. JET B also found that students mostly found benefits in peer assessment, so such positive attitudes towards peer assessment made the class more harmonious:

> I also enjoyed observing students' cheerful mood in class because I could find out how they were enjoying reading others' writing. I assume that students considered peer assessment to be a communication tool between friends. They also seemed to understand that it was a tool to motivate them to improve their English skills because the assessment criteria listed the important items for English writing. (JET B)

To sum up, both student and teacher interviewees described the learning effect from peer assessment on students' mindset and English writing positively (Theoretical Code E). All of the student interviewees reported that they had learned from their peers' writing, such as writing skills and expressions. Five student interviewees praised their peers' writing (Category E1), while four student interviewees reported that they enjoyed sharing their ideas with their peers (Category E2). The two teachers also confirmed the learning effect from peers (Theoretical Code E), such as students' positive attitude towards learning English. Additionally, the teachers found that peer assessment worked as a communication or study tool in class.

5.3.2.2 Development of Metacognition (Theoretical Code F). Theoretical Code F (the development of metacognition) comprises two categories: (1) awareness of readers' presence (Category F1) and (2) becoming reader friendly (Category F2). For Category F1 (awareness of readers' presence), all six student interviewees stated that they had become aware of the presence of peers as readers and that they were conscious of this perception when carrying out peer assessments. An average-scoring student in the peer assessment group (PMSS1) described a change in his writing attitudes because of his consciousness of his peers when writing English. To be specific, he became more strongly aware of peer reading and peer assessment (Category F1), so he made efforts to write more convincingly, using examples, than before. It is also noteworthy that he evaluated his writing attitude as having changed substantially as a result of peer assessment. It appears that peer assessment made him conscious of readers' presence and motivated him to become a better writer for his readers:

> As I knew that a classmate would read my composition, I was careful about what I wrote. Especially, I tried to be objective in order to make my friends understand my ideas, so I made it a rule to present examples to help my friend understand my experience from my writing. I often reflected on myself and I thought such a reflection could be good for me. (PMSS1)

Consistent with the student above, another high-scoring student in the peer assessment group 2 (PHSS2) commented that her motivation was raised by peer assessment because she wanted to be praised by her peers (Category F1). Additionally, she realized that the presence of a peer assessor influenced her attitude towards writing. This consciousness was connected with the recognition of others and led to the objective evaluation of her own consciousness. In other words, it is considered that peer assessment worked to stimulate her metacognition (Theoretical Code F):

> Peer assessment always made me be conscious of readers. If it had not been for peer assessment, the content of my writing would have been different. I am not sure whether the peer readers or assessors were good, but it is a fact that I always thought

> about the presence of peers while writing and tried to gain compliments from others. And I also understood that I was thinking of the peers' assessment while writing. (PHSS2)

Thus, it is noticeable that peer reading and assessment made the students try to be clear and persuasive (Category F1) when they wrote compositions. This is because the students knew that their writing would be read and evaluated by their peers according to the assessment criteria, so their writing needed to be understandable and impressive for their peers. In short, peer assessment made the students think of the presence of peer assessors and made them conscious that their writing would be evaluated by peers in turn. In other words, peer assessment might provide students with multiple perspectives. Hence, some students recognized the metacognitive process while conducting peer assessments (Theoretical Code F).

In the teacher interviews, the two teacher interviewees also mentioned the improvement of metacognition of peer assessment group students. For instance, JET B evaluated the effect of peer assessment on the process of metacognition because she found that students changed the status of agents in class, but such turn taking did not occur intentionally. While writing or assessing, the students in the peer assessment group naturally changed their perception from assessors to writers, readers, or friends (Theoretical Code F). Such self-perception helped the students to monitor their learning attitude and find a strategy for writing and studying English.

> I found that students naturally took different roles, such as writers, readers, instructors, assessors, and friends, during peer assessment. I think that such an

> experience improved students' strategy of writing and knowledge about themselves. (JET B)

JET B stated that peer assessment gave the students an opportunity to look at themselves from different perspectives and helped them to improve their writing skills by enhancing their self-understanding in terms of study strategies, such as defining a study goal and setting priorities to attain it.

Hence, all six student interviewees mentioned the effect of peer assessment on the improvement of metacognition, such as knowledge about oneself as a learner and awareness of cognition (Theoretical Code F). For example, students became conscious of the presence of readers and reflected on their own writing as writers, and they valued the effect of peer assessment on the improvement of consciousness. The two teacher interviewees also agreed on the improvement of their students' metacognition. They reported that peer assessment improved students' cognitive knowledge and cognitive regulation; in other words, peer assessment helped students to understand the requirement of writing tasks and led students to self-directed thoughts while assessing peers' writing. Therefore, students' metacognition seemed to be stimulated more than before the session.

5.3.2.3 Effects on Learners' Perception of English Writing (Theoretical Code G). Positive and negative psychological effects on English writing were proposed by all of the student interviewees. Positive reactions to peer assessment (Categories G1 and G2) were more frequently observed than negative comments (Categories G3 and G4). To be specific, five of the six student interviewees presented positive comments on peer assessment. According to them, peer assessment motivated the students to learn English and stimulated the students to have positive attitudes towards reading and writing in English. In fact, two student interviewees (PMSS2 and PHSS1) reported the potential benefits to their performance in English examinations that they may take in the future due to their engagement in peer assessment. It is difficult to find evidence of such positive effects of peer assessment; however, the students positively explained that these successes were related to peer assessment. They tended to connect the effect of peer assessment with their aspiration for obtaining successful results in English scores in regular school tests or entrance examinations in the future. For example, a high-scoring student in peer assessment group 1 (PHSS1) commented on a positive effect of peer assessment on his perception of English study (Category G1) and reported that he was motivated to study all the more (Category G2). He also linked the positive effect of peer assessment to his future Recommendation Entry Exam (called the Admission Office exam in Japanese) as follows:

> Peer assessment had a positive effect on my study strategies and it worked well to pass the Recommendation Entry Exam. This is because it helped me to realize the importance of studying. Peer assessment would work well even in a workplace in my future. (PHSS1)

An average-scoring student in the peer assessment group (PMSS2) supported the idea that peer assessment stimulated her motivation (Category G2) because it encouraged her to study English, as Table 5.3 shows. Another average-scoring student in the peer assessment group (PMSS1) agreed with the idea that exchanging feedback between peers produced a good atmosphere in class (Category G1). Hence, positivity about peer assessment was reported by five of the six student interviewees.

On the other hand, two student interviewees made negative comments about peer assessment. They indicated the difficulty in reporting their honest opinions to their peers who did not show effort in their writing (Category G4). They were worried about whether their honest assessment would discourage or distress their peers. They also feared it could damage their relationship with each other in the classroom (Category G3). For example, an average-scoring student in the peer assessment group (PMSS2) indicated the difficulty of peer assessment when she had to assess an incomplete piece of writing because it made her think about how to give her peers an honest evaluation as follows:

> Peer assessment sometimes made me motivated to study and learn a lot from others. But some students did not complete their task as writers at all. In this case, I was not sure how to react to those people. (PMSS2)

Additionally, PMSS2 was doubtful about the reliability of peer assessment (Category G3) because it could be influenced by the relationship of the assessor with his or her peers. Specifically, PMSS2 wondered whether peers' compliments regarding her writing were honest or whether their true judgement was hidden.

A low-scoring student in the peer assessment group 2 (PLSS2) also mentioned doubts about the effectiveness of peer assessment from a different perspective (Category G4). She gave doubtful comments on the reliability of peer assessment because of the lack of her English ability to assess peers' writing (Category G3). She also felt pressurized by peer assessment because it was difficult for her to judge her peers' writing by herself as follows:

> I do not have the competence to judge others' writing because I am not good at English, so my assessment was not good: very simple and not reliable. I felt sorry for my peers, so I do not like peer assessment. I am not sure whether their writing is accurate. I cannot judge it by myself. (PLSS2)

Hence, depending on individual students' English proficiency and attitudes towards English, students presented different reactions to writing and peer assessment. However, even negative comments (Categories G3 and G4) were related to the development as a writer. A lack of confidence could be a negative affect (Category G3). At the same time, however, it was connected with the motivation to make efforts in the future (Category G2).

As shown above, the student interviewees presented contrastive ideas about peer assessment. Consistent with the student interviewees, the two teachers stated that peer assessment had different aspects for students, such as positivity and negativity, so they suggested that teachers need to be careful when implementing it in class. Specifically, the Japanese English teacher (JET B) acknowledged the usefulness and difficulty of peer assessment as follows, particularly concerning the relationship between students, the class atmosphere, and the matter of courtesy:

> Peer assessment is very useful to nurture students' objectivity and motivation because peer assessment makes students competitive. It helps students work better. However, I was very careful or sometimes nervous about the relationship between students. Who is a friend with whom? Who has a good relationship with whom?

> Isn’t it a problem to make pairs between them? Like that. Namely, peer assessment is influenced by class atmosphere and human relationships between students. Teachers should teach a rule before peer assessment to make it successful. (JET B)

Hence, teachers are expected to have a good understanding of human relationships between students because the success of peer assessment depends on the combination of students in the class. Moreover, teachers are encouraged to guide students to follow rules to make peer assessment fruitful. Compatible with the opinion of JET B, NET C also commented on the necessity of teacher guidance for peer assessment in terms of class courtesy, that is, politeness, as follows:

> Students tend to take peers’ comments and assessment very seriously. Some students did not understand how those comments hurt others’ feelings, so teachers should be very careful in the classroom. (NET C)

On the whole, many more student interviewees mentioned positive effects on learners’ perception of English study (Categories G1 and G2), but negative opinions, such as the unreliability of peer assessment and the possibility of damaging the relationships among peers, were offered (Categories G3 and G4). The teachers made positive and negative comments concerning Theoretical Code G (effects on learners’ perception of English study) similar to those of the student interviewees. The teachers also insisted on the necessity of teacher guidance in peer assessment in class in terms of human relationships and class courtesy to make peer assessment beneficial.

5.3.2.4 Synthesis: Development of Peer Assessment Group Students as Writers. The three theoretical codes (emergent themes), that is, (E) learning effect from peers, (F) development of metacognition, and (G) learners' perception of English writing, coalesce around a core category of norms, status, and hierarchy. To be specific, as was the case for the self-assessment group, one overarching theme seems to emerge from these theoretical codes (three emergent themes): development as a writer (Figure 5.2).

When the results for the peer assessment group are compared with those for the self-assessment group, on the one hand, one core category of peer assessment, development as a writer, was the same as that of the self-assessment group in the sense that every theoretical code was concentrated in it. On the other hand, different theoretical codes (emergent themes) and relationships were found between them. The results for the peer assessment group are presented schematically in Figure 5.2.

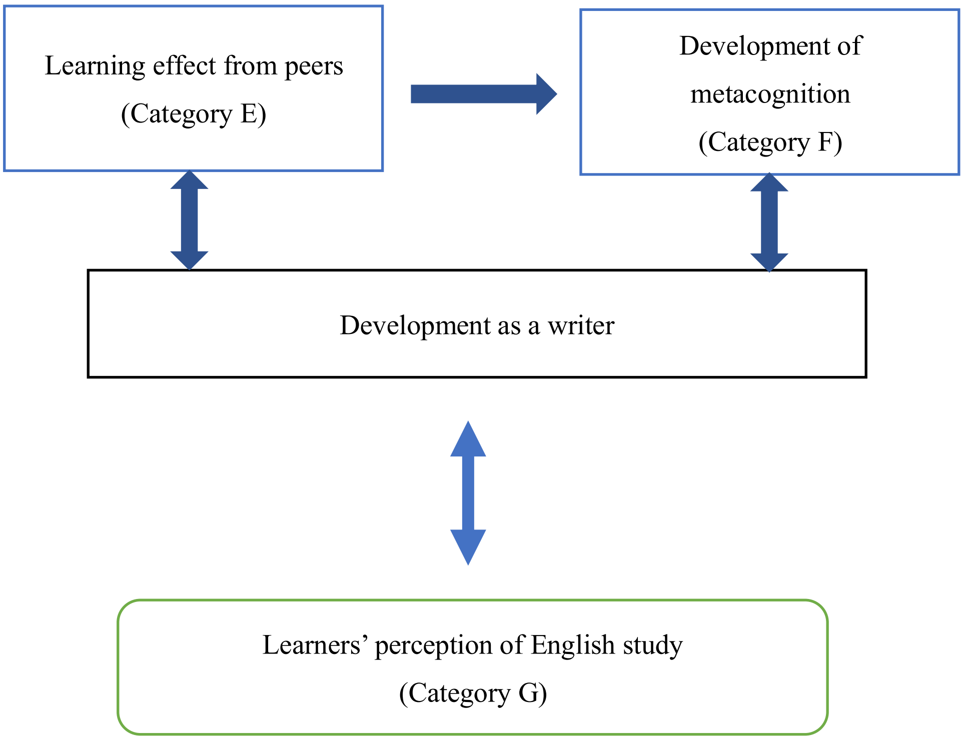


Figure 5. 2 A schematic representation of Learners' Development as a Writer through peer assessment

A primary difference between the peer and the self-assessment group concerned Category F (development of metacognition). Peer assessment group students tended to recognize themselves as a reader, assessor, friend, and writer. Their self-recognition was more diversified than that of the self-assessment group students. As Figure 5.2 shows, development as a writer is the central concept linking the three theoretical codes to one another: (E) learning effect from peers, (F) development of metacognition, and (G) learners' perception of English writing. In contrast with the self-assessment group, the most prominent feature of the peer assessment group is the effect of peers on writing performance and learner affect because the peer assessment group students interacted with their peers and learned from them. To be specific, learning from peers influenced and boosted students' metacognition as writers, and the interaction through peer assessment made students motivated to write more clearly and persuasively. Therefore, in Figure 5.2, the direction of the learning effect from peers (Category E) turns towards the development of metacognition (Category F). These two theoretical codes have double directions to the core code, development as a writer. This is because the transactions will continue as students develop themselves as writers. On the other hand, the effect of peer assessment on learners' perception of English writing (Category G) is located at the bottom of Figure 5.2 because it forms the basis of learners' affect. Learners' perception of English writing (Category G) is composed of two factors: positive and negative affect. These affective factors are closely related to individual students' human relationships with their classmates and personality. In other words, these factors are delicate and fragile as they have the possibility of being influenced by environmental conditions, so they could be personal. Compared with the self-assessment group, the students in the peer assessment group tended to express positive affect, such as enjoyment of peer assessment. The comments related to negative affect caused by peer assessment were made by three of the six student interviewees, while one of them also made positive comments on peer assessment. Moreover, such negative affective responses were found less frequently in the results of the open-ended questionnaire for this group (Table 5.3). Therefore, learners' perception of English study (Category G) comprehensively represents such positive and negative affect. In sum, the model of peer assessment presents the development as a writer as being located at the centre of the three theoretical codes.

5.3.3 A Comparison of the Semi-structured Interview Results between Self-assessment and Peer Assessment

As previously stated, self- and peer assessment group student interviewees commonly aimed at becoming good English writers, yet the make-up of the theoretical codes (emergent themes) differed between them. For readers' convenience, Figure 5.3 presents comparison of the models between the self- assessment and the peer assessment group side by side. Self-assessment and peer assessment have two commonalities and three differences in the construct of the theoretical codes (emergent themes).

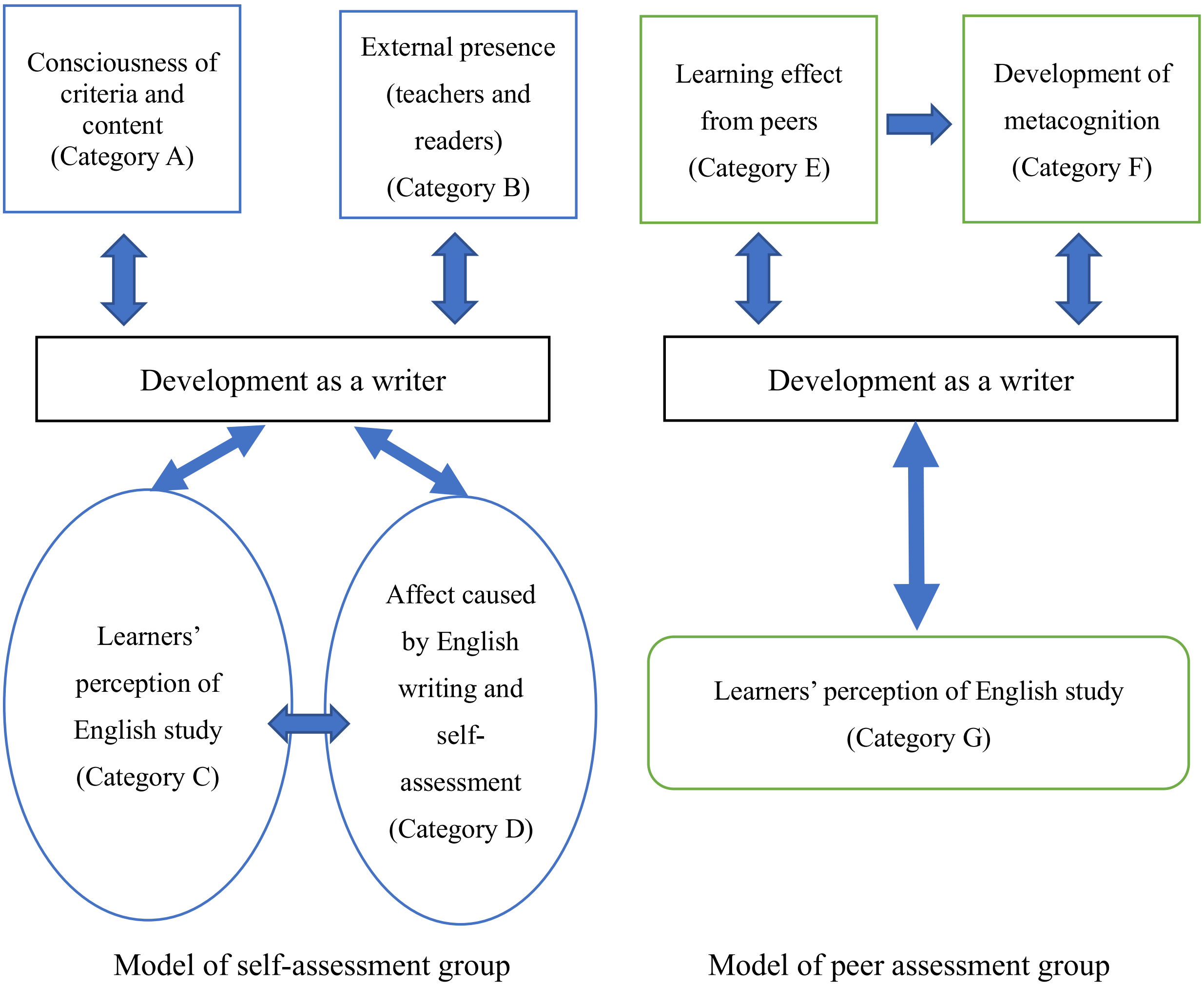

Consciousness of criteria and content (Category A)
External presence (teachers and readers) (Category B)
Learning effect from peers (Category E)
Development of metacognition (Category F)
Development as a writer
Development as a writer
Learners' perception of English study (Category C)
Affect caused by English writing and self-assessment (Category D)
Learners' perception of English study (Category G)
Model of self-assessment group
Model of peer assessment group

Figure 5.3 Comparison of models between the self-assessment group and the peer assessment group

Some notable commonalities and differences can be found between the two models presented in Figure 5.3. There are two commonalities. First, the subcategories in both assessment types converged in the same core category, that is, development as a writer. In other words, those subcategories all work towards achieving the same goal: the development of the learner as a writer. Second, one subcategory, that is, learners' perception of English writing (Category C and Category G), was identified in both student assessment types. To be specific, it presents the effect of the student assessment type on learner affect, such as motivation, writing attitude, and self-confidence. In brief, the two models are made up of partly different subcategories but they aim to achieve the same goal.

On the other hand, three differences were notable. Firstly, the numbers of constituents, or theoretical codes (emergent themes), are different from each other. The model of the self-assessment group is composed of four theoretical codes (emergent themes), while the model of the peer assessment group comprises three theoretical codes (emergent themes). Secondly, the effects of external presence, such as teachers, peers, and readers, on self-assessment group students was weaker than those on peer assessment group students. Self-assessment group students were more specifically conscious of the assessment criteria and writing content than peer assessment group students. In the peer assessment group, the learning effect from peers worked to improve metacognition. To be precise, self-assessment had a more specific impact on students' consciousness of the assessment criteria, while peer assessment had a more general effect on students' learning strategies. As O'Malley and Chamot (1990) stated, a learning strategy works as a cognitive process to increase understanding, learning, or memory. Therefore, it is considered that the students in the peer assessment group tried to gather and understand information from peers' writing. Such a cognitive process nurtured the learning strategy. Thirdly, the theoretical codes (emergent themes) related to learner affect differed between the two groups. Self-assessment group students tended to show positive affect towards self-assessment because it motivated them to study and feel self-attainment. However, a lack of confidence and doubt in the reliability of self-assessment were found in some of the comments made by student interviewees in the self-assessment group as well. On the other hand, students in the peer assessment group tended to note concrete benefits to their mindset of peer assessment. Compared with the challenges, the benefits were reported more frequently by the student interviewees in the peer assessment group, so learners' perception of English study (Category G) inclusively displays positive and negative affect in the model (Figure 5.3) of the peer assessment group. These student interviewees mostly supported the effectiveness of peer assessment in terms of cheerfulness, fun, learning, and communication with peers, but the difficulty and unreliability of peer assessment were partly indicated in the responses to the open-ended questionnaire and the comments from two student interviewees.

Thus, the comparison of the models between the self-assessment and the peer assessment group suggests that the two assessment types share the same core category in spite of their different make-up. The results of the teacher interviews concurred with those of the student interviews as well. The teacher interviewees supported the findings from the student interviews and added deeper insights into the effect of both assessment types on learning attitude and affect. Regarding self-assessment, both teacher interviewees agreed on the four categories of the self-assessment group (Figure 5.1) in terms of four points. First, both teacher interviewees recognized the positive effect of peer assessment on students' consciousness of assessment criteria, especially the usage of vocabulary and grammar. However, JET B was doubtful about the reliability of self-assessment because some students tended to evaluate their writing highly due to self-satisfaction or self-esteem. Second, both teacher interviewees indicated that no presence of readers or assessors highlighted external presence for the students in the self-assessment group. They added that some students apparently hoped to rely on teacher guidance. Third, they mentioned both positive and negative effects on learners' perception of English writing. For instance, as a positive effect, self-assessment enhanced self-reflection and self-revision. As a negative effect, intensive self-reflection might amplify a sense of inferiority or demotivation for students, so the teacher interviewees insisted on the necessity of adequate teacher support. Finally, the teacher interviewees mentioned that affect caused by English writing and self-assessment differed among students. For instance, self-assessment might boost some students' motivation to study English but cause others to lose confidence in English writing.

With respect to peer assessment, the teacher interviewees supported the comments from the student interviewees as well as giving statements from different perspectives in terms of three points. First, both teacher interviewees agreed on the positive learning effect from peers and recognized peer assessment as an effective tool to improve writing ability. Moreover, peer assessment encouraged students to interact with their peers and make a cheerful atmosphere in class, so the teacher interviewees valued peer assessment to activate their class. Second, both teacher interviewees also stated that peer assessment worked to develop students' metacognition because peer assessment provided students with multi-perspective views, such as those of writers, readers, assessors, learners, and friends. The teacher interviewees also reported that such turn changing naturally occurred in peer assessment sessions, so students could nurture metacognition through peer assessment. Finally, the teacher interviewees stated that the success of peer assessment might depend on peer combinations because peer assessment would be influenced by human relationships in class. In addition, assessments reflecting the honest opinions of students might not be carried out by peers so that students may avoid face-threatening assessment, so there is a possibility that some students might not accept the peer assessment as a true assessment. Therefore, the teacher interviewees stressed the necessity of teacher guidance to succeed in peer assessment.

In sum, the findings from the student interviews were reinforced by the reports from the teacher interviews because the comments made by the teacher interviewees revealed hidden aspects of each student assessment type. Self- and peer assessment have similar and different qualities but are directed at the same goal: development as a writer.

5.4 Summary of the Qualitative Analyses

This chapter examined the effects of self-assessment and peer assessment on writing performance and learner affect. The self-assessment and peer assessment groups were compared by means of an analysis of students' open-ended questionnaire responses and comments obtained from semi-structured interviews with students and teachers. The key results of these qualitative analyses can be summarized in terms of five points. First, the results of the analysis of open-ended questionnaire responses showed that the self-assessment group students tended to turn more intensive attention on themselves than the peer assessment group students. Stated differently, self-introspection or self-reflection (*hansei* in Japanese) on writing content and language use was intensively conducted by the self-assessment group. When the members of this group reflected on their writing, they referred more frequently to the specific assessment criteria, such as vocabulary, grammar, structure, and coherence, than the peer assessment group students did. A possible explanation for this is that self-assessment encouraged students to appraise their writing by themselves, which required them to be conscious of the assessment criteria and content. The teacher interviewees agreed that the students in the self-assessment group were intensively conscious of specific areas of the assessment criteria, that is, vocabulary and grammar. The teacher interviewees also commented about a positive effect of self-assessment on self-revision.

On the other hand, the open-ended questionnaire responses and semi-structured interviews showed that the peer assessment students were conscious of the presence of peers not only as assessors but also in other roles, like readers, writers, and friends. Through peer assessment, they seemed to develop metacognition by being aware of others instead of concentrating on themselves. The teacher interviewees also agreed that the development of metacognition occurred for the students in the peer assessment group because peer assessment made students take on multiple roles, such as assessors, readers, writers, and friends. Moreover, the teacher interviewees found that some of the students in the peer assessment group valued the effect of having different perspectives on their writing performance.

Second, while external presence emerged as a key issue in both groups, the students' comments differed considerably between the two assessment types. The self-assessment group students mentioned that the external presence of people such as teachers and imaginary readers made their writing become more reader friendly, while the peer assessment group students recognized the learning effect from peers who served as the readers of their writing. Namely, peer assessment enabled students to be practically aware of the presence of readers, while the self-assessment students were obliged to imagine their presence. On the one hand, in the case of the self-assessment group, two theoretical codes, external presence and consciousness of criteria and content, did not involve any mutual interaction between the students (Figure 5.1). On the one hand, the relationship between the two theoretical codes in the peer assessment group differed from that observed in the self-assessment group (Figure 5.2). To put it concretely, learning English through interaction with peers had an effect on the development of metacognition for the peer assessment group students.

The teacher interviewees supported the comments from the student interviewees on the external presence and indicated the necessity of teacher support for student assessment in class. With respect to self-assessment, demotivated or low-scoring students might perceive self-assessment negatively because those students tended to rely on external support, such as support from peers and teachers. Regarding peer assessment, its success might be influenced by peer relationships, individual students' courtesy, and motivation. Therefore, the teacher interviewees insisted on the importance of teacher guidance for making peer assessment more meaningful. As shown in the different constructions of Figure 5.1 and Figure 5.2, both assessment types had some effect on the development as a writer, but learners' awareness of the assessment criteria and external presence as readers were influenced differently by the student assessment types, namely the self-assessment or peer assessment methods.

Third, the results suggest that self-assessment had a different effect on the students' development as a writer in terms of affective conditions. As a comparison of Figures 5.1 and 5.2 demonstrates, the emotional condition of the self-assessment group consists of two theoretical codes: learners' perception of English study (Category C) and affect caused by English writing and self-assessment (Category D). In contrast, the peer assessment group showed only learners' perception of English study (Category G). The comments from the teacher interviewees offer an explanation for the commonality and difference between the groups. Regarding the commonality, both assessment types might develop motivation for English writing and make students feel unconfident in their English ability and doubt the reliability of student assessment. However, according to the open-ended questionnaire responses (Table 5.2), the students in the self-assessment group evaluated their efforts and attainment for the study highly while the students in the peer assessment did not show such high self-esteem. A possible explanation for this is that self-assessment made the students focus more intensely on their writing attitude and achievement of the goal of the study than the students in the peer assessment group.

Fourth, both the self-assessment and the peer assessment group students commented on a lack of confidence as an assessor because of their lack of English proficiency. Though this statement was common to both assessment types, the self-assessment group students seemed to have a stronger sense of it than the peer assessment group students, as shown by Table 5.3, in which a lack of confidence ranks third in the frequency of occurrence in open-ended survey item responses for the self-assessment group while ranking seventh for the peer assessment group. Furthermore, the members of the peer assessment group mentioned that they enjoyed reading and assessing their peers' compositions while admitting a lack of confidence in evaluating others' writing at the same time. To put it differently, the self-assessment group students seemed to become much more tense or worried about English writing and the assessment of compositions than the peer assessment group students did. Regarding the peer assessment group, the affective condition of the peer assessment group students seemed to coexist with the interaction with peers or enjoyment as well as worries caused by a lack of English proficiency.

With respect to the lack of confidence, the comments from the teacher interviewees elucidated a difference between the two groups. To be specific, the students in the peer assessment group mostly received compliments from their peers, and these peers' good comments encouraged the students to study. However, self-assessment might intensify the level of self-reflection, which might negatively influence learners' self-confidence in English ability and English writing. In brief, the results of both the open-ended responses and the semi-structured interviews in the self-assessment and peer assessment groups included a lack of confidence as a category (Table 5.1 and 5.2), but the lack of confidence in each assessment type might be induced through a different process.

Finally, the reliability of student assessment was partly questioned in the responses to open-ended questionnaires and semi-structured interviews in both assessment types. Regarding self-assessment, the teacher interviewees indicated that some students evaluated self-attainment or self-satisfaction highly and that affect might influence the reliability of self-assessment. With respect to peer assessment, the responses to the open-ended questionnaire and one student interviewee showed that the reliability of peer assessment might be influenced by the relationships among peers, so the reliability of peer assessment was doubtful.

To sum up, both self-assessment and peer assessment encouraged students to develop as writers. Moreover, while the peer assessment group students could enhance their metacognition, influenced by learning from peers, the intensive focus on writing through self-assessment made it possible for the self-assessment group students to concentrate on specific or general assessment criteria. Hence, both assessment types were effective in enhancing the development as a writer, but each assessment type affected the development as a writer in different ways.

Chapter 6

DISCUSSION

6.1 Introduction

This chapter discusses two research questions and integrates the results of quantitative and qualitative analyses obtained from the present mixed-method study on the efficacy of two types of student assessment on L2 writing: self-assessment and peer assessment. The first main research question concerned the similarities and differences between self-assessment and peer assessment: *How do self-assessment and peer assessment compare with each other?* It comprised three sub-research questions that were postulated to investigate the effects of student assessment types on assessment score reliability, learner writing ability, and learner affect. The second main research question was *How can student assessment work as the formative assessment in the classroom?* To address these main research questions, a convergent mixed-method research design was adopted.

A quantitative analysis was conducted to compare self-assessment and peer assessment (Research Question 1). To investigate the reliability of student assessment (Hypothesis 1.1), a many-facet Rasch measurement (MFRM) analysis was employed to examine the reliability of the two student assessment methods against teacher assessment. It was revealed that both assessment types were not as reliable as teacher assessment, but the severity of student assessment was closer to that of teacher assessment. The MFRM was also conducted with respect to the effects of student assessment on writing ability (Hypothesis 1.2). It was found that both assessment types had positive effects on the improvement of grammatical accuracy, but student assessment also had a positive influence on the composite score. Regarding the effects of student assessment on learner affect (Hypothesis 1.3), MANCOVA was carried out to analyse the change in learner affect after the intensive student assessment sessions. It was found that neither assessment type showed a change in writing anxiety but that learner autonomy was positively influenced by both types of student assessment methods.

The second research question asked about the relationship between student assessment and formative assessment. It was explored mainly by employing qualitative analyses. Student responses to an open-ended questionnaire were analysed by conducting term extraction. The transcripts of semi-structured interviews with 12 students and two teachers were also analysed in terms of the grounded theory. The results of the qualitative analyses indicated that both assessment types promoted the development of students as writers, though the assessment types supported students in the process of becoming good writers in different ways. In this chapter, the quantitative and qualitative data analysis results obtained in this study will be triangulated to answer the main research questions and discuss the efficacy of the two types of student assessment: self-assessment and peer assessment.

6.2 Research Question 1: How Do Self-Assessment and Peer Assessment Compare with Each Other?

The results suggested that self-assessment and peer assessment had different effects on student writing performance and learner affect. The quantitative analysis results were endorsed by qualitative analysis.

It is important for teachers in particular to become acquainted with these differences and similarities to make effective use of student assessment types. Understanding the features of student assessment facilitates teachers and students in the adoption of more effective learning strategies. A discussion on the three specific hypotheses of the first research question is presented in the following sections.

6.2.1 Hypothesis 1.1: Reliability of Student Assessment

Hypothesis 1.1 concerned the effects of student assessment on the reliability of scores compared with teacher assessment. The reliability was considered from the perspective of the stringency and consistency of teacher and student assessments. A many-facet Rasch measurement (MFRM) analysis of the reliability of student assessment showed that all four teachers in the data of the MFRM performed similarly to one another with high consistency and severity of the ratings that they assigned to students' writing. The students in the self-assessment group exhibited a similar level of strictness to the teachers in assessing their own work. It was also found that peer assessment group students were the most lenient among all the groups of assessors, that is, self-assessors, peer assessors, and teachers. However, with respect to the consistency of ratings, the self-assessment and peer assessment groups did not meet the standard of the teachers.

The results obtained from the qualitative data of the semi-structured interviews also revealed that the students in both student assessment type groups felt doubtful about the reliability of student assessment since they did not have confidence in their evaluation of their own or their peers' writing because of their lack of proficiency in writing English. In other words, they felt that neither student assessment type was as reliable as teacher assessment.

Regarding self-assessors, the qualitative analysis of the open-ended questionnaires and semi-structured interviews showed that they tended to pay special attention to language use. To be specific, self-assessors frequently reflected on their written production and writing performance, especially in terms of language use, even after the sessions. It can therefore be concluded that strict self-reflection was related to the severity of self-assessment because many of the self-assessors commented on their use of grammar and vocabulary after the sessions in their responses to the open-ended questionnaire.

In contrast, the peer assessors did not tend to give specific comments on language use. This was reflected in the peer assessors' perceived difficulty in undertaking peer assessment because of their lack of confidence in English proficiency, the reliability of peer assessment, and their relationships with their peers. However, the number of negative comments from the students in the peer assessment group was much smaller than that from the self-assessment group. Rather, the students in the peer assessment group demonstrated a positive attitude towards learning language use from peers' writing. For instance, the students in the peer assessment group expressed their enjoyment of peer assessment. The teacher interviewees also stated that peer assessment encouraged students to learn actively through the interaction with their peers.

Based on the previous studies on the reliability of student assessment before the present study, it was hypothesized that peer assessment was more reliable than self-assessment. However, the qualitative results highlighted students' comments about the doubtfulness of student assessment reliability. Furthermore, the quantitative analysis did not demonstrate the consistency of self- and peer assessment, so it was concluded that self- and peer assessment were not reliable. However, each assessment type showed different characteristics, such as the degree of severity in student assessment, and the self-assessment results were found to be comparable to those of teacher assessment.

6.2.2 Hypothesis 1.2: Effects on the Improvement of Writing Ability

Hypothesis 1.2 concerned the effects of the student assessment type on the improvement of writing ability. The results of the quantitative analysis based on the MFRM analysis of writing ability suggested that both assessment types resulted in an improvement in certain aspects of writing ability between the pretest and the post-test, conducted after the intensive student assessment sessions. Both groups improved their scores on grammatical accuracy, while only the self-assessment group increased the composite scores. One reason for the development of the composite score of the self-assessment group was a significant improvement in grammatical accuracy after the session. Thus, in general, it was found that the students in the self-assessment group were able to improve their writing ability more than those in the peer assessment group.

The qualitative analysis of the semi-structured interview protocols presented positive comments on the improvement of writing skills and English proficiency from both assessment types. The students in the self-assessment group indicated the usefulness of self-assessment in improving their writing ability, reporting that it prompted them to be aware of weaknesses such as grammar and vocabulary and to be more conscious of their readers. They also mentioned that this type of awareness helped them to reflect on what they needed to do to improve their English proficiency. On the other hand, the peer assessment group students commented that they could learn how to write from peers' writing, in particular their use of language, but how they might learn from specific comments on mistakes or advice offered by their peers was not frequently reported. In other words, the responses of the peer assessment group tended to be a general overview or impression of the peers' writing, including many compliments and much encouragement. Comments were not frequently found concerning specific suggestions relating to peers' writing, especially on the use of language.

The results of the qualitative analysis of the semi-structured interview data supported the improvement of students' grammatical accuracy in both student assessment groups found in the MFRM analysis. For the students in the self-assessment group, a habit of self-revision and consciousness of the assessment criteria were commented on by the student interviewees. The teacher interviewees also stated that the assessment criteria helped students to reflect on their writing, especially their usage of grammar. With respect to the students in the peer assessment group, peer assessment encouraged them to understand the assessment criteria for evaluating peers' writing. In other words, such comprehension of the assessment criteria also enhanced the students' reflective attitude towards grammatical accuracy. The students in the peer assessment group also learned accurate usage of grammar through peer reading and assessment. Hence, both types of student assessment facilitated the improvement of grammatical accuracy. Moreover, the learning awareness of the self-assessment group, particularly with regard to grammatical accuracy, seemed to provide a viable explanation for the grammatical improvement observed in the self-assessment group. The students in the peer assessment group also mentioned the influence of peers' writing on their own writing. However, the effect of peers' suggestions and their awareness of writing deficiencies seemed to be ambiguous. This may indicate that the peer assessment group students were not able to identify the problem areas in their writing compared with the self-assessment group.

It was hypothesized before conducting the present study that self-assessment and peer assessment would both have positive effects on writing ability. This hypothesis was partly confirmed because both student assessment types were effective in improving grammatical accuracy. Above all, self-assessment demonstrated its strengths with regard to the improvement of grammatical accuracy because self-reflection enabled the students to be conscious of their weaknesses in grammar and motivated them to be more accurate in the next session. In contrast, peer assessment positively influenced students' grammatical accuracy due to learning from peers and adopting a reflective attitude, but specific indications on writing problems were found less frequently than for the students in the self-assessment group.

6.2.3 Hypothesis 1.3: Effects on Writing Anxiety and Learner Autonomy

With respect to Hypothesis 1.3, the effects of student assessment on writing anxiety and learner autonomy was investigated. The results of a quantitative analysis gained from MANCOVA suggested that both assessment types had some effects on learner autonomy while their effects on writing anxiety did not present differences between the pretest and the post-test.

Writing anxiety, defined in this study based on Daly and Wilson (1983), comprised three subscale components: somatic, avoidance, and cognitive anxieties. The results of a quantitative analysis based on MANCOVA showed that neither self-assessment nor peer assessment resulted in an increase or a decrease in writing anxiety between the questionnaires conducted as the pretest and post-test. In terms of the mean ratings (Table 4.3.1), both assessment types indicated low level writing anxiety in the questionnaire conducted as the pretest: mean for the self-assessment group: 2.12; mean for the peer assessment group: 2.22. In the questionnaire conducted as post-test, those means of writing anxiety were also considered to be low level writing anxiety: mean for the self-assessment group: 2.16; mean for the peer assessment group: 2.21. After the intensive student assessment sessions, the level of writing anxiety seemed to be similar in terms of the mean ratings in both assessment groups. Furthermore, the self-assessment and peer assessment groups had a similar pattern of correlations between the subscales of writing anxiety (Chapter 4.3), showing that there were correlations between somatic anxiety and cognitive anxiety and between avoidance and cognitive anxiety. Therefore, both student assessment types can be considered to have had similar effects on writing anxiety.

In the qualitative analysis of the semi-structured interview data, negative comments associated with writing anxiety, such as worries, nervousness, doubtfulness, and lack of confidence, were reported by both student assessment groups, although the frequency of negative comments was lower than that of positive comments on other categories. These negative responses were mostly related to the achievement of tasks and the reliability of student assessment. According to the analysis of the responses to the open-ended questionnaire, the comments related to writing anxiety were not among the highest-ranking ones except the worries about the time restriction for peer assessment, which ranked tenth of ten items (Table 5.1). Hence, the results of the qualitative analysis corroborated the results of the quantitative analysis.

The reason for neither assessment type showing effects on writing anxiety is considered to be associated with the learners' perception of English writing during student assessment sessions. To be specific, the students in both groups seemed to switch individual writing anxiety or change it into a different affect in spite of feeling anxiety or nervousness throughout the intensive sessions. For instance, the students in the self-assessment group commented on self-attainment and self-satisfaction as well as a lack of confidence and inferiority (Chapter 5). In other words, these students felt anxious but at the same time valued their efforts and work highly, so they transformed their anxiety into self-attainment through writing. On the other hand, the peer assessment group students generally enjoyed reading and assessing peers' writing, even though they confessed some anxiety about a lack of confidence in peer assessment. In other words, the students in the peer assessment group valued the interaction with peers through English writing despite feeling anxious. Therefore, the results of the qualitative analysis suggested that writing anxiety was converted into self-attainment for the students in the self-assessment group and rapport with peers for the students in the peer assessment group.

Regarding learner autonomy, its improvement was recognized after the intensive student assessment sessions in both assessment types, as shown in the results of the quantitative analysis. In terms of descriptive statistics, both assessment types indicated similar mean ratings: 1.93 for the self-assessment group and 1.96 for the peer assessment group in the questionnaire conducted as the pretest. In the questionnaire conducted as the post-test, the self-assessment group scored 2.23 while the peer assessment group achieved 2.33. Thus, both assessment types increased their mean ratings. The results of the MANCOVA also presented statistical significance in the difference in learner autonomy between the questionnaires conducted as the pre- and post-tests for both assessment types. Furthermore, the learner autonomy level on the pretest was positively correlated with that on the post-test, so those students who had higher learner autonomy before the intensive student assessment sessions tended to have higher learner autonomy after the intensive student assessment sessions.

As the results of the quantitative analysis demonstrated the positive effects of student assessment on learner autonomy, the results of the qualitative analysis supported the improvement of learner autonomy. However, the student assessment type might influence the promotion of learner autonomy in different ways. To be specific, the results of the qualitative analysis of the semi-structured interviews and the responses to the open-ended questionnaire suggested that the motivation of the self-assessment group students was heightened by self-reflection and an awareness of weaknesses in their writing. The qualitative analysis also indicated that the peer assessment group students were motivated by their peers' encouragement and compliments. This type of stimulation, which occurred in the student assessment sessions, influenced the development of learner autonomy.

Before the current study, it was hypothesized that self-assessment and peer assessment would positively influence the development of learner autonomy but that only writing anxiety would be affected by self-assessment. Positive effects of both assessment types on learner autonomy were found, but a positive effect on writing anxiety was not identified in the study. The reason behind learner autonomy being related to both types of student assessment will be discussed in the next section.

6.2.4 Interpretation of the Results of Research Question 1

One of the aims of the study was to explore the efficacy of student assessment by examining the difference between self-assessment and peer assessment. It was found that both assessment methods were effective in improving writing performance and learner affects, but self-assessment and peer assessment were somewhat different from each other in their influences.

First, the above synthesis of quantitative and qualitative analysis results suggests that self-assessment improved students' writing ability more than peer assessment. While peer assessment was also useful for improving grammatical accuracy, self-assessment resulted in the improvement of overall English writing as well as having a positive effect on grammatical accuracy. This may be because self-reflection and self-awareness were ongoing for students in the self-assessment group during the writing of a composition and even after the intensive sessions. Such self-focus through self-assessment encouraged the students to identify the shortcomings of their writing and motivated them to perform better in the subsequent session. The severity of the self-assessment group students as assessors might also have prompted the improvement of grammatical accuracy in particular. Therefore, self-assessment can be considered to have been a more useful assessment type from the perspective of improving students' writing of English.

Regarding learner affect, learner autonomy was positively affected by both assessment types while writing anxiety was not decreased by either assessment type. The development of learner autonomy was found in both assessment types, but each type influenced its development in different ways. To be specific, ego involvement, such as self-esteem, acted on the improvement of learner autonomy for the self-assessment group while socialization positively influenced the promotion of learner autonomy in the peer assessment group.

These findings are supported by self-determination theory (SDT; Deci & Ryan, 1985, 2000). Self-determination theory proposes that people share three basic psychological needs: autonomy, competence, and relatedness. The theory predicts that, when these three needs are supported by social contexts and can be fulfilled by individuals, well-being is intensified. According to Ryan and Deci (2017), "autonomy is based on self-endorsement; it is supported and deepened by authentic self-reflection and mindfulness" (p. 59). The more "automatic" one's behaviour, the more one is at risk of being controlled. This is one reason why mindfulness is considered to relate to greater autonomy (Brown & Ryan, 2003). It has also been observed that "the act of engaging in self-reflection itself can deepen a sense of self-determination or autonomy" (Ryan & Deci, 2017, p. 56).

In reference to the comments made by the students in the self-assessment group, words related to self-reflection were frequently reported in the responses to interviews and questionnaires. Therefore, it is evident that the students in the self-assessment group were able to develop greater learner autonomy through the intensive writing sessions. Additionally, the students in the self-assessment group used words related to pride and self-esteem (Table 5.2). On the one hand, they reported a lack of confidence. At the same time, however, they expressed self-satisfaction and self-attainment. These results might be interpreted as indicating that the self-assessment students displayed an aspect of ego involvement. According to Greenwald (1982), in psychology, ego involvement is categorized into three types: (1) a struggle with threats to esteem made by others; (2) threats to self-esteem, specifically one's ego being endangered with respect to self-evaluation rather than evaluation by others; and (3) a more general or undifferentiated phenomenon because of personal importance (Ryan & Deci, 2017, p. 168). The second category, threats to self-esteem, can be considered to represent the ego involvement of the students in the self-assessment group, so it might show that the learner autonomy of the self-assessors developed because of the threats to their self-esteem.

With respect to peer assessment, it is supposed that the students in the peer assessment group nurtured their learner autonomy differently from those in the self-assessment group. Socialization is closely related to the development of learner autonomy for the students in the peer assessment group. Friedman (2000, p. 39) stated that the human capacity for autonomy develops in the course of socialization. Relationship motivation theory (RMT), which was proposed within SDT, suggests that the need to relate to others is intrinsic and inclines people to be volitionally engaged in close relationships. If relatedness or social interaction were undermined, the autonomy to connect with others would decrease. In the peer assessment sessions in the present study, students interacted with peers. One teacher commented on the cheerful atmosphere created by peer assessment and the interaction involved in exchanging assessments. This view might suggest that the learner autonomy of the students in the peer assessment group was enhanced by socialization.

Autonomy is also concerned with integrated, self-endorsed actions, so there is “a willingness to act as one does and an endorsement of the motivation that leads one to act in this way” (Ryan & Deci, 2017, p. 57). Students in both assessment groups showed willingness to engage in or self-endorsement of student assessment; therefore, both self-assessment and peer assessment can be considered to promote learner autonomy.

In sum, both assessment types were effective in improving writing ability and learner autonomy, but it was found that they acted differently on the development of writing ability and learner autonomy. Concerning the improvement of writing ability, self-assessment was a more useful tool. However, the qualitative analyses of the responses to the open-ended questionnaire and semi-structured interview protocols illuminated different aspects of peer assessment, which was perceived to have made the writing classes more interactive and communicative. While it is difficult to state which assessment method is superior, the most important outcome would be to understand the characteristic features of both student assessment types and to implement them effectively in the classroom.

6.3 Research Question 2: How Can Student Assessment Work as Formative Assessment in the Classroom?

Previous studies have presented several definitions of formative assessment in terms of process, instrument, and contrast with summative assessment (Black & Wiliam, 2003; Garrison & Ehringhaus, 2013; Rea-Dickins, 2000; Scriven, 1967). In the present study, formative assessment was defined as "assessment carried out during the instructional process for the purpose of improving teaching and learning" (Shepard, Hammerness, Darling-Hammond, Rust, Snowden, & Gordon, 2005, p. 275). Formative student assessment is frequently adopted in the process of classroom learning.

Formative assessment provides teachers and students with information or evidence of student achievement. In concrete terms, formative assessment means monitoring students' learning progress and the meeting of their goals. In other words, formative assessment offers teachers and students a bridge between teaching and learning. Leahy and Wiliam (2012) also suggested five important strategies to conduct formative assessment successfully: (1) clarifying learning intentions and sharing criteria for success; (2) drawing evidence out of learning; (3) providing feedback that promotes learning; (4) encouraging learners to take on roles of instructional resources for one another; and (5) promoting learners to become owners of their own learning (p. 7). These five key strategies are used to analyse the second research question of the current study concerning the relationship between formative assessment and student assessment through qualitative analyses (Research Question 2).

First, regarding the clarification and understanding of the assessment criteria, the students in the self-assessment group tried to understand where they were in the process of study in relation to the learning goal. Many of the students in the self-assessment group mentioned their awareness of the gaps between the study goal and their present stage of learning. This is because understanding the assessment criteria helped the students to clarify the present standpoint of their learning in relation to the goal.

The same applied to the peer assessment group in the sense that the students in the peer assessment group were also encouraged to understand the criteria to evaluate the writing of their peers. However, in the quantitative analyses based on the many-facet Rasch measurement, peer assessors were the most lenient assessors and the reliability of the assessment was doubtful (Chapter 4.1). The results of the qualitative analyses of semi-structured interview data also revealed that the students in the peer assessment group made general comments on their partners' writing while the students in the self-assessment group noted their own weaknesses specifically. This is because some of the students in the peer assessment group felt the difficulty of evaluating a peer's composition because of a lack of English ability or a lack of confidence in making an evaluation as they did not want to criticize their classmates by using negative evaluative words (Freeman, 1995; Orsmond et al., 1996; Vu & Dall'Alba, 2007).

Ultimately, both assessment types could help students to clarify and understand the learning goals through the assessment criteria, but the self-assessment group students carried out Leahy and Wiliam (2012) first strategy of formative assessment, that is, clarifying, sharing, and understanding learning intentions and criteria, more successfully. This may be because the students in the self-assessment group attended more consciously to their weak areas, such as language use, than the students in the peer assessment group. In other words, self-assessment enabled students to concentrate on self-testing using the assessment criteria, which helped the students to determine where they had difficulties in their writing, despite their lack of confidence in assessing writing. The results of the quantitative analysis of the reliability of student assessment also showed a difference in the severity of the evaluation (Chapter 4.1) between the self-assessment and the peer assessment group. The MFRM analysis identified greater severity among the self-assessment group than among the peer assessment group.

Second, it is questionable whether both assessment types could reliably draw evidence out of learning (Leahy & Wiliam, 2012), because the results of the quantitative analysis of the MFRM revealed that neither student assessment type was as reliable as teacher assessment. However, the results of the qualitative analysis of semi-structured interviews showed that both types of student assessment moderately enabled students to find specific strengths and weaknesses in their writing but that the elicitation of the self-assessment group seemed to be more specific than that of the peer assessment group.

According to the results of qualitative study, the elicitation of evidence was not reliably ensured by the students in the self-assessment group because of a lack of confidence in English ability and self-assessment. This is because the students in the self-assessment group endorsed the necessity of the teacher's effective instructional adjustments to respond to their learning needs. These students might have overestimated the reliability of the teacher's evaluation, have strong psychological dependence on the teacher's guidance, or both. To be specific, the students in the self-assessment group identified their learning needs, such as a shortage of vocabulary and grammatical accuracy, and pointed out the lack of trustworthiness of their assessments if appropriate feedback from teachers was lacking.

On the other hand, fewer students in the peer assessment group provided more comments that may be considered to be specific corrective feedback as evidence of their partners' learning than the students in the self-assessment group did. As mentioned above, the students in the peer assessment group were the most lenient in assessing their partners' writing. Some students commented on the difficulty of making comfortable and pertinent remarks on their peers' work. As a result, it was found that peer assessment tended to be vague and lenient, even though they hoped to share learning. Based on these findings, Leahy and Wiliam's (2012) second strategy, elicitation from learning, was accomplished by both assessment groups while the peer assessment group's elicitation seemed to have been weaker than that of the self-assessment group.

Leahy and Wiliam's (2012) third key strategy of effective formative assessment concerns whether student assessment provides feedback that moves learning forwards. It should be noted, before discussing this point, that self-assessment and peer assessment have a notable difference. In self-assessment, students themselves reflect on their work after the writing sessions. Therefore, this does not constitute feedback from others, although self-reflection provides students with insights into their learning. The peer assessment group students, however, mutually exchanged positive or encouraging feedback, and this feedback motivated the students to study more effectively. From this perspective, only peer assessment might lead to the provision of feedback that may move students' learning forwards.

The purpose of feedback is to provide students with a constructive evaluation of writing quality (Biber, Nekrasova, & Horn, 2011). According to Finn and Metcalfe (2010), effective feedback provides students with information about the correct answer to a question that they answered incorrectly. Feedback also needs to direct attention to what is next for the student, rather than focusing only on past achievement. In other words, feedback should not be just a judgement or punishment but a formative tool to help learners move forwards (Wiliam, 2018). If feedback works formatively, learners can use the information to improve their performance (Wiliam, 2018).

Nyquist (2003) categorized feedback into five types of effect: (a) weak feedback only; (b) feedback only; (c) weak formative assessment; (d) moderate formative assessment; and (e) strong formative assessment. Weak formative assessment means just giving information. Moderate formative assessment is defined as giving information and suggestions for how to improve, and strong formative assessment means giving information, an explanation, and a specific activity to undertake in the future. In terms of the effect of peer assessment on formative assessment, peer assessment tended to be categorized as weak or moderate formative assessment because the students in the peer assessment group indirectly gave their partners feedback, praise for good points or advice for the next session, and mostly encouragement or compliments. It was rare for peer assessment group students to correct the errors of their peers directly. Therefore, some students were doubtful about the reliability of peer assessment because of a lack of specific criticism of their writing. Previous studies have stated that praise is not necessarily a good thing to affect students' achievement (Butler, 1987; Good & Grouws, 1975). Rather, high levels of task involvement or scaffolded responses help students to learn more effectively. In short, peer assessment encouraged the students to learn in the next session but the feedback did not clearly suggest how to improve or provide a specific activity to undertake in the future. Peer assessment could thus be categorized as weak or modest formative assessment.

The fourth key point regarding formative assessment proposed by Leahy and Wiliam's (2012) is activating learners as instructional resources for one another. Slavin, Hurley, and Chamberlain (2003) stated that the effectiveness of cooperative learning is fourfold. First, well-structured cooperative learning encourages students to learn. Second, social interaction with peers prompts students to make efforts. Third, students could learn more through cooperative learning because their more capable peers are experiencing the same particular difficulties as them. Finally, students who help others could enhance their cognition because they have to think and describe the ideas more clearly to make their peers understand. As far as student assessment is concerned, peer assessment shares these elements because students are expected to work to achieve a common goal with their peers. Furthermore, individual students need to have accountability as individual assessors when they evaluate a peer's writing. Self-assessment, however, does not provide an opportunity to be activated as instructional resources for others, although the students in the self-assessment group could also internalize learning resources and criteria through self-reflection. In this sense, therefore, peer assessment functions more effectively to promote formative assessment than self-assessment.

Finally, in terms of Leahy and Wiliam's (2012) fifth point, it is possible for both types of student assessment to activate learners as owners of their own learning. In other words, students could promote formative assessment through self-assessment and peer assessment. During the intensive student assessment sessions, students in both assessment groups became more reflective about their writing performance. In the present study, two teacher interviewees also reported that students became used to monitoring their learning and trying to compare and synthesize their English proficiency and the goal of the study into assessment during the student assessment sessions.

In particular, the peer assessment group students evaluated their peers' compositions by comparing them against their own compositions through peer assessment (Table 5.1). The comments made by the peer assessment group showed that they reflected on their writing skills in contrast with those of their peers, so the peer assessment could be said to be a variation of self-assessment (Huerta-Macias, 1995). More specifically, the peer assessment group students tended to show more awareness of others and multiple perspectives in terms of readers, writers, assessors, and friends than those in the self-assessment group. In other words, the students in the peer assessment group more frequently presented metacognitive reviews than the students in the self-assessment group, as shown in the summaries of the qualitative analyses in Figures 5.1 and 5.2. The students in the self-assessment group also mentioned metacognitive aspects, such as consciousness of imaginary readers, but it was found that the peer assessment group students presented stronger metacognitive aspects than the self-assessment group.

As can be seen in the main results summarized above, both assessment types enabled students to be aware of readers' presence and reflective on their study. In addition, the present study suggested that both types of student assessment gave students a chance to take responsibility for their own assessment, and this helped them to manage their learning. It is thus considered that both self-assessment and peer assessment accelerated the process of students becoming self-regulating learners. According to Boekaerts (2006), self-regulated learning enables the learner to coordinate cognitive resources, emotions, and actions in the service of his or her learning goals. In the present study, it was observed that students made efforts to bring their awareness of learning, knowledge, emotion, and strategies into the ideal goal after student assessment. Student assessment thus also plays a useful role in formative assessment because student assessment motivates learners to start to manage their learning process and outcomes. In short, students could learn to develop insights into their own writing through student assessment.

Following the five key stages of formative assessment based on Leahy and Wiliam (2012), the findings of the present study are summarized in Table 6.1. Peer assessment mostly covers Leahy and Wiliam's (2012) five strategies for formative assessment, while self-assessment only covers three strategies. However, it is difficult to judge whether peer assessment is superior to self-assessment in terms of its formative effects. This is because the students in the self-assessment group concentrated more specifically on monitoring their learning than those in the peer assessment group. Sadler (1989) stated that formative assessment includes both feedback and self-monitoring, so teachers should facilitate the transition from feedback to self-monitoring. Feedback is information about "the gap between the actual level and the reference level of a system parameter, which is used to alter the gap in some way" (Ramaprasad, 1983, p. 4). Students are expected to use feedback to identify the strengths and weaknesses of their performance and to improve their performance. Stated explicitly, peer assessment group students broadly comprehended the five key strategies, but self-monitoring was weaker than among those in the self-assessment group. On the other hand, the students in the self-assessment group showed strong self-monitoring, while feedback from others could not be expected.

Table 6.1

Leahy and Wiliam's (2012) Five Key Strategies of Formative Assessment by Student Assessment Type

	Self-assessment	Peer assessment
Clarifying learning intentions and sharing criteria for success	+	+/-
Eliciting evidence of learning	+	+/-
Providing feedback that moves learning forwards	-	+/-
Activating learners as instructional resources for one another	-	+
Activating learners as owners of their own learning	+	+

Note. + stands for good; - stands for bad; +/- stands for average.

A typical class has teachers and students who rely more or less on their teachers to be helped in the classroom. The classroom assessment environment is created mostly by the teacher (Stiggins & Conklin, 1992). Wiliam (2018) also mentioned that "the most important factor influencing learning is what the learner already knows, and that the teacher's job is to ascertain what the learner already knows and to teach accordingly" (p. 122). In the present study, teachers participated not as instructors but as observers, so some students showed less positive attitudes towards self-assessment or peer assessment because they were sceptical about the meaningfulness of student assessment (Freeman, 1995; Orsmond et al., 1996; Vu & Dall'Alba, 2007). The absence of the teacher from classroom assessment illuminated, if anything, the role of the teacher in classroom assessment.

To make formative assessment successful, teachers are expected to intervene in self-assessment and peer assessment. This is because the self-assessment and peer assessment group students did not put great trust in the reliability of student assessment, so they expected teacher assessments. Formative assessment has been used to gather information about the degree of success of students' respective efforts in the classroom. It allows "teachers to diagnose students' strengths and weaknesses in relation to specific curricular objectives and thus guides them in organising and structuring instructional material" (d'Anglejan, Harley, & Shapson, 1990, p. 107). Therefore, teacher assessment is required to be part of both types of student assessment and to make formative assessment complete. Teachers are expected to give individual students reliable information on their progress and learning goals.

Those findings about the relationship between formative assessment and student assessment could also be interpreted from the perspective of the theoretical framework of learning-oriented assessment (LOA: Turner & Purpura, 2016), which has recently gained attention as a comprehensive theoretical framework, because student assessment enables students to be centred in learning. In addition, student assessment makes it possible to bridge learning and teaching formatively with sharing formative information with teachers. According to Turner and Purpura (2016), LOA has been promoted by formative assessment research on content learning. In the theoretical framework of LOA, the responsibilities of stakeholders in classroom assessment, that is, teachers and students, are highlighted. To be specific, teachers and students are expected to share responsibilities for learning and assessment (Turner & Purpura, 2016, pp. 256–258). As LOA advocates, student assessment could be effective based on the evidence to present specific feedback and guide students towards a learning goal. It means that the efficacy of student assessment could be demonstrated with teachers' assistance based on definite evidence related to individual students' development. Therefore, teachers are encouraged to link learning and assessment formatively by gathering information from student assessment. This is because the data obtained from student assessment support teachers in clarifying individual students' learning perception, process, and attainment of their learning goal.

Hence, the role of teachers deserves attention in relation to student assessment in terms of formative assessment. Therefore, it is important for teachers to know the features of each student assessment type to utilize it successfully in class because self-assessment and peer assessment have different functions and differential effects on formative assessment. As Rea-Dickins and Gardner (2000) discussed regarding the validity and appropriateness of formative assessment, formative assessment might be "an informal procedure rather than being systematically integrated into the curriculum and classroom practices" (p. 231). In addition, student assessment might be considered to be informal and unsystematic as an assessment method. However, if teachers had a solid plan on how to intervene during the process of student assessment, student assessment would be fruitful as formative assessment.

The present study aimed to extract the effects of student assessment on students' writing performance and learner affect; however, the semi-structured interviews and the students' responses to open-ended questionnaires indicated the necessity of teacher assessment due to students' lack of confidence. This qualitative analysis supported the results of the MFRM analysis because neither student assessment type presented reliability. As successful formative assessment should provide immediate and appropriate feedback to students and as such feedback could lead to the improvement of students' academic skills (Davis, Elder, Hill, Lumley, & McNamara, 1999, p. 65), teacher assessment or guidance is also essential to complete formative assessment.

6.4 Summary of the Findings

In the present study, teachers participated as observers in the classroom, so they did not instruct students at all during the intensive sessions. This procedure was followed to eliminate the effect of teachers on the survey responses and to abstract the effects of student assessment as much as possible. The absence of teachers from the classroom illuminated the role of teachers in classroom assessment. Teachers are expected to pilot students towards their learning goal. The efficacy of both types of student assessment was established and their relative strengths were clarified, yet the efficacy of student assessment would be triangulated into formative assessment by teachers' tact based on their teaching experience and elaborate teaching plans. Teaching experience nurtures the ability of teachers to observe individual students' growth, ability, and needs, while complex, well-prepared teaching plans before the class develop student assessment into successful formative assessment.

Chapter 7

CONCLUSION

7.1 A Review of the Study Objectives

The present study aimed to break through the stagnant condition of English writing education for Japanese high school students. As one possible approach to contributing to the reform of English writing instruction, two types of student assessment methods, self-assessment and peer assessment, were examined. The rationale behind this approach was that students' involvement in classroom assessment would encourage them to become more active and responsible learners. It was also assumed that the change of agents in assessment from teachers to students would improve formative assessment, which involve the integration of learning and teaching. Thus, the second aim of the present study was to investigate the effect that student assessment would have on formative assessment.

The review of the relevant literature presented in Chapter 2 suggested that the similarities and differences between the two types of student assessment, that is, self-assessment and peer assessment, have not yet been analysed thoroughly, though many previous studies have discussed student assessment with a focus on the possibilities for student assessment to replace teacher assessment from the perspective of score reliability.

However, as discussed in Chapter 2, the effect of student assessment might differ depending on the learners' ages, cultural backgrounds, and levels of English proficiency, so the results of previous studies differed across learners' conditions and educational environments. Furthermore, little empirical evidence concerning Japanese adolescent students on this topic was available. In addition, few previous studies have focused on the relationship between learner affect and student assessment. Therefore, it is meaningful that the present study aimed to contribute to the improvement of English writing education by elucidating the effect of individual student assessment methods on writing performance and learner affect within a single study.

7.2 Major Findings

The results of the quantitative analyses of reliability showed that neither assessment type was as reliable as teacher assessment, while the rating severity of self-assessment was found to be similar to that of teacher assessment. Peer assessment was considered to be the most lenient method compared with teacher and self-assessment. Both self-assessment and peer assessment had positive effects on the improvement of grammatical accuracy, while only self-assessment presented a development of overall writing ability. With respect to the effects of learner affect, the quantitative analysis of MANCOVA indicated that there was no remarkable increase or decrease in writing anxiety in either assessment condition. However, learner autonomy significantly improved in both assessment type groups after the intensive assessment sessions. Furthermore, it was also found that the students in both groups who had higher learner autonomy tended to have higher learner autonomy after the intensive student assessment sessions. In sum, the results of the quantitative analyses presented some similarities and differences between self-assessment and peer assessment.

The results of the qualitative analyses of the semi-structured interview protocols and open-ended responses to the questionnaire on writing anxiety and learner autonomy indicated that both types of assessment could exert positive effects but that the two student assessment methods differed in efficacy as formative assessment methods. Specifically, self-assessment and peer assessment had different effects on students' involvement in self-reflection. The self-assessment group presented self-reflection outstandingly in their writing, in their academic achievements, and in their learning strategies. In contrast, the peer assessment group showed enjoyment in reading and exchanging comments on their writing with their classmates. Such reciprocity of assistance between peers deepened their awareness of accurate language use and of the presence of readers. In terms of the five strategies of formative assessment (Leahy and Wiliam, 2012), peer assessment covered all five strategies, though it showed moderate conformity with three of them, while self-assessment covered three of them (Table 6.1). This is because self-assessment does not allow for external feedback and the exchange of instructional resources.

Therefore, both assessment methods could have made a contribution to improving formative assessment in class because formative assessment could provide important information regarding learners' strengths and weaknesses that the teacher could use as feedback to improve his or her subsequent instructional decisions (Poehner, 2008). To enhance formative assessment, the self-assessment group used self-assessment as self-mediation, which enabled the students to notice their weaknesses and strengths in writing, while the interactive feedback in the peer assessment, in particular, peers' respect and enjoyment of learning with peers, functioned to motivate students to progress.

7.3 Formative Assessment

Formative assessment is constructed by the collaboration of teachers and students, specifically in the process of teaching and learning. In formative assessment, teachers are supposed to focus their classroom assessment more directly on learners' development and can actively involve learners in this process (Harlen & Winter, 2004). Teachers are also expected to know the students' framework for learning to connect students' learning experiences and capacity appropriately with their framework (Harlen & Winter, 2004, p. 392). In other words, teachers need to understand individual students' existing capacity to match their learning needs effectively with a learning goal. Therefore, it is necessary to consider the role of the teacher in formative assessment in relation to student assessment. Teacher assessment is considered to be everyday assessment, that is, standard assessment as classroom assessment. The present study proposes that the change of agent in classroom assessment enables teachers to have new perspectives on individual students' attainment of a learning goal. This is because both assessment type students also commented on the necessity for teacher assessment in terms of reliability.

In the present study, students in the self-assessment group expressed worries about whether their effort was appropriately directed to the learning goals because, in this study, there was no instruction from teachers between the intensive writing practice sessions. This result signifies the importance of the teacher's role, which is to determine whether the students are following the right path in their study. In other words, teachers need to show students a map that clarifies the learning goals and their progress towards them.

The self-assessment group students tended to be strict assessors, as revealed by the quantitative analyses of reliability (Chapter 4). Specifically, they did not adjust the assessment criteria appropriately to the present students' abilities, so teachers should indicate and amend the severe judgements of students about their own writing products to more realistic ones through teacher assessment. Furthermore, teachers could add materials or criteria that suit individual students' levels and needs; it is important for teachers to analyse students' thinking before assuming that they have understood something (Wiliam, 2018, p. 87). In summary, teachers' appropriate intervention in the self-assessment process makes formative assessment more meaningful.

With respect to the role of peer assessment, it promotes more communication within the context of classroom assessment because peer assessment gives students a chance to communicate with each other about their writing. Students learn how to write English through social interaction with friends. Through peer assessment, students explain, justify, and take responsibility for their assessment on the basis of equality between their status and that of their peers. Compared with the relationship between teachers and students, peer learners are equal in the classroom. Thus, in comparison with teacher assessment, peer assessment reduces tension among the students. In the present study, when peer assessment enabled students to recognize their strengths and weaknesses, it was found that students cooperatively taught and learned from each other. For instance, the peer assessment group students commented that they did not feel that their grades were threatened in peer assessment because teacher assessment is more formal and demanding as well as providing information for the school record. Thus, peer assessment offers different aspects for students and teachers within the context of classroom assessment. It makes writing classes more active, communicative, cooperative, and sometimes relaxing as a result of students exchanging assessments among themselves.

The present study proposed that student assessment has a positive impact on students' learning and that it provides teachers and students with a linkage between teachers' instruction and students' learning. Student assessment also makes students conscious of the gap between the final learning goal and their actual abilities. In that sense, student assessment is effective in enhancing students' internalization and self-regulation in learning while its reliability is questionable. Moreover, the enhancement of self-regulation helps students to take responsibility for learning (Turner & Purpura, 2016). Therefore, appropriate teacher guidance or assessment would help students to understand accurately the distance between their actual abilities and their learning goals.

According to Black (2001), social interaction and language discourse are connected with formative assessment:

> The dependence of learning on teacher–pupil interaction is a very specific one, linked to the nature of the teacher's guidance. Most theorists emphasize the importance of language in learning. Interaction takes place through language discourse, which is learned and understood in particular social contexts. It would follow that the nature of our learning depends on the particular "communities of discourse". (pp. 15–16)

If teachers paid attention to the results of student assessments and interacted with students while referring to the results of these assessments, the interactions would provide them with useful information about what the learners know. Furthermore, the focus on the gap between teacher assessment and student assessment would enhance the insights into students rather than producing a judgmental evaluation (Torrance & Pryor, 1998). If student assessment were used as a tool for formative assessment, the interaction between teachers and students would make formative assessment more productive. This is because formative assessment should be constructed jointly by the collaboration between teachers and students. Student assessment could also generate information to guide teachers and students towards the learning goal.

7.4 Efficacy of Student Assessment

Self-assessment and peer assessment have positive effects on students' writing performance and learner affect, yet each assessment method presented different yet similar processes in their influence on the learners' performance and affect. The stringency of self-assessment directly acts on learners' self-reflection or self-introspection. Stringent assessment encourages students to revise and review their writing and leads to the improvement of writing ability and grammatical accuracy. On the other hand, peer assessment provides learners with opportunities to adopt multiple roles, such as assessor, reader, and friend. Those experiences encourage students to develop metacognitive views. In addition, the enjoyment gained from reading peers' writing gives students pure motivation to read and write because they learn how to write from their peers' writing and they enjoy interacting with their peers. Thus, each student assessment method has different but similar positive functions, so teachers and curriculum developers should complement each other to implement each assessment method in class depending on the class conditions.

To sum up, self-assessment should be adopted regularly after writing tasks to nurture individual students' internalization and self-regulation. On the other hand, peer assessment should be conducted before students present what they have written to the whole class. Peer assessment would work as an icebreaker and could deepen in-class discussions. Teachers are expected to read the atmosphere of the class and the human relationships among students to implement student assessment effectively as part of classroom instruction.

7.5 Contributions and Implications of the Study

As the previous literature reviewed in the present study showed, student assessment has been implemented in writing classes; however, the effects of student assessment on writing performance and learner affect have not been analysed for Japanese adolescent students. Furthermore, similarities and differences have not been distinguished specifically in previous studies. In that sense, the present study contributes to clarifying the important roles that student assessment plays in the classroom and provides useful information for teachers. In particular, based on the results of this study, one can expect either student assessment type to develop grammatical accuracy. Furthermore, learner autonomy is expected to be improved after the implementation of either student assessment type. Therefore, these findings are noteworthy for teachers and for students because it is important for teachers to understand the features of each assessment method to apply it effectively in class.

7.6 Limitations to the Study and Recommendations for Future Research

One limitation of the study was the focus on the effect of student assessment during an intensive but limited time span to prevent students from being exposed to other activities, such as other English learning activities as part of the school curriculum and in different environments. An intensive short-duration design made it possible to extract specific and genuine effects of student assessment on students' writing performance and learner affect. However, longitudinal designs have another possibility to illuminate changes in learners, such as their learning attitudes towards English writing in the long term. Given that the results of longitudinal designs would include other effects on the students' writing performance and learner affect, careful consideration of the social and educational backgrounds is advisable when conducting research with longitudinal designs.

In addition, the role of teachers in student assessment should be analysed in further studies since formative assessment involves the integration of teaching and learning, that is, collaboration among teachers and students. The present study revealed the function of student assessment as formative assessment, but the relationship between teacher assessment and student assessment as formative assessment needs to be investigated thoroughly. Therefore, the exact roles of teachers in student assessment should be explored to offer teachers effective guidance on how to conduct formative assessment.

7.7 Conclusion

The present study was an exploratory investigation into the effects of self-assessment and peer assessment on writing performance and learner affect. It also attempted to interpret the relationship between student assessment and formative assessment. Additionally, it discussed how teacher assessment should be associated with student assessment in terms of formative assessment. It was found that (a) self-assessment was comparable to teacher assessment in terms of severity of the composite score, while peer assessment was not similar to teacher assessment. In terms of analytic rating scales, the scores obtained in both student assessment types were not as reliable as teacher assessment (Hypothesis 1.1); (b) both assessment types had positive effects on the improvement of grammatical accuracy, while only self-assessment showed the development of the composite scores of writing (Hypothesis 1.2); (c) both assessment types had positive effects on learner autonomy, while writing anxiety was not influenced by either type of assessment (Hypothesis 1.3). As for Research question 2, it was found that both student assessment types are promising as methods for the successful implementation of formative assessment. In particular, self-assessment was found to increase students' internalization and self-regulation while peer assessment could lead to the development of students' self-regulation and metacognitive awareness.

It is my hope that the conclusions support a vision to reform English writing classes in Japan. Teachers especially may discover new ways to focus on their instruction in English writing classes to help learners to be responsible for their learning.

References

Andrade, H. & Du, Y. (2007). Student responses to criteria referenced self-assessment. *Assessment & Evaluation in Higher Education, 32*(2), 159-181. doi: 10.1080/02602930600801928

Andrade, H., Du, Y., & Mycek, K. (2010). Rubric-referenced self-assessment and middle school students' writing. *Assessment in Education, 17*(2), 199-214. doi: 10.1080/09695941003696172

Andrade, H. & Valtcheva, A. (2009). Promoting Learning and Achievement Through Self-Assessment. *Theory Into Practice, 48*(1), 12-19. doi: 10.1080/00405840802577544

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage Publications.

Arter, J. A., McTighe, & Guskey, T. R. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Newbury Park, CA: Sage Publications.

Ary, D., Jacobs, L. C., Razavieh, A. (2002). *Introduction to Research in Education, 6th Ed.* Belmont, CA: Wadsworth-Thomson Learning.

Ashraf, H., & Mahdinezhad, M. (2015). The Role of peer-assessment versus self-assessment in promoting autonomy in language use: A case of EFL Learners. *Iranian Journal of Language Testing, 5*(2), 110-120. Retrieved from http://www.semanticscholar.org/paper/The-Role-of-Peer-assessment-versus-Self-assessment-Ashraf-Mahdinezhad/a7f4764fc0bb730681690ab628f544a2a78aebc0?p2df

Bandura, A. (1986). *Social Foundations of Thought and Action: a Social Cognitive Theory.* Englewood Cliffs, NJ: Prentice Hall. Retrieved from http://en.wikipedia-on-ipfs.org/wiki/Social_Foundations_of_Thought_and_Action%3A_A_Social_Cognitive_Theory.html

Bandura, A. (1989). Regulation of cognitive processes through perceived self-efficacy. *Developmental Psychology, 25*(5), 729-735. doi: 10.1037/0012-1649.25.5.729

Bannister, L. (1992). *Writing apprehension and anti-writing: A naturalistic study of composing strategies used by college freshmen*. San Francisco, CA: Mellen Research University Press. Retrieved from http://digitallibrary.usc.edu/cdm/ref/collection/p15799coll20/id/216889

Barkhuizen, G., Wette, R. (2008). Narrative frames for investigating the experiences of language teachers. *System*, *36*(3), 372-387. doi: 10.1016/j.system.2008.02.002

Barnard, R. & Viet, H. (2010). Task-Based Language Teaching (TBLT): A Vietnamese

Case Study Using Narrative Frames to Elicit Teachers' Beliefs. *Language Education in Asia, 1*, 77-86. Retrieved from http://www.researchgate.net/profile/Gia_Viet_Nguyen/publication/228864316_Task-Based_Language_Teaching_TBLT_A_Vietnamese_Case_Study_Using_Narrative_Frames_to_Elicit_Teachers_Beliefs/links/544a7ae90cf24b5d6c3cc89b/Task-Based-Language-Teaching-TBLT-A-Vietnamese-Case-Study-Using-Narrative-Frames-to-Elicit-Teachers-Beliefs.pdf

Bazerman, C., Applebee, A. N., Berninger, V. W., Brandt, D., Graham, S., Matsuda, Pl K., Schleppegrell, M. (2017). Taking the long view on writing development. *Research in the Teaching of English, 51*, 351-360. Retrieved from http://library.ncte.org/journals/rte/issues/v51-3

Becker, A. (2011). Examining rubrics used to measure writing performance in US intensive English programs. *The CATESOL Journal, 22*(1), 113-130. Retrieved from http://www.catesoljournal.org/wp-content/uploads/2014/06/CJ22_becker.pdf

Behjat, F. & Yamini, M. (2012). Self- vs. Peer-editing: One Step Forward from Assessment to Building EFL Students' Writing Skill. *Journal of Studies in Learning and Teaching English, 1*(1), 65-85. Retrieved from http://jslte.iaushiraz.ac.ir/article_518867.html

Benesse Educational Research and Development Institute (2015). Retrieved from http://berd.benesse.jp/up_images/research/Eigo_Shido_all.pdf

Benesse Educational Research and Development Institute (2018). Retrieved from http://berd.benesse.jp/global/research/detail1.php?id=4356

Benson, S. H. (2000). Make mine an A. *Educational Leadership*, *57*(5), 30-32. Retrieved from link.gale.com/apps/doc/A62149127/AONE?u=waseda&sid=AONE&xid=f457090c.

Benson, P. (2001). *Teaching and researching autonomy in language learning*. London, UK: Longman.

Berry, V. (1994). Current assessment issues and practices in Hong Kong: A preview. In D. Nunan, R. Berry, & V. Berry. (Eds.), *Bringing about change in language education*: *Proceedings of the International Language in Education Conference 1994* (pp. 31-34). Hong Kong: University of Hong Kong.

Betz, N.E. (1978). Prevalence, distribution, and correlates of math anxiety in college students. *Journal of Counseling Psychology, 25*, 441-448. doi: 10.1037/0022-0167.25.5.441

Biber, D., Nekrasova, T., & Horn, B. (2011). *The Effectiveness of Feedback for L1-English and L2-Writing Development: A Meta-Analysis*. (ETS Research Report RR-11-05). Princeton, NJ: ETS. http://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2333-8504.2011.tb02241.x

Bishop, K. L., Holm, J.E., Borowiak, D.M., & Wilson, B.A. (2001). Perceptions of pain in women with headache: A laboratory investigation of the influence of pain-related anxiety and fear. Headache: *The Journal of Head and Face Pain, 41*, 494-499. doi: 10.1046/j.1526-4610.2001.01087.x

Black, P. (2001). Formative assessment and curriculum consequences. In Scott, D. (Eds.), *Curriculum and assessment* (pp. 7-23). Westport, CT: Ablex Publishing.

Black, P. & Wiliam, D. (1998a). Assessment and Classroom Learning. Assessment in Education: Principles. *Policy & Practice, 5*(1), 7-74. Retrieved from http://assess.ucr.edu/sites/g/files/rcwecm2336/files/2019-02/blackwiliam_1998.pdf

Black, P., & Wiliam, D. (1998b). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139–147. Retrieved from http://d1wqtxts1xzle7.cloudfront.net/40440148/Black_Wiliam_blackbox1998_1.pdf

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for Learning: putting it into practice*. Buckingham: Open University Press.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working Inside the Black Box: Assessment for Learning in the Classroom. *Phi delta kappan, 86*(1), 8-21. Retrieved from http://www.researchgate.net/profile/Dylan_Wiliam/publication/44835745

Black, P. & Wiliam, D. (2003). In Praise of Educational Research: formative assessment. *British Educational Research Journal, 29*(5), 623-637. doi: 10.1080/0141192032000133721

Black, P. & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 1*(1), 1-40. doi: 10.1037/0022-0167.25.5.441

Blanche, P., & Merino, M. (1989). Self-assessment of foreign-language skills: Implications for teachers and researchers. *Language Learning, 39* (3), 313-340. doi: 10.1111/j.1467-1770.1989.tb00595.x

Bline, D., Lowe. D. R., Meixner, W.F., Nouri, H. & Pearce, K. (2001). A research note dimensionality of Daly and Miller’s writing apprehension scale. *Written Communication, 18*(1), 61-79. doi: 10.1177/0741088301018001003

Bloom, B. S. (1969). Some theoretical issues relating to educational evaluation. In R. W. Tyler (Eds.), *Educational evaluation: new roles, new means: the 63rd yearbook of the National Society for the Study of Education (part II).* (Vol. 69(2), pp. 26-50). Chicago, IL: University of Chicago Press.

Bloom, B.S., Hastings, J.T. & Madaus, G.F. (Ed.) (1971). *Handbook on the Formative and Summative it into practice.* Buckingham: Open University Press.

Bloom, L. Z. (1980). *The composing processes of anxious and non-anxious writers: A naturalistic study. Paper presented at the annual meeting of the Conference on College Composition and Communication*. Retrieved from http://eric.ed.gov/?id=ED185559

Blue, G. (1994). Self-assessment of foreign language skills: Does it work? *CLE Working Papers, 3*, 18-35. Retrieved from http://files.eric.ed.gov/fulltext/ED396569.pdf

Boekaerts, M. (2006). Self-regulation and effort investment. In K. A. Renninger & I. E. Sigel (Ed.), *Handbook of child psychology: Volume 4–Child psychology in practice* (pp. 345-377). New York: Wiley.

Bond, T. & Fox, C. (2015). *Applying the Rasch Model Fundamental Measurement in the Human Sciences*. London: Routledge.

Bong, M. (2002). Predictive utility of subject-, task-, and problem-specific self-efficacy judgments for immediate and delayed academic performances. *The Journal of Experimental of Education, 70* (2), 133–162. doi: 10.1080/00220970209599503

Borg, S and Al-Busaidi, S (2012). *British Council ELT Research Paper. 12–07 Learner Autonomy: English Language Teachers' Beliefs and Practices*. London: British Council. Retrieved from http://www.teachingenglish.org.uk/sites/teacheng/files/b459%20ELTRP%20Report%20Busaidi_final.pdf

Boud, D. (1986). Facilitating Learning in Continuing Education: some important sources. *Studies in Higher Education, 11* (3), 237-243. doi: 10.1080/03075079.1986.10721161

Boud, D. (1990). Assessment and the promotion of academic values. *Studies in Higher Education, 15*(1), 101-111. doi: 10.1080/03075079012331377621

Boud, D. (1992). The use of self-assessment schedules in negotiated learning. *Studies in Higher Education, 17* (2), 185-200. doi: 10.1080/03075079212331382657

Boud, D. & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *Higher Education*, *18*, 529-549. Retrieved from http://link-springer-com.ez.wul.waseda.ac.jp/article/10.1007/BF00138746

Brentani, E. & Golia, S. (2007). Unidimensionality in the Rasch Model: How to Detect and Interpret. *STATISTICA*, *67*(3), 253-261. doi: 10.6092/issn.1973-2201/3508

Brookhart, S.M. (2003). Developing Measurement Theory for Classroom Assessment Purposes and Uses. *Educational Measurement: Issues and Practice, 22*(4), 5-12. doi: 10.1111/j.1745-3992.2003.tb00139.x

Brousseau, G. (1984). The crucial role of the didactical contract in the analysis and construction of situations in teaching and learning mathematics, In: H.-G. STEINER (Eds.), *Theory of Mathematics Education: ICME 5 topic area and miniconference* (pp. 110-119). Germany: Institut fur Didaktik der Mathematik der Universitat Bielefeld.

Brown, J.D. (1998). *New ways of classroom assessment*. Alexandria, VA: Teachers of English to Speakers of Other Languages.

Brown, H. (2005). Self-assessment of writing in independent language learning programs: The value of annotated samples. *Assessing Writing*, *10*(3), 174-191. doi: 10.1016/j.asw.2005.06.001

Brown, H. & Hudson, T. (1998). The Alternatives in Language Assessment. *TESOL quarterly, 32*(4), 653-675. doi: 10.2307/3587999

Brown, S. & Knight, P. (1994). *Assessing Learners in Higher Education*. London & New York: Rouledge Falmer Reader.

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology, 84*(4), 822-848. doi: 10.1037/0022-3514.84.4.822

Bryman, A. (2004). Qualitative research on leadership: A critical but appreciative review. *The Leadership Quarterly*, *15*, 729-769. doi: 10.1016/j.leaqua.2004.09.007

Buck, G. (1988). Testing listening comprehension in Japanese university entrance examinations. *JALT Journal*, *10*, 12-42. Retrieved from http://jalt-publications.org/sites/default/files/pdf-article/jj-10.1-art1.pdf

Burgoon, J., & Hale, J. L. (1983a). Dimensions of communication reticence and their impact on verbal encoding. *Communication Quarterly, 31*(4), 302-311. doi: 10.1080/01463378309369519

Burgoon, J., & Hale, J. L. (1983b). A research note on the dimensions of communication reticence. *Communication Quarterly, 31*(3), 238-248. doi: 10.1080/01463378309369510

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology, 79*(4), 474-482. doi:

10.1037/0022-0663.79.4.474

Butler, Y. G. & Lee, J. (2010). The effects of self-assessment among young learners of English. *Language Testing, 27*(1), 5-31. doi: 10.1177/0265532209346370

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Eds.), *Spoken and written language: Exploring orality and literacy* (pp. 35-53). Norwood, NJ: Ablex.

Chan, V., Spratt, M., Humphreys, G. (2002). Autonomous language learning: Hong Kong tertiary students' attitudes and behaviors. *Evaluation and Research in Education, 16*(1), 1-18. doi: 10.1080/09500790208667003

Chang, C. C., Su, J. A., Tsai, C. S., Yen, C. F. , Liu, J. H., & Lin, C. Y. (2015). Rasch analysis suggested three unidimensional domains for Affiliate Stigma Scale: additional psychometric evaluation. *J Clin Epideiol, 68*, 674-683. doi: 10.1016/j.jclinepi.2015.01.018

Chang, L. (2007). The influences of group processes on learners' autonomous beliefs and behaviors. *System, 35*(3)*,* 322-337. doi: 10.1016/j.system.2007.03.001

Chang, W-C & Chan, C. (1995). Rasch Analysis for Outcomes Measures: Some Methodological Considerations. *Arch Phys Med Rehabil, 76*, 934-939. Retrieved from http://www.archives-pmr.org/article/S0003-9993(95)80070-0/pdf

Charmaz, K. (1996). Grounded Theory. The search for Meanings–Grounded Theory. In J. A. Smith, R. Harre, & L. Van Langehove (Eds.), *Rethinking Methods in Psychology* (pp. 27-49). London: Sage Publications.

Charmaz, K. (2006). *Constructing grounded theory: a practical approach through qualitative analysis.* Thousand Oaks, CA: Sage Publications.

Cheng, Y-s. (2002). Factors associated with foreign language writing anxiety. *Foreign Language Annals*, *35*(6), 647-656. doi: 10.1111/j.1944-9720.2002.tb01903.x

Cheng, Y-s. (2004). A measure of second writing anxiety: scale development and preliminary validation. *Journal of Second Language Writing, 13*(4), 313-335. doi: 10.1016/j.jslw.2004.07.001

Cheng, Y., Horwitz, E., Schallert, E. (1999). Language anxiety: differentiating writing and speaking components. *Language Learning, 49*(3), 417-116. doi: 10.1111/0023-8333.00095

Cheng, W. & Warren, M. (1997). Having second thoughts: student perceptions before and after a peer assessment exercise. *Studies in Higher Education, 22*(2), 233-239. doi: 10.1080/03075079712331381064

Cheng, W. & Warren, M. (2005). Peer assessment of language proficiency. *Language Testing*, *22*(1), 93-121. doi: 10.1191/0265532205lt298oa

Cho, K. & MacArthur, C. (2010). Student revision with peer and expert. *Learning and Instruction, 20*, 328-338. doi: 10.1016/j.learninstruc.2009.08.006

Cho, K., Schunn, C.D., & Wilson, R.W. (2006). Validity and reliability of scaffolded peer assessment of writing from instructor and student perspectives. *Journal of Educational Psychology, 98*(4), 891-901. doi: 10.1037/0022-0663.98.4.891

Christie, F. & Derewianka, B. (2008). *School Discourse: Learning to Write Across the Years of Schooling*. London: Bloomsbury.

Clément, R. (1980). Ethnicity, contact and communicative competence in a second language. In Giles, H., Robinson, W. P., Smith, P. M. (Eds.), *Language: Social Psychological Perspectives* (pp. 147-154). Oxford: Pergamon.

Clément, R., Gardner, R. C., & Smythe, P. C. (1977). Motivational variables in second language acquisition: A study of Francophones learning English. *Canadian Journal of Behavioral Science, 9*, 123-133. Retrieved from http://www.researchgate.net/profile/Richard_Clement2/publication/216308568

Cohen, A. D. (1994). *Assessing language ability in the classroom.* Boston, MA: Heinle and Heinle.

Cohen, J. (1988). *Statistical power analysis for the behavioural sciences 2nd edition*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Colby-Kelly, C. & Turner, C.E. (2007). AFL research in the L2 classroom and evidence of usefulness: Taking formative assessment to the next level. *Canadian Modern Language Review, 64*(1), 9-38. doi: 10.3138/cmlr.64.1.009

Corbin, J. & Strauss, A. (2008) *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory, 3rd edition.* Thousand Oaks, CA: Sage Publications.

Corbin, J. & Strauss, A. (2015). *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. London: Sage Publications.

Cotterall, S. (1995). Readiness for autonomy: investigating learner beliefs. *System*, *23*(2), 195-205. doi: 10.1016/0346-251X(95)00008-8

Cotterall, S. (1999). Key Variables in Language Learning: what do learners believe about them? *System, 27*(4), 493-513. doi: 10.1016/S0346-251X(99)00047-0

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment.* Cambridge: Cambridge University Press.

Creswell, J. W. (2007). *Qualitative inquiry & research design: Choosing among five approaches*. Thousand Oaks, CA: Sage Publications.

Creswell, J. W. (2015). *A Concise Introduction to Mixed Methods Research*. LA: Sage Publications.

Creswell, J. W. & Clark, V. L. P. (2017). *Designing and Conducting Mixed Methods Research*. Thousand Oaks, CA: Sage Publications.

Crismore, A., Markkanen, R., & Steffensen, M. S. (1993). Metadiscourse in persuasive writing: a study of texts written by American and Finnish university students. *Written Communication*, 10*(1)*, 39-71. doi: 10.1177/0741088393010001002

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58(4),* 438-481. doi: 10.3102/00346543058004438

Cumming, A. (1998). Theoretical perspective on writing. *Annual Review of Applied Linguistics, 18*, 61-78. doi: 10.1017/S0267190500003482

Cumming, A., Kantor, R., & Powers, D. (2001). Scoring TOEFL essays and TOEFL 2000 prototype tasks: An investigation into raters' decision making and development of a preliminary analytic framework. *TOEFL Monograph Series, Report No. 22*. Retrieved from http://www.ets.org/research/policy_research_reports/publications/report/2001/icbg

Dafei, D. (2007). An exploration of the relationship between learner autonomy and English proficiency. In P. Robertson, P. & R. Nunn (Eds.), *Asian EFL Journal Press* (pp. 1-23). Retrieved from http://www.asian-efl-journal.com.

Daly. J. A. & Miller, M.D. (1975). The empirical development of an instrument of writing apprehension. *Research in the Teaching of English, 9*(3), 242-249. Retrieved from http://library.ncte.org/journals/rte/issues/v9-3/20067

Daly, J.A., & Shamo, W.G. (1976). Writing apprehension and occupational choice. *Journal of Occupational Psychology, 49*(1), 55-56. doi: 10.1111/j.2044-8325.1976.tb00329.x

Daly, J.A., & Shamo, W.G. (1978). Academic decisions as a function of writing apprehension. *Research in the Teaching of English, 12*(2), 119-126. Retrieved from https://library.ncte.org/journals/rte/issues/v12-2/17890

Daly, J. A., & Wilson, D. A. (1983). Writing apprehension, self-esteem, and personality. *Research in the Teaching of English, 17*(4), 327-341. Retrieved from http://library.ncte.org/journals/rte/issues/v17-4/15695

d'Anglejan, A., Harley, B. & Shapson, S. (1990). Student evaluation in a multidimensional core French curriculum. *The Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 47*(1), 106-124. doi: 10.3138/cmlr.47.1.106

Daugherty, R. (1996). In search of teacher assessment–its place in the National Curriculum assessment system of England and Wales. *The Curriculum Journal,* *7*(2), 137-152. doi: 10.1080/0958517960070202

Davis, B. (1997). Listening for differences: an evolving conception of mathematics teaching. *Journal for Research in Mathematics Education*, *28*, 355-376. doi: 10.2307/749785

Davis, B., A., Elder, C., Hill, K., Lumley, T. & McNamara, T. (1999). *Dictionary of language testing.* Cambridge: Cambridge University Press.

Davison, C. (2007). Views From the Chalkface: English Language School-Based Assessment in Hong Kong. *Language Assessment Quarterly, 4*(1), 37-68. doi: 10.1080/15434300701348359

Davison, F. & Henning, G. (1985). A self-rating scale of English difficulty: Rasch scalar analysis of items and rating categories. *Language Testing*, *2*, 164-179. doi: 10.1177/026553228500200205

Davison, C. & Leung, C. (2009). Current Issues in English Language Teacher-Based Assessment. *TESOL Quarterly, 43*(3), 393-415. doi:10.1002/j.1545-7249.2009.tb00242.x

Deci, E. L. (1995). *Why We Do what We Do: The Dynamics of Personal Autonomy*. New York: Putnam's Sons.

Deci, E. L., & Ryan, R. M. (1985). The general causality orientations scale: Self-determination in personality. *Journal of Research in Personality, 19*(2), 109-134. doi: 10.1016/0092-6566 (85)90023-6

Deci, E. L., & Ryan, R. M. (1991). Nebraska symposium on motivation: Vol. 38. Perspectives on motivation. In R. Dienstbier (Eds.) *A motivational approach to self: Integration in personality.* (pp. 237–288). Lincoln, NE: Univ. of Nebraska Press.

Deci, E. L. & Ryan, R. M. (2000). The "what" and "why" of goal pursuits: Huma needs and the self-determination of behaviour. *Psychological Inquiry, 11*(4), 227-268. doi: 10.1207/S15327965PLI1104_01

Deci, E. L. & Ryan, R. M. (2006). Self-Regulation and the Problem of Human Autonomy: Does Psychology Need Choice, Self-Determination, and Will? *Journal of Personality,* 1557-1586. doi: 10.1111/j.1467-6494.2006.00420.x

Dennis, B., Lowe, D., Meixner,W., Nouri,H., & Pearce, K. (2001). A research note on the dimensionality of Daly and Miller's writing apprehension scale. *Written Communication, 18*(1), 61-79. doi: 10.1177/0741088301018001003

DePascale, C. A. (2003). *Putting Large-Scale Assessment in Perspective: The Ideal Role of Large-Scale Assessment in a Comprehensive Assessment System.* Paper presented at the 2003 Annual Meeting of the American Educational Research Association, Chicago: Illinois.

Dickinson, L. (1987). *Self-instruction in language learning*. Cambridge: Cambridge University Press.

Dieten, A. J. (1989). The development of a test of Dutch as a second language: the validity of self-assessment by inexperienced subjects. *Language Testing, 6* (1), 30-46. doi: 10.1177/026553228900600105

Dörnyei, Z. (2007) Creating a motivating classroom environment. In J. Cummins and C. Davison (Eds.), *International Handbook of English Language Teaching (Vol. 2)* (pp. 719-731). New York: Springer.

Dörnyei, Z. & Malderez, A. (1999). The role of group dynamics in foreign language learning and teaching. In: Arnold, J. (Eds.), *Affect in Language Learning* (pp. 155-169). Cambridge University Press: Cambridge.

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

Earl, S. E. (1986). Staff and Peer Assessment-Measuring An Individual's Contribution To Group Performance. *Assessment & Evaluation in Higher Education, 11*(1), 60-69. doi: 10.1080/0260293860110105

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments*. Frankfurt am Main: Peter Lang Edition.

Edmondson, A. C. (1999). Psychological safety and learning behaviour in teams. *Administrative Science Quarterly, 44*, 250-282. doi: 10.2307%2F2666999

EIKEN. (2018). EIKEN Grades. Retrieved from http://www.eiken.or.jp/eiken/exam/about/

EIKEN. (2018). *San-kyu no kakomon to taisaku* [Previous questions of the third grade of EIKEN]. Retrieved from http://www.eiken.or.jp/eiken/exam/grade_3/solutions.html

Ellis, R. (2003). *Task-based language learning and teaching.* Oxford: Oxford University Press.

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

Esfandiari, R. & Myford, C. (2013). Severity differences among self-assessors, peer-assessors, and teacher assessors rating EFL essays. *Assessing Writing, 18*(2), 111-131. doi: 10.1016/j.asw.2012.12.002

Evans, A. W., McKenna, C., & Oliver, M. (2005). Trainees' perspectives on the assessment and self-assessment of surgical skills. *Assessment & Evaluation in Higher Education*, *30*(2), 163-174. doi: 10.1080/0260293042000264253

Eysenck, M.W. (1979). Anxiety, learning and memory: A reconceptualization. *Journal of*

Research in Personality, 13, 363-385. doi: 10.1016/0092-6566(79)90001-1

Faigley, L., Daly, J.A., & Witte, S.P. (1981). The role of writing apprehension in writing performance and competence. *Journal of Educational Research, 75*, 16-21. doi: 10.1080/00220671.1981.10885348

Falchikov, N. (1995). Peer feedback marking: developing peer assessment. *Innovations in Education and Training International, 32*, 175–187. doi :10.1080/1355800950320212

Falchikov, N. & Boud, D. (1989). Student Self-Assessment in Higher Education: A Meta-Analysis. *Review of Educational Research, 59*(4), 395-430. doi: 10.3102%2F00346543059004395

Falchikov, N. & Goldfinch, J. (2000). Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks. *Review of Educational Research, 70*(3), 287-322. doi :10.3102%2F00346543070003287

Filed, A. (2009). *Discovering Statistics Using SPSS*. London: SAGE Publications.

Finn, B. & Metcalfe, J. (2010). Scaffolding feedback to maximize long term error correction. *Memory and Cognition, 38*(7), 951-961. Retrieved from http://link.springer.com/article/10.3758/MC.38.7.951

Fisher, W. R. (1985). The narrative paradigm: An elaboration. *Communication Monographs, 52*, 347-367. doi: 10.1080/03637758509376117

Fischer, G. H., & Scheiblechner, H. (1970). Algorithmen und Programme fuer das probabilistische Testmodell von Rasch. *Psychologische Beitrage, (12)*, 23–51. Retrieved from https://psycnet.apa.org/record/1971-07228-001

Flick, U. (2014). *An Introduction to Qualitative Research*. London: Sage Publications.

Freeman, M. (1995). Peer assessment by groups of group work. *Assessment & Evaluation in Higher Education, 20*(3), 289-300. doi: 10.1080/0260293950200305

Fremouw, W.J., & Breitenstein, J.L. (1990). Speech anxiety. In H. Leitenberg (Eds.), *Handbook of social and evaluation anxiety* (pp.455-474). New York: Plenum Press.

Friedman, M. (2000). Education for world citizenship. *Ethics, 110*(3), 586-601. doi: 10.1086/233325

Fulcher, G. (2010). *An introduction to Assessment for Learning*. Retrieved from http://languagetesting.info/features/afl/formative.html

Gardner, R.C. (1985). *Social Psychology and Language Learning*. London: Edward Arnold.

Garrison, C., & Ehringhaus, M. (2009). Formative and summative assessment in the classroom. *National Middle School Association, 1–3.* Retrieved from: http://schools.nyc.gov/NR/rdonlyres/33148188-6FB5-4593-A8DF-

8EAB8CA002AA/0/2010_ 11_Formative_Summative_Assessment.pdf

Garrison, C., & Ehringhaus, M. (2013). Formative and Summative Assessments in the Classroom. Retrieved: from http://www.amle.org/BrowsebyTopic/WhatsNew/WNDet/TabId/270/ArtMID/888/ArticleID/286/Formativeand-Summative-Assessments-in-the

Genesee, F. & Upshur, J. (1996). *Classroom-based evaluation in second language education.* Cambridge: Cambridge University Press.

Glaser, B. (1978). *Theoretical sensitivity*. Mill Valley, CA: Sociology Press.

Glaser, B. G. & Strauss, A.L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research.* Chicago, IL: Aldine.

Goto, K. (2005). Oral Communication Yearly plan. *Eigo-kyoiku, 54*(1) *April*, 25-28. Retrieved from http://ci.nii.ac.jp/naid/40006662199/

Good, T. L. & Grouws, D.A. (1975). *Process-product relationships in fourth grade mathematics classrooms (Grand No. NEG-00-3-0123)*. Columbia: University of Missouri.

Goodman, S. G. & Cirka, C. C. (2009). Efficacy and anxiety: An examination of writing attitudes in a first-year seminar. *Journal on Excellence in College Teaching, 20*(3), 5-29. Retrieved from http://www.researchgate.net/profile/Carol_Cirka/publication/264496818_

Graesser, A. C., Pearson, N.K., & Magliano, J. P. (1995). Collaborative dialogue patterns in naturalistic one-to-one tutoring. *Applied Cognitive Psychology, 9*(6), 495-522. doi: 10.1002/acp.2350090604

Grant, L., & Ginther, L. (2000). Using computer-tagged linguistic features to describe L2 writing differences. *Journal of Second Language Writing*, *9*, 123-145. doi: 10.1016/S1060-3743(00)00019-9

Greenwald, A. G. (1982). Ego task analysis: An integration of research on ego-involvement and self-awareness. In A. Hastorf & A. M. Isen (Eds.), *Cognitive social psychology* (pp.109-147). New York: Elsevier.

Gregersen, T. & Horwitz, E. (2002). Language Learning and Perfectionism: Anxious and Non-Anxious Language Learners’ Reactions to Their Own Oral Performance. *The Modern Language Journal, 86*(4), 562-570. doi: 10.1111/1540-4781.00161

Guiora, A. Z. (1983). Language and concept formation: a cross-lingual analysis. *Cross-Cultural Research, 18*(3), 228-256. doi: 10.1177%2F106939718301800304

Güneyli, A. (2016). Analyzing Writing Anxiety Level of Turkish Cypriot Students. *Education and Science, 41*, 163-180. Retrieved from http://www.researchgate.net/profile/Ahmet_Guneyli/publication/295258722

Gungle, B. W., & Taylor, V. (1989). Writing apprehension and second language writers. In D. M. Johnson & D. H. Roen (Eds.), *Richness in writing: Empowering ESL students* (pp. 235-249). New York: Longman.

Halliday, M. A. K. (2005). Spoken and written modes of meaning. In Graddol, D. & Boyd-Barrett, O (Eds.), *Media Texts: Authors and Readers* (pp. 41-58). London: The Open University.

Hanrahan, S. & Isaacs, G. (2001). Assessing self- and peer assessment: the students' views. *Higher Education Research and Development*, *20*(1), 53-70. doi: 10.1080/07294360123776

Harlen, W. (1996). Editorial. *The Curriculum Journal, 7*(2), 129. doi: 10.1080/0958517960070201

Harlen, W. & Winter, J. (2004). The development of assessment for learning: learning from the case of science and mathematics. *Language Testing, 21*(3), 390-408. doi: 10.1191%2F0265532204lt289oa

Harrell, M. C. & Bradley, M. A. (2009). *Data Collection Methods. Semi-Structured Interviews and Focus Groups.* Santa Monica, CA: Rand National Defense Research Institute.

Harvey, L. & Knight, P. (1996). *Transforming Higher Education*. London: Society for Research into Higher Education, Ltd.

Hassan, B. A. (2001). The Relationship of Writing Apprehension and Self-Esteem to the Writing Quality and Quantity of EFL University Students. *Mansoura Faculty of Education Journal, 39*, 1-36. Retrieved from http://files.eric.ed.gov/fulltext/ED459671.pdf

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: Theories, methods, individual differences, and applications* (pp. 1–27). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heine, S.J. & Hamamura, T. (2007). In Search of East Asian Self-Enhancement. *Personality and Social Psychology Review, 11*(4), 4-27. doi: 10.1177%2F1088868306294587

Heywood, J. (2000). *Assessment in Higher Education Student Learning, Teaching, Programmes and Institutions*. London: Jessica Kingsley Publishers.

Hiratsuka, T. (2014). A study into how high school students learn using narrative frames. *ELT Journal Volume, 68*(2), 169-178. doi: 10.1093/elt/cct096

Ho, J. & Crookall, D. (1995). Breaking with Chinese cultural traditions: learner autonomy in English language teaching. *System, 23*(2), 235-243. doi: 10.1016/0346-

251X(95)00011-8

Holec, H. (1981). *Autonomy and foreign language learning*. Oxford: Pergamon. (First published [1979], Strasbourg: Council of Europe.

Holec, H. (ed.) (1989). *Autonomy and Self-Directed Learning*: *Present Fields of Application.* Strasbourg: Council of Europe.

Holland, M. (1978). *Studies of students in UCLA's composition courses: A final report.* (Unpublished manuscript). University of California, Los Angeles.

Horwitz, E. K. (1986). Preliminary evidence for the reliability and validity of a foreign language anxiety scale. *TESOL Quarterly, 20*, 559-562. doi: 10.2307/3586302

Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics, 21*, 112-126. doi: 10.1017/S0267190501000071

Horwitz, E. K., Horwitz, M .B., & Cope, J.A. (1986). Foreign language classroom anxiety. *Modern Language Journal, 70*(2), 125-132. doi: 10.2307/327317

Hu, M. & Liu, B. (2004). Mining and summarizing Customer Reviews. *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, 168–177. doi: 10.1145/1014052.1014073

Huerta-Macias, A. (1995). Alternative assessment: Responses to commonly asked questions. *TESOL Journal, 5*, 8-10. Retrieved from http://books.google.co.jp/books?hl=ja&lr=lang_ja|lang_en&id=VxnGXusQlI8C&oi=fnd&pg=PA338&dq

Hung, Y., Samuelson, B. L., & Cheng, S. (2016). The Relationship between Peer and Self-Assessment and Teacher Assessment of Young EFL Learners' Oral Presentations. In M. Nikolov (Eds.), *Assessing Young Learners of English: Global and Local Perspectives* (pp. 317-338). New York: Springer.

Hyland, K. (1996). *Second Language Writing*. Cambridge: Cambridge University Press.

Jacobs, H. L., Zingraf, S., Wormuth, D. R., Hartfiel, V. F., & Hughey, J. B. (1981). *Testing ESL composition: A practical approach. Rowley.* MA: Newbury House.

Jonsson, A., and G. Svingby. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130–44. doi: 10.1016/j.edurev.2007.05.002

Kean, D., Gylnn, S., & Britton, B. (1987). Writing persuasive documents: The role of students' verbal aptitude and evaluation anxiety. *Journal of Experimental Education, 55*, 95-102. doi: 10.1080/00220973.1987.10806440

Kim, J.-H. (2000). *Foreign language listening anxiety: A study of Korean students learning English*. (Unpublished doctoral dissertation). The University of Texas, Austin. Retrieved from http://www.researchgate.net/publication/35732188_Foreign_language_listening_anxiety_a_study_of_Korean_students_learning_English

Kim, K.J. (2009).Motivational Challenges of Adult Learners in Self-Directed e-Learning. *Journal of Interactive Learning Research, 20*(3), 317-335. Retrieved from http://www.semanticscholar.org/paper/Motivational-Challenges-of-Adult-Learners-in-Kim/370ed97d6f5989e103db7dec9a15265bb17518ec

Kim, S.-Y., & Kim, W.-K. (2005). An investigation into learner autonomy in relation to practical English skills. *English Language Teaching, 17*(3), 107-129. Retrieved from http://www.ccsenet.org/journal/index.php/elt/issue/archives

King, B. M. & Minium, E. M. (2003). *Statistical Reasoning In Psychology and Education*. Fourth Edition. New York: Wiley.

Kondo-Brown, K. (2002). A FACETS analysis of rater bias in measuring Japanese second language writing performance. *Language Testing, 19*(1), 3-31. doi: 10.1191%2F0265532202lt218oa

Krashen, S.D. (1985). *The input hypothesis: Issues and implications.* New York: Longman.

Kuiken, F., & Vedder, I. (2008). *The influence of task complexity on linguistic performance in L2 writing and speaking: the effect of mode. Cognitive approaches to second/foreign language processing: theory and pedagogy.* Paper presented at 33rd International LAUD Symposium, Landau/Pfalz, Germany. (pp. 386-390). Retrieved from http://pure.uva.nl/ws/files/4267252/64168_294147.pdf

Kwan, K. & Leung, R. (1996). Tutor Versus Peer Group Assessment of Student Performance in a Simulation Training Exercise. *Assessment & Evaluation in Higher Education, 21*(3), 205-214. doi: 10.1080/0260293960210301

Landau, J. K., Vohs, J. R., & Romano, C. A. (1999). *Statewide Assessment: Policy Issues, Questions, and Strategies. PEER Information Brief.* Boston, MA: The Federation for Children with Special Needs.

Lang, P.J. (1971). The application of psychophysiological methods to the study of psychotherapy and behavior modification. In A.E. Bergin & S.L. Garfield (Eds.), *Handbook of psychotherapy and behavior change* (pp. 75-125). New York: Wiley.

Larson-Hall, J. (2008). Weighing the benefits of studying a foreign language at a younger starting age in a minimal input situation. *Second Language Research, 24*(1), 35-63. doi:10.1177%2F0267658307082981

Larson-Hall, J., & Herrington, R. (2010). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368– 390. doi:10.1093/applin/amp038

Leach, L. (2012). Optional self-assessment: Some tensions and dilemmas. *Assessment & Evaluation in Higher Education, 37*(2), 137-147. doi:10.1080/02602938.2010.515013

Léger, D. (2009). Self-Assessment of Speaking Skills and Participation in a Foreign Language Class. *Foreign Language Annals, 42*(1), 158-178. doi:10.1111/j.1944-9720.2009.01013.x

Leahy, S., Lyon, C., Thompson, M., and Wiliam, D. (2005). Classroom assessment: Minute-by-minute, day-by-day. *Educational Leadership, 63*(3), 18-24. Retrieved from http://www.rbteach.com/sites/default/files/classroom-assessment-minute-by-minute-day-by-day.pdf

Leahy, S., & Wiliam, D. (2012). From teachers to schools: Scaling up professional development for formative assessment. In J. Gardner (Eds.), *Assessment and learning* (pp. 49-72). London: Sage Publications.

Lee, H. (2017). The Effects of University English Writing Classes Focusing on Self and Peer Review on Learner Autonomy. *The Journal of Asia TEEL, 14*(3), 464-481. Retrieved from http://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.18823%2Fasiatefl.2017.14.3.6.464

Lee, S.-Y. (2001). The relationship of writing apprehension to the revision process and topic preference: A student perspective. In P.-H. Chen & Y.-N. Leung (Eds.), *Selected papers from the tenth international symposium on English teaching* (pp. 504-516). Taipei, Taiwan: Crane.

Leki, I. (1991). The preferences of ESL students for error correction in college-level writing classes. *Foreign Language Annals, 24*. 203- 18. doi:10.1111/j.1944-9720.1991.tb00464.x

Leki, I. (1999). Techniques for reducing second language writing anxiety. In D. J. Young (Eds.), *Affect in foreign language and second language learning: A practical guide to creating a low-anxiety classroom atmosphere* (pp. 64-88). Boston: McGraw-Hill.

Leung, C. & Mohan, B. (2004). Teacher formative assessment and talk in classroom contexts: assessment as discourse and assessment of discourse. *Language Testing, 21*(3), 335-359. doi:10.1191%2F0265532204lt287oa

Lew, M, Alwis, W., & Schmidt, H. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education, 35*(2), 135-156. doi:10.1080/02602930802687737

Liebert, R. M. & Morris, L. M. (1967). Cognitive and emotional component of test anxiety: a distinction and some initial data. *Psychological Reports, 20*, 975-978.

doi:10.2466/pr0.1967.20.3.975

Linacre, J. M. (1989). *Many-facet Rasch measurement.* Chicago: II: MESA Press.

Linacre, J. M. (1993). *Generalizability Theory and Many-Facet Rasch Measurement*. Paper presented at the 1993 Annual Meeting of the American Educational Research Association Atlanta, Georgia. Retrieved from http://files.eric.ed.gov/fulltext/ED364573.pdf

Linacre, J. M. (1998). Structure in Rasch residuals: why principal component analysis? *Rasch measurement transactions, 21*(2), 636. Retrieved from http://www.rasch.org/rmt/rmt122m. htm.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878. Retrieved from http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. M. (2014). *A user's guide to Winsteps Ministep Rasch-model computer programs (3.81.0)*. Retrieved from http://www. winsteps. com/winman/reliability. htm

Linacre, J. M. & Wright, T. F. (1993). *A user's guide to FACETS: Rasch-measurement computer program. Version 2.62.* Chicago, IL:MESA Press.

Little, D. (1991). *Learner autonomy 1: Definitions, issues and problems*. Dublin: Authentik.

Little, D. (1995). Learning as dialogue: The dependence of learner autonomy on teacher autonomy. *System, 23*(2), 175-181. doi:10.1016/0346-251X(95)00006-6

Little, D. (2002). Learner autonomy and second/foreign language learning, *Good Practice Guide. LTSN Subject Centre for Languages, Linguistics and Area Studies*. Retrieved from http://www.lang.ltsn.ac.uk/ resources/goodpractice.aspx?resourceid=1409

Little, D. (2009). Language learner autonomy and the European Language Portfolio: Two L2 English examples. *Language Teaching, 42*(2), 222-233. doi:10.1017/S0261444808005636

Littlewood, W. (1999). Defining and developing autonomy in East Asian context. *Applied Linguistics, 20*(1), 71-94. doi:10.1093/applin/20.1.71

Liu, MF. & Carless, D. (2006). Peer feedback: The learning element of peer assessment. *Teaching in Higher Education, 11*(3), 279-290. doi:10.1080/13562510600680582

Logan, E. (2009). Self and peer assessment in action. *Practitioner Research in Higher Education, 3*(1), 29-35. Retrieved from http://194.81.189.19/ojs/index.php/prhe

Lord, E. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin, 72*, 304-305. doi:10.1037/h0025105

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*(1), 54-71. doi:10.1177%2F026553229501200104

Lundstrom, K. & Baker, W. (2009). To give is better than to receive: the benefits of peer review to the reviewer's own writing. *Journal of Second Language Writing, 18*, 30-43. doi:10.1016/j.jslw.2008.06.002

Lynch, B. K. & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15* (2), 158-180. doi:10.1177%2F026553229801500202

MacIntyre, P.D. (1992). *Anxiety And Language Learning From A Stages Of Processing Perspective* (Doctoral dissertation). The University of Western Ontario, Canada. Retrieved from http://ir.lib.uwo.ca/cgi/viewcontent.cgi?article=3154&context=digitizedtheses

MacIntyre, P.D. (1999). Language Anxiety: a review of research for language teachers. In: Young, D.J. (Eds.), *Affect in Foreign Language and Second Language Learning: a Practical Guide to Creating a Low Anxiety Classroom Atmosphere*. Boston: McGraw-Hill College.

MacIntyre, P.D. , & Gardner, R. C. (1989). Anxiety and second language learning: Toward a theoretical clarification. *Language Learning, 39*(2), 251-275. doi:10.1111/j.1467-1770.1989.tb00423.x

MacIntyre, P.D., & Gardner, R. C. (1991).Methods and results in the study of anxiety and language learning: A review of the literature. *Language Learning, 4*. 85-117. Retrieved from http://faculty.cbu.ca/pmacintyre/research_pages/journals/methods_results1991.pdf

MacIntyre, P. D., Clément, R., Dörnyei, Z., & Noels, K. (1998). Conceptualizing willingness to communicate in a L2: A situated model of confidence and affiliation. *Modern Language Journal, 8*2(4), 545–562. doi:10.2307/330224

MacIntyre, P. D., Noels, K. A., & Clément, R. (1997). Biases in self-ratings of second language proficiency: The role of language anxiety. *Language Learning, 47*, 265-287. Retrieved from http://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1111%2F0023-8333.81997008

Madignan, R., Linton,P., & Johnson, S. (1996). The paradox of writing apprehension. In L. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 295-307).

Hillsdale, NJ: Lawrence Erlbaum.

Mandinach, E. B., & Gummer, E. (2015). Data-driven decision making: Components of the enculturation of data use in education. *Teachers College Record, 117*(4). Retrieved from http://cdn.tc-library.org/Rhizr/Files/FkE9DdrKdtH7PAQaw/files/Mandinach.pdf

Masny, D., & Foxall, J. (1992). *Writing apprehension in L2*. Retrieved from https://eric.ed.gov/?id=ED352844

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174. Retrieved from https://link.springer.com/article/10.1007/BF02296272

Matsuno, S. (2009). *Self-, peer-, and teacher-assessments in Japanese university EFL writing classrooms. Language Testing, 26*(1), 75-100. doi:10.1177%2F0265532208097337

McCroskey. J.C. (1970). Measures of communication-bound anxiety. *Speech Monographs, 37*, 269-277. doi.org/10.1080/03637757009375677

McDonald, B. & Boud, D. (2003). The Impact of Self-assessment on Achievement: The effects of self-assessment training on performance in external examinations. *Assessment in Education: Principles, Policy & Practice, 10*(2), 209-220. doi:10.1080/0969594032000121289

McKain, T. L. (1991). *Cognitive, affective, and behavioural factors in writing anxiety* (Doctoral dissertation). Washington, D.C., Catholic University of America. Retrieved from http://www.worldcat.org/title/cognitive-affective-and-behavioral-factors-in-writing-anxiety/oclc/25196822

McLaughlin, P. & Simpson, N. (2004). Peer assessment in first year university: How the students feel. *Studies in Educational Evaluation, 30*(2), 135-149. doi:10.1016/j.stueduc.2004.06.003

McMillan, J. & Hearn, J. (2008). Student Self-Assessment: The Key to Stronger Student Motivation and Higher Achievement. *Educational Horizons, 87*(1), 40-49. Retrieved from http://www.jstor.org/stable/42923742

McNamara, T. (1996). *Measuring second language performance*. London: Longman.

Messick, S. (1989). Validity. In R. L. Linn (Eds.), *The American Council on Education/Macmillan series on higher education. Educational measurement* (pp. 13-103). New York: Macmillan.

Ministry of Education, Culture, Sports, Science and Technology. (2003). *Regarding the establishment of an action plan to cultivate "Japanese with English abilities."* Retrieved from http://mext.go.jp/english/topics/03072801.htm

Ministry of Education, Culture, Sports, Science and Technology. (2010). *The Designated*

"Course of Study" for Senior High Schools. Retrieved from http://www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/__icsFiles/afieldfile/2012/10/24/1298353_3.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2013). *The second basic plan for the promotion of education*. Retrieved from http://www.mext.go.jp/english/lawandplan/135530.htm

Ministry of Education, Culture, Sports, Science and Technology & National Institute for Educational Policy Research (NIER). (2015). *Heisei 27 nendo zenkoku gakuryoku gakushuu joukyou chosa houkokusho [2015 Academic year report from the National Academic Ability and Situation Assessment]*. Retrieved from http://www.nier.go.jp/15chousakekkahoukoku

Ministry of Education, Culture, Sports, Science and Technology. (2018a). *Koutougakkou gakushu shidouyouryou (Heisei 30nenndo kokuji) [The Designated "Course of Study" for Senior High Schools].* Retrieved from http://www.mext.go.jp/content/1384661_6_1_3.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2018b). *Eigo kyoiku jisshi chosa [The National Survey of English Education Enforcement Situation]*. Retrieved from http://www.mext.go.jp/component/a_menu/education/detail/__icsFiles/afieldfile/2018/04/06/1403469_02.pdf

Ministry of Education, Culture, Sports, Science and Technology. (2018c). *Heisei 29nenndo eigoryoku chousa kekka [The Results of Fiscal 2007 Survey of Proficiency in English]*. Retrieved from http://www.mext.go.jp/a_menu/kokusai/gaikokugo/__icsFiles/afieldfile/2018/04/06/1403470_01_1.pdf

Minium, E. W. & King, B. M. (2002). *Statistical Reasoning In Psychology And Education Fourth Edition*. New Jersey: John Wiley & Sons, Inc.

Mizumoto, A. & Takeuchi, O. (2008). Basics and Considerations for Reporting Effect Sizes in Research Papers. *Eigo Kyoiku Kenkyu, 31*, 57-66. Retrieved from http://ci.nii.ac.jp/naid/120005685973

Morris, L. W. & Engle, W. B. (1981). Assessing various coping strategies and their effects on test performance and anxiety. *Journal of Clinical Psychology, 37*(1), 165-171. doi:10.1002/1097-4679(198101)37:1%3C165::aid-jclp2270370133%3E3.0.co;2-8

Morris, L. W., Davis, M.A., & Hutchings, C.H. (1981). Cognitive and emotional components of anxiety: literature review and a revised worry-emotionality scale.

J.Educ.Psychol, *73*, 541-555. doi:10.1037/0022-0663.73.4.541

Myford, C. M. & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part 1. *Journal of Applied Measurement, 4*, 386-422. Retrieved from http://psycnet.apa.org/record/2003-09517-007

Nadamitsu, Y., Asai, A., & Koyanagi, S. (2014). Shitsuteki kenkyu nitusite kangaeru [A comparison of three qualitative research methods: Grounded Theory Approach, Narrative Analysis, and Action Research]. *Cross-cultural communication, 12*, 67-84.　灘光洋子, 浅井亜紀子, 小柳志津著。「質的研究について考える」異文化コミュニケーション論集。doi/10.14992/00011112

Nanine, A. E., van Gennip, Segers, & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of international variables and conceptions. *Learning and Instruction, 20*(4), 280-290. Retrieved from http://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1016%2Fj.learninstruc.2009.08.010

Negari, G. M. & Rezaabadi, O. T. (2012). Too Nervous to Write? The Relationship between Anxiety and EFL Writing. *Theory and Practice in Language Studies, 2*(12), 2578-2586. doi:10.4304/TPLS.2.12.2578-2586

Nicol, D. & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *High Education, 31*(2)*,* 199-218. doi:10.1080/03075070600572090

Niemiec, C. P., & Ryan, R. M. (2009). Autonomy, competence, and relatedness in the classroom: Applying self-determination theory to educational practice. *Theory and Research in Education, 7*, 133–144. doi:10.1177%2F1477878509104318

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

Nyquist, J. B. (2003). *The benefits of reconstructing feedback as a larger system of formative assessment: A meta-analysis.* (Unpublished Master of Science thesis). Vanderbilt University, Nashville, TN. Retrieved from http://www.worldcat.org/title/benefits-of-reconstruing-feedback-as-a-larger-system-of-formative-assessment-a-meta-analysis/oclc/174149300&referer=brief_results#borrow

Oi. S. Y. (2018). The Relationship between Writing Tasks in Textbooks and CAN-DO Lists in terms of Task Difficulty. *Journal of Pan-Pacific Association of Applied Linguistics, 22*(2), 53-70. Retrieved from http://files.eric.ed.gov/fulltext/EJ1201793.pdf

Oi, S. Y. (2019a). Japanese High School English Teachers' Perspectives on Classroom

Writing Assessment Criteria: A Needs Analysis. *The bulletin of the Graduate School of Education of Waseda University. Separate volume*, *27*(1), 159-176. Retrieved from http://hdl.handle.net/2065/00063295

Oi, S. Y. (2019b). *How do Self- & Peer assessment have Effect on High School students' Writing Anxiety?* Unpublished manuscript, Waseda University, Tokyo.

O'Malley, J. M. & Chamot, A. U. (1990). *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.

O'Malley, J.M. & Valdez Pierce, L. (1996). *Authentic assessment for English language learners: practical approaches for teachers.* New York, NY: Addison-Wesley.

Orsmond, P., Merry, S., & Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. *Assessment & Evaluation in Higher Education*, *21*(3), 239-250. doi:10.1080/0260293960210304

Orsmond, P., Merry, S., & Reiling, K. (2000). The Use of Student Derived Marking Criteria in Peer and Self-assessment. *Assessment & Evaluation in Higher Education,* *25*(1), 23-38. Retrieved from http://www.researchgate.net/deref/http%3A%2F%2Fdx.doi.org%2F10.1080%2F02602930050025006

Oxford Languages. (2010). *Oxford Dictionary of English*. Oxford: Oxford University Press.

Peirce, B.N., Swain, M., & Hart, D. (1993). Self-Assessment, French Immersion, and Locus of Control. *Applied Linguistics, 14*(1), 25-42. Retrieved from http://faculty.educ.ubc.ca/norton/Norton%20Peirce,%20Swain,%20&%20Hart%201993.pdf

Perrenoud, P. (1998). From Formative Evaluation to a Controlled Regulation of Learning Processes. Towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice, 5*(1), 75-102. doi:10.1080/0969595980050105

Philips, E.M. (1992). The effects of language anxiety on students' oral test performance and attitudes. *Modern Language Journal, 76*, 14-26. doi:10.2307/329894

Poehner, M. (2008). *Dynamic Assessment A Vygotskian Approach to Understanding and Promoting L2 Development*. New York: Springer.

Pope, N. (2001). An examination of the use of peer rating for formative assessment in the context of the theory of consumption values. *Assessment & Evaluation in Higher Education*, *26*(3), 235-246. doi:10.1080/02602930120052396

Proulx, P. (1991). Anxiety in language learning: Recognition and prevention. *Canadian Journal of Native Education, 18*, 53-64. doi:10.1111%2F0023-8333.00095

Qiu, G., Liu, B., Bu, J. and Chen, C. (2011). Opinion Word Expansion and Target

Extraction through Double Propagation. *Computational Linguistics, 37*(1)*, 9-27.* doi:10.1162/coli_a_00034

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13. Retrieved from http://www.researchgate.net/profile/Arkalgud_Ramaprasad/publication/227634769_On_the_Definition_of_Feedback/links/59fa08ceaca272026f6ed416/On-the-Definition-of-Feedback.pdf

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research.

Ravand, S. & Ravand, H. (2016). Investigating the Effect of Self-, Peer, and Teacher Assessment in Second Language Writing Over Time: A Multifaceted Rasch Approach. *Iranian Journal of Applied Language Studies, 8*(1), 97-112. doi:10.22111/ijals.2016.3022

Rea-Dickins, P. (2000). Mirror, mirror on the wall: identifying processes of classroom assessment. *Language Testing, 18*, 429-462. doi:10.1177%2F026553220101800407

Rea-Dickins, P. (2004). Editorial Understanding teachers as agents of assessment. *Language Testing, 21*(3), 249-258. doi:10.1191%2F0265532204lt283ed

Rea-Dickins, P. & Gardner, S. (2000). Snares and Silver bullets: disentagling the construct of formative assessment. *Language Testing, 17*(2), 215-243. doi:10.1177%2F026553220001700206

Rose, M. (1984). *Writer's block: The cognitive dimension*. Carbondale: Southern Illinois University Press.

Ross, A. (2006). The reliability, validity, and utility of self-assessment. *Practical Assessment, Research & Evaluation, 11*(10), 1-13. doi:10.7275/9wph-vv65

Ross, J. S. (1998). Self-assessment in second language testing: a meta-analysis and analysis of experiential factors. *Language Testing, 15*(1), 1-20. doi:10.1177%2F026553229801500101

Rossa, J. A., Rolheisera, C., Hogaboam-Graya, A. (1998). Skills training versus action research in-service: impact on student attitudes to self-evaluation. *Teaching and Teacher Education, 14*(5), 463-477. doi:10.1016/S0742-051X(97)00054-1

Runnels, J. (2014). Japanese English Learner Self-assessments on the CEFR-J's A-level Can-do Statements Using Four and Five-Point Response Scales. *The Asian Journal of Applied Linguistics*, *1*(2), 167-77. Retrieved from http://www.academia.edu/8958707/Japanese_English_learner_self_assessments_on_the_CEFR_J_s_A_level_can_do_statements_using_four_and_five_point_respo nse_scales

Rushton, C., Ramsey, P. & Rada, R. (1993). Peer assessment in a collaborative hypermedia environment: a case study. *Journal of Computer-Based Instruction*, *20*(3), 75. Retrieved from http://eric.ed.gov/?id=EJ476366

Ryan, R. M, & Deci, E. L. (2006). Self-regulation and the problem of human autonomy: does psychology need choice, self-determination, and will? *J Pers, 74*, 1557–85. doi:10.1111/j.1467-6494.2006.00420.x

Ryan, R. M. & Deci, E. L. (2017). *Self-Determination Theory*. The Guilford Press. NY.

Sadler, R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119-144. Retrieved from http://link.springer.com/article/10.1007%252FBF00117714

Sadler, P. R. & Good, E. (2006). The impact of self- and peer-grading on student learning. *Educational Assessment, 11*(1), 1-31. Retrieved from http://www.cfa.harvard.edu/sed/staff/Sadler/articles/Sadler%20and%20Good%20EA.pdf

Saiki, C. S. (2018). Grounded theory approach. Tokyo: Shinyosha.「グランデッド・セオリー・アプローチ　改訂版　理論を生みだすまで」戈木クレイグヒル滋子著。新曜社。

Saito, H. (2008). EFL classroom peer assessment: Training effect on rating and commenting. *Language Testing, 25*(4). 553-581. doi: 10.1177%2F0265532208094276

Saito, H. & Fujita, T. (2004). Characteristics and user acceptance of peer rating in EFL writing classrooms. *Language Teaching Research*, *8*(1), 31-54. doi: 10.1191%2F1362168804lr133oa

Saito, H. & Fujita, T. (2009). Peer-assessing peers' contribution to EFL group presentation. *RELC Journal, 40*(2), 149-171. doi: 10.1177%2F0033688209105868

Saito, H. & Inoi, S. (2017). Junior and Senior High School EFL Teachers' Use of Formative Assessment: A Mixed-Methods Study. *Language Assessment Quarterly, 14*(3), 213-233. doi:10.1080/15434303.2017.1351975

Sarason, I. G., & Ganzer, V. J. (1962). Anxiety, reinforcement, and experimental instructions in a free verbalization situation. *Journal of abnormal and Social Psychology*, *65*, 300-307. doi: 10.1037/h0048977

Sarason, I. G., & Sarason, B. R. (1990). Text anxiety. In H. Leiternberg (Eds.), *Handbook of social and evaluation anxiety* (pp.475-495). New York: Plenum Press.

Sawaki, Y. & Xi, X. (2019). Univariate generalizability theory in language assessment. In Aryadous, V. & Raquel, M. (Eds.), *Quantitative data analysis for language* *assessment Volume I Fundamental Techniques* (pp. 30-53). New York: Routledge.

Schensul, S., Schensul, J., and Lecompte, M. D. (1999). *Essential ethnographic methods: Observations, interviews, and questionnaires.* CA: Sage Publications.

Schneider, M. C., Egan, K. L., & Julian, M. W. (2013). Classroom Assessment in the Context of High-Stakes Testing. In J. H. McMillan (Eds.), *SAGE Handbook of Research on Classroom Assessment* (pp.55-70). CA: Thousand Oaks.

School Curriculum and Assessment Authority (SCAA). (1995). *Consistency in teacher assessment: guidance for schools, English Speaking and Listening Key Stages 1 to 3.* London: SCAA.

Schoonen, R., Van Gelderen, A., DeGlopper, K., Hulstijin, J., Snellings, P., Simis, A., & Stevenson, M. (2002). Linguistic knowledge, metacognitive knowledge, and retrieval speed in L1, L2 and EFL writing. A structural equation modelling approach. In S. Ransdell & M.-L. Barbier (Eds.), *New directions for research in L2 writing* (pp. 101-122). Dordrecht: Kluwer Academic.

Schunk, D. H. (2007). *Learning Theories: An Educational Perspective.* New York: Prentice Hall.

Scovel, T. (1978). The effect of affect on foreign language learning: A review of the anxiety research. *Language Learning, 28*, 129-142. doi:10.1111/j.1467-1770.1978.tb00309.x

Scriven, M. (1967). The methodology of evaluation. In R. W. Tyler, R.M. Gagne, & M. Scriven (Eds.), *Perspectives of curriculum evaluation* (pp. 39-83). Chicago: Rand McNally.

Searby, M., & Ewers, T. (1997). An evaluation of the use of peer assessment in higher education: a case study in the school of music, Kingston University. *Assessment and Evaluation in Higher Education, 22*, 371-383. doi: 10.1080/0260293970220402

Shaver, J.P. (1990). Reliability and validity of measures of attitudes toward writing and toward writing with the computer. *Written Communication, 7*(3)*,* 375-392. doi: 10.1177%2F0741088390007003004

Shell, D.F., Murphy, C. C., & Bruning, R.H. (1989). Self-efficacy and outcome expectancy mechanisms in reading and writing achievement. *Journal of Educational Psychology, 81,* 91-100. Retrieved from doi:10.1037/0022-0663.81.1.91

Shepard, L. A., Hammerness, K., Darling-Hammond, L., Rust, F., Snowden, J. B., Gordon, E., et al. (2005). Assessment In L. Darling-Hammond & J. Bransford (Eds.), *Preparing teachers for a changing world: What teachers should learn and be able* *to do* (pp.275-326). San Francisco: Jossey-Bass.

Shinn, H. & Good III. (1993). 6. CBA: An Assessment of Its Current Status and Prognosis for Its Future. *Curriculum-based Measurement*, 139-178. Retrieved from http://digitalcommons.unl.edu/buroscurriculum/8/

Slavin, R. E., Hurley, E. A., & Chamberlain, A. M. (2003) Cooperative learning and achievement. In W. M. Reynolds & G. J. Miller (Eds.), *Handbook of psychology, vol 7: educational psychology* (pp. 177-198). Hoboken, NJ: Wiley. doi: 10.1002/0471264385.wei0709

Sluijsmans, D., Brand-Gruwel, S. & Van Merriënbor, J. (2002) Peer assessment training in teacher education: effects on performance and perceptions. *Assessment and Evaluation in Higher Education, 27*(5), 443-454. doi: 10.1080/0260293022000009311

Sluijsmans, D., Moerkerke, G., Van Merriënbor, J. & Dochy, F. (2001) Peer assessment in problem-based learning, *Studies in Educational Evaluation, 27*, 153-173. Retrieved from http://sluijsmans.net/wp-content/uploads/2019/01/Sluijsmans-Peer-Assesment-in-Problem-Based-Learning-2001_9.pdf

Sluijsmans, D. & Prins, F. (2006). A conceptual framework for integrating peer assessment in teacher education. *Studies in Educational Evaluation, 32*(1), 6-22. doi:10.1016/j.stueduc.2006.01.005

Smith, A. B., Rush, R., Fallowfield, L. J., Velikova, G, & Sharpe, M. (2008). Rasch fit statistics and sample size consideration for polytomous data. *BMC Medical Research Methodology, 8* (33). doi: 10.1186/1471-2288-8-33

Smith, R.E., & Smoll, F.L. (1990). Sport performance anxiety. In H. Leitenberg (Eds.), *Handbook of social and evaluation anxiety* (pp.417-454). New York: Plenum Press.

Spielberger, C.D. (1983). *Manual for the state-trait anxiety inventory (form Y).* Palo Alto California: Consulting Psychologists Press.

Spratt, M., Humphreys, G., & Chan, H. (2002). Autonomy and motivation: Which comes first? *Language Teaching Research, 6*(3), 245-266. doi: 10.1191%2F1362168802lr106oa

Stefani, L. (1992). Comparison of collaborative self, peer and tutor assessment in a biochemistry practical. *Biochemistry and Molecular Biology Education, 20*(3), 148-151. doi: 10.1016/0307-4412(92)90057-S

STEP. (2006). *The Eiken Can-do list*. Tokyo: Society for Testing English Proficiency. Retrieved from http://www.eiken.or.jp/eiken/exam/cando/list.html

Stiggins, R.J. (2002). Assessment Crisis: The Absence of Assessment For Learning. A Special Section on Assessment. *Phi Delta Kappan, 83*(10), 758-765. doi:

10.1177%2F003172170208301010

Stiggins, R. J. & Conklin, N. F. (1992). *In Teachers' Hands: investigating the practice of classroom assessment*. Albany, NY: SUNY Press.

Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Techniques and procedures for developing grounded theory* (2nd edition.). Thousand Oaks, CA: Sage Publications.

Sudweeks, R.R., Reeve, S., & Bradshaw, W.S. (2004). A comparison of generalizability theory and many-facet Rasch measurement in an analysis of college sophomore writing. *Assessing Writing, 9*, 239-261. doi: 10.1016/j.asw.2004.11.001

Sung, Y., Chang, K., Chiou, S., & Hou, H. (2005). The design and application of a web-based self- and peer- assessment system. *Computers & Education, 45*, 187-202. doi: 10.1016/j.compedu.2004.07.002

Swain, M., & Watanabe, Y. (2013). Languaging: Collaborative dialogue as a source of second language learning. In C. Chapelle (Eds.), *The encyclopedia of applied linguistics* (pp. 1-8). Oxford: Wiley-Blackwell.

Tabachnick, B. G. & Fidell, L. S. (2014). *Pearson New International Edition Using Multivariate Statistics*. UK: Pearson.

Takahashi, T. (2004). Gakush-sha-tachi ha nani-wo kitai shiteiruka? *Eigo-kyoiku*, *53*, 28-29.「学習者たちは何を期待しているか？」高橋妙子著。大修館書店。

Takahashi, S., Nakano, N., & Suzuki, K. (2014). What Multivariate Analysis should you select for a research in physical education and sports science? *Taiiku-sokutei-hyoka-kenkyu, 13*(0), 41-52. doi: 10.14859/jjtehpe.13.41

Teasdale, A. & Leung, C. (2000). Teacher assessment and psychometric theory: a case of paradigm crossing? *Language Testing, 17*(2), 163-184. doi: 10.1177%2F026553220001700204

Tindle, R. & Longstaff, M. G. (2015). Writing, Reading, and Listening Overload Working Memory Performance Across the Serial Position Curve. *Advances in Cognitive Psychology, 11*(4), 147-151. doi: 10.5709%2Facp-0179-6

Topping, K. J. (1998). Peer assessment between students in college and university. *Review of Educational Research, 68*(3), 249-276. doi: 10.3102%2F00346543068003249

Topping, K. J. (2000). Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169. doi: 10.1080/713611428

Topping, K. J. (2003). Self- and peer- assessment in school and university: Reliability, validity and utility. In: M. Segers & E. Cascallar (Eds.), *Optimizing new methods* *of assessment: In search of qualities and standards* (pp.55-87). Dordrecht, Netherlands: Kluwer Academic Publishers.

Topping, K. J. (2009). Classroom Assessment. *Theory Into Practice*, *48*(1), 20-27. doi: 10.1080/00405840802577569

Topping, K. J. (2013). Peers as a source of formative and summative assessment. In J. H. McMillan (Ed.), *Sage Handbook of Research on Classroom Assessment.* (pp. 395-412). Los Angeles: Sage Publications

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative Peer Assessment of Academic Writing Between Postgraduate Students. *Assessment & Evaluation in Higher Education, 25*(2), 149-169. doi: 10.1080/713611428

Torrance, H. & Pryor, J. (1998). *Investigating formative assessment: teaching, learning and assessment in the classroom.* Buckingham: Open University Press.

Truscott, J., & Hsu, A. Y. (2008). Error correction, revision, and learning. *Journal of Second Language Writing, 17*, 292-305. doi: 10.1016/j.jslw.2008.05.003

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. 255-274). Berlin, Germany: De Gruyter.

Tsai, H. M. (2008). The development of an English writing anxiety scale for institute of technology English majors. *Journal of Education and Psychology, 31*(3), 81-107. Retrieved from http://www.fed.cuhk.edu.hk/ceric/jep/200800310003/0081.htm

Tsao, J., Tseng, W., & Wang, C. (2017). The Effects of Writing Anxiety and Motivation on EFL College Students' Self-Evaluative Judgments of Corrective Feedback. *Psychological Reports, 20*(2), 219-241. doi: 10.1177%2F0033294116687123

Tsui, A. & Ng, M. (2000). Do Secondary L2 Writers Benefit from Peer Comments? *Journal of Second Language Writing*, *9*(2), 147-170. doi: 10.1016/S1060-3743(00)00022-9

VanDerHeide, J. & Newell, G. (2013). Instructional Chains as a Method for Examining the Teaching and Learning of Argumentative Writing in Classrooms. *Written Communication, 30*, 300-329. doi: 10.1177%2F0741088313491713

van Gennip, N. A., Segers, M. S. S., & Tillema, H. H. (2010). Peer assessment as a collaborative learning activity: The role of interpersonal variables and conceptions. *Learning and Instruction, 20*(4), 280-290. doi: 10.1016/j.learninstruc.2009.08.010

Vogely, A.J. (1998). Listening comprehension anxiety: Students' reported sources and solutions. *Foreign Language Annals, 31*, 67-80. doi: 10.1111/j.1944-9720.1998.tb01333.x

Vu, T., & Dall'Alba, G. (2007). Students' experience of peer assessment in a professional course. *Assessment & Evaluation in Higher Education*, *32*(5), 541-556. doi: 10.1080/02602930601116896

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing*, *11*(3), 322-344. doi: 10.1177%2F026553229401100305

Warwick, P., & Maloch, B. (2003). Scaffolding speech and writing in the primary classroom: A consideration of work with literature and science pupil groups in the USA and UK. *Reading Literacy and Language, 37*(2), 54-63. doi: 10.1111/1467-9345.3702003

Weaver, B.D. & Esposto, A. (2012). Peer assessment as a method of improving student engagement. *Assessment and Evaluation in Higher Education, 37*(7), 805-816. doi: 10.1080/02602938.2011.576309

Weaver, B.D., Van Keer, H., Schellens, T. Valcke, M. (2011). Assessing Collaboration in a wiki: The reliability of university students' peer assessment. *The Internet and Higher Education, 14*(4). 201-206. Retrieved from http://hdl.handle.net/1854/LU-2079317

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing*, *15* (2), 263-287. doi: 10.1177%2F026553229801500205

Weigle, S. C. (2002). *Assessing writing*. Cambridge: Cambridge University Press.

Wenden, A. (1987). Conceptual background and utility. In Wenden, A. and Rubin, J. (Eds.), *Learner Strategies in Language Learning* (pp. 103-118). Englewood Cliffs, NJ: Prentice-Hall International.

White, E.M. (1985). *Teaching and assessing writing*. San Francisco, CA: Jossey-Bass.

Widdowson, H. G. (1998). Context, community, and authentic language. *TESOL Quarterly, 32*(4), 705-716. doi: 10.2307/3588001

Wiliam, D. (1998). *Enculturaling learners into communities of practice: Raising achievement through classroom assessment.* Paper presented at the European Conference for Educational Research, University of Ljubjana, Slovenia. Retrieved from http://www.leeds.ac.uk/educol/documents/000000874.htm

Wiliam, D. (2000). *Integrating summative and formative functions of assessment*. Paper presented to Working group 10 of the International Congress on Mathematics Education, Makuhari, Tokyo. Retrieved from http://www.kcl.ac.uk//depsta/education/hpages/dwliam.html

Wiliam, D. (2007). *Five "key strategies" for effective formative assessment*. Paper presented to National Council of Teachers of Mathematics assessment research brief. Reston, VA: NCTM. Retrieved from http://www.nctm.org/uploadedFiles/Research_and_Advocacy/research_brief_and _clips/Research_brief_04_-_Five_Key%20Strategies.pdf

Wiliam, D. (2018). *Embedded formative assessment.* IN. U.S.: Solution Tree Press.

Wiseman, C. (2012). A Comparison of the Performance of Analytic vs. Holistic Scoring Rubrics to Assess L2 Writing. *Iranian Journal of Language Testing, 2*(1), 59-92. Retrieved from http://cdn.ov2.com/content/ijlte_1_ov2_com/wp-content_138/uploads/2019/07/408-2012-2-1.pdf

Wolfe-Quintero, K., Inagaki, S., & Kim, H-Y. (1998). *Second language development in writing: Measure of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii.

Woodrow, L. (2006). Anxiety and speaking English as a second language. *RELC Journal, 27* (3), 308-328. doi: 10.1177%2F0033688206071315

Woodrow, L. (2011). College English writing affect: Self-efficacy and anxiety. *System. 39*(4), 510-522. doi: 10.1016/j.system.2011.10.017

Wray, D., & Lewis, M. (1997). Teaching factual writing. *The Australian Journal of Language and Literacy, 20*(2), 43-52. Retrieved from http://search.informit.com.au/documentSummary;dn=264292620341185;res=IELHSS

Wright, B. (1984). Despair and hope for educational measurement. *Contemporary Education Review, 3*(1), 3-24. Retrieved from http://www.rasch.org/memo41.htm

Wright, B. (1998). Rating Scale Model (RSM) or Partial Credit Models (PCM)? *Rasch Measurement Transactions*, *12*(3), 641-642. Retrieved from http://www.rasch.org/rmt/rmt1231.htm

Wright, B. (2006). Comparing groups in a before-after design: When *t* test and ANCOVA produce different results. *British Journal of Educational Psychology, 76*, 663-675. doi: 10.1348/000709905X52210

Wright, B.D. & Masters, G.N. (1981). *The measurement of knowledge and attitude (Research Memorandum No. 30)*. Chicago, IL: University of Chicago, MESA Psychometric Laboratory.

Wu, Y. (1992). *First and second language writing relationship: Chinese and English*. (Unpublished doctoral dissertation), Texas A & M University, College Station. Retrieved from http://hdl.handle.net/1969.1/DISSERTATIONS-1348972

Xiao, Y. & Wong, K. F. (2014). Exploring heritage language anxiety: a study of Chinese Heritage Language Learners. *The Modern Language Journal, 98*(2), 589-611. doi: 10.1111/modl.12085

Yamashita, S. O. (1996). *Six measures of JSL pragmatics.* Honolulu, HI: University of Hawai'i Press.

Yin, M. (2010). Understanding Classroom Language Assessment Through Teacher Thinking Research. *Language Assessment Quarterly, 7*(2), 175-194. doi: 10.1080/15434300903447736

Young, D. J. (1986).The relationship between anxiety and foreign language oral proficiency ratings. *Foreign Language Annals, 19*(5), 439-445. doi: 10.1111/j.1944-9720.1986.tb01032.x

Zarei, A. & Mahdavi, A. (2014). The Effect of Peer and Teacher Assessment on EFL Learners' Grammatical and Lexical Writing Accuracy. *Journal of Social Issues & Hummanities*, *2*(9), 92-97. Retrieved from http://www.researchgate.net/publication/312471482_The_Effect_of_Peer_and_Teacher_Assessment_on_EFL_Learners'_Grammatical_and_Lexical_Writing_Accuracy

Zarei, A. & Usefli, Z. (2015). The Effect of Assessment Type on EFL Learners' Goal-Orientation. *Journal of Language, Linguistics and Literature, 1*(4), 112-119. Retrieved from http://creativecommons.org/licenses/by-nc/4.0/

Zheng, Y. (2008). Anxiety and Second/Foreign Language Learning Revisited. *Canadian Journal for New Scholars in Education, 1*, (1), 1-12. Retrieved from http://cdm.ucalgary.ca/index.php/cjnse/article/view/30393

Zuckerman, M. (1960). The development of an affect adjective check list for the measurement of anxiety. *Journal of Consulting Psychology, 24*, 457-462. doi: 10.1037/h0042713

Appendices

Appendix A: Questionnaire about Writing Anxiety and Learner Autonomy

これは､ライティングの授業を改善するためのアンケートです｡最後までどうぞご協力をお願いします｡
もし､わかりにくい言葉があったら､質問をしてください｡

1. 英語で文章を書いている間、自分は全然神経質ではない。　（神経質：イライラする）

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

2. 時間制限があるなかで英作文を書くと、心臓がどきどきする。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

3. 英作文を書いている間、評価されることがわかると心配で不安に感じる。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

4. 英語で考えを書きとめることをよくする。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

5. ふだんから英作文を書くことをできるだけ避けるようにしている。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

6. 英作文を書こうとすると頭がよく真っ白になる。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

7. 自分の英作文が他の人の英作文よりも非常に悪いということは心配していない。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

8. 時間のプレッシャーがあって英作文を書くと震えて汗が出てくる。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

9. もし自分の英作文を評価されることになれば、悪い評価をとるのではないかと心配だ。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

10. 英語で書かなければならない状況を避けるのにできるだけのことをする。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

11. 時間制限があるなかで英作文を書くと、考えがごちゃごちゃになる。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

12. もし、選択肢がなければ、作文を書くのに英語は使わないだろう。

1　2　3　4
全然そうではない ⟷ 本当にその通りだ

13.　時間制限があるなかで英作文を書くと、よくパニックになる。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

14.　自分が書いた英作文を他の生徒が読んだらばかにされるのではないかと心配だ。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

15.　予期しない中で英作文を書くように言われると、体が固まる。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

16.　英作文を書くように言われると、（やりたくなくて）言い訳をするのに最善を尽くしてしまう。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

17.　自分が書いた英作文について他人が考えることについて心配しない。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

18.　教室外でふだんから英作文が書けるあらゆるチャンスを探している。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

19.　英作文を書くとふだんから体全体がこわばり緊張する。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

20.　自分が書いた英作文を授業の話し合いの例として選ばれるのではないかということが心配だ。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

21.　自分の英作文が大変悪いと評価されても全然心配しない。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

22.　可能ならいつでも、英語を使って作文を書くつもりだ。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

23.　自分の長所短所を把握している。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

24.　自分自身の学習目標を設定している。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

25.　授業以外で何を学ぶべきかを自分で決めている。

1　2　3　4

全然そうではない ↔ 本当にその通りだ

26.　自分の学びや進歩を自己評価している。

1　2　3　4

27. 英語学習に対する興味を自分で刺激し意欲を高めている。

1　　2　　3　　4

全然そうではない ↔ 本当にその通りだ

28.　先生からだけではなく、友達からも学んでいる。

1　　2　　3　　4

全然そうではない ↔ 本当にその通りだ

29.　自分の勉強に対してより自己管理をするようになってきた。

1　　2　　3　　4

全然そうではない ↔ 本当にその通りだ

30.　学習教材に関して意見を言える。

1　　2　　3　　4

全然そうではない ↔ 本当にその通りだ

31.　先生から教えてもらう知識をまつよりむしろ自分で英語関する知識を得ている。

1　　2　　3　　4

全然そうではない ↔ 本当にその通りだ

32.　教室で何を学ぶべきかに関して意見を言える。

1　　2　　3　　4

全然そうではない ↔ 本当にその通りだ

自己評価（他己評価）の効果的な点と問題点は何だと思いますか。

効果的な点　（　　　　　　　　　　　　　　　　　　　　　　　　）

問題点　（　　　　　　　　　　　　　　　　　　　　　　　　　　）

ご協力を有難うございました。

年　　　組　　番　　　氏名(　　　　　　　　　　　　　)

Appendix B: Writing Assessment Rubric

Components	Assessment criteria	Scores
Task Fulfillment （課題で求められている内容が含まれているかどうか）	Does English composition fulfill the task? - present one's idea or opinion - present two reasons -present concrete explanation and examples 課題で求められている内容が含まれていますか？ 自分の考えや意見を明示していますか。 その理由が二つ書かれていますか。 理由には具体的な例や説明が加えられていますか。	Good Poor 4 3 2 1
Structure and Coherence （構成と内容の一貫性）	Is composition written coherently and logically? -easy to understand -use discourse markers effectively -well organized 英作文は内容に一貫性があり、論理的ですか。 内容や表現に一貫性があり、読んでいてわかりやすいですか。 伝えたい情報の流れや展開を示す表現（接続詞など）を効果的に使っていますか。 関係のない考えや情報が書かれていず、まとまりがありますか。	Good Poor 4 3 2 1
Appropriate Usage of Vocabulary （語彙の適切な使用）	Is vocabulary used appropriately and accurately? -accurate spelling -appropriate usage of vocabulary 語彙は適切で正確に使われていますか。 単語の綴りは正確ですか。 単語は正しい意味で使われていますか。 英語以外の言葉を使う場合は、その言葉を知らない人でも理解できるように説明が加えられていますか。	Good Poor 4 3 2 1
Grammatical Accuracy （正確な文法の使用）	Is grammar used to effectively convey ideas and reasons? 自分の考えとその理由を効果的に伝えられるように文法は効果的に使われていますか。	Good Poor 4 3 2 1
Sum（合計）		/16

＊次に全体をふりかえって自分の英作文の評価を日本語か英語でコメントしてください。

My English writing is ______________________________________.

私（ぼく）の英作文は全体的に______________________________________。

NAME (   )

Appendix C: Writing Work sheet

My Favorite Food

Write an English composition about your favorite food, titled “My Favorite Food”. You should present reasons and examples in your writing in approximately 40 - 50 words in 10 minutes without using dictionaries.

□ words
語数を書いてください。

年　　　組　　　番　氏名

Appendix D: Narrative Frame for Peer Assessment Group

友達の英作文を読み、思ったことを下の枠組みにそって書き入れてください。その際、英語か日本語のどちらで書いてもかまいません。この紙をパートナーに見せる必要はありません。なお、こちらの紙は、授業終了後に提出となります。

Hello. I enjoyed reading your writing. From now, let me talk about your writing.
（こんにちは。英作文を読ませてもらって楽しかったです。〇〇君・さんの書いた英作文ついて私（ぼく）が思ったことについてこれから話しをさせてください）

My overall impression is ______________________________________.
私（ぼく）の全体的な印象は__________________________________です。

Next, let me talk about four components.
次に、４つの項目について話しをさせてください。

At first, about task fulfillment, I think __________________________________.
最初に、課題の達成についてですが、__________________________だと思いました。

Secondly, the structure and coherence of your writing are _________________________.
二つ目として、構成や内容の一貫性についてですが、__________________________だと思います。

Thirdly, you used vocabulary _____________________________________.
三つ目に、単語を_________________________________に使っていたと思います。

Finally, your grammatical correctness is ________________________________.
最後に文法は____でした。

Thank you for reading. Do you have any comments or questions?
読んでくれて有難う。何か質問やコメントはありますか？

Your peer's name　　　　　　　　Your name
友達の名前　(　　　　　)　　　　　あなたの名前 (　　　　　)

Appendix E: Summary Rasch Statistics of Composite Score

Statistic	Students	Raters	Tasks
M (measure)	.92	.00	.00
SD (measure)	.61	.45	.02
M (*SE*)	.26	.04	.02
RMSE	.27	.04	.02
Adj. (true) *SD*	.5	.45	.02
df	292	5	1
Separation ratio (*G*)	2.26	11.58	.98
Separation (strata) index (*H*)	3.36	15.78	1.64
Separation reliability (*R*)	.84	.99	.49

Note. *RMSE* =root mean-square measurement error. “Examinees with non-extreme scores only. “The rater and tasks facets were each constrained to have a mean element measure of zero. ** $p < .01$.

Appendix F: The Bias Analysis of the Composite Scores of Four Analytic Rating Scales

Rater	Observed score	Expected score	Measure ment-N	Obs-Exp Task Average	Bias+	Size measure ment	Model *SE*	*t*	*DF*	Probability	Infit Sq	*M*	*M* Sq
Peers of Pre-task	2180	2033.49	-1.01	1.00	.88	.03	.08	10.41	145	.00		3.5	2.7
Self of Post-task	1898	1819.90	.21	.53	.38	-.03	.07	5.30	145	.00		.6	.4
Teacher 4 of Post-task	3610	3601.67	.20	.06	.04	-.03	.05	.79	290	.4250		.6	.6
Teacher 1 of Post-task	3630	3612.73	.20	.06	.04	-.03	.05	.79	291	.4277		.6	.6
Teacher 2 of Post-task	3630	3612.73	.20	.06	.04	-.03	.05	.79	291	.4277		.6	.6
Teacher 3 of Post-task	3641	3623.80	.20	-.06	-.03	-.03	.05	-.76	292	.4305		.6	.6
Teacher 3 of Pre-task	3579	3595.87	.20	-.06	-.03	-.03	.04	-.77	292	.4468		.6	.6
Teacher 1 of Pre-task	3568	3584.93	.20	-.06	-.03	.03	.04	-.77	291	.4439		.6	.6
Teacher 2 of Pre-task	3568	3584.93	.20	-.06	-.03	.03	.04	-.77	291	.4439		.6	.6
Teacher 4 of Pre-task	3557	3574.00	.20	-.06	-.03	.31	.04	-.77	290	.4411		.6	.6
Self of Pre-task	1716	1793.95	.21	-.54	-.031	.31	.06	-.520	144	.00		4.0	4.6
Peers of Post-task	1898	2044.30	-1.01	-1.00	-.74	-.03	.07	-10.75	145	.00		.7	.5
Mean	3040.3	3040.19	243.3	.00	.02		.05	-.01				1.1	1.1
SD (Population)	796.3	793.00	66.9	.47	.36		.01	4.86				1.2	1.2
SD (Sample)	831.7	828.26	72.0	.49	.38		.01	5.08				1.2	1.3

Fixed (all = 0) chi-square: 283.7 *DF:* 12 significance (probability): .00

Appendix G: Summary Rasch Statistics of Analytic Rating Scales of Pre- & Post-tests

Statistic	Examinees (Students)	Raters	Occasion * Rating Scale
M (measure)	1.56	.00	.00
SD (measure)	.65	.17	.77
M (*SE*)	.22	.03	.04
RMSE	.22	.03	.04
Adj. (true) *SD*	.65	.37	.77
Df	291	5	7
Separation ratio (*G*)	2.92	11.12	20.43
Separation (strata) index (*H*)	4.23	15.16	27.58
Separation reliability (*R*)	.90	.99	1.00

Note. *RMSE* =root mean-square measurement error. “Examinees with non-extreme scores only. “The rater and tasks facets were each constrained to have a mean element measure of zero. ** $p < .01$.

Appendix H: The Bias Analysis of Four Analytic Rating Scales

Rater	Observed score	Expected score	Measurement-N	Obs-Exp Task Average	Bias+	Size measurement	Model *SE*	*t*	*DF*.	Probability	Infit Sq	*M*	*M* Sq
Self Pre-Grammatical correctness	968	772.76	-.15	.67	1.38	1.26	.09	15.35	291	.000		1.2	1.1
Peers Post-Grammatical correctness	523	480.19	-.77	.29	.79	.69	.15	5.28	145	.000		1.2	1.5
Peers Pre-Grammatical correctness	968	864.62	-.77	.35	.75	1.26	.09	8.40	291	.000		1.6	1.1
Peers Post-Appropriate usage of vocabulary	530	499.30	-.77	.21	.63	.37	.16	4.04	145	.0001		1.2	1.4
Self Pre-Appropriate usage of vocabulary	964	919.74	-.15	.15	.33	.25	.09	3.76	291	.0002		1.5	1.2
Teacher 2 Post-Task Fulfilment	1069	1042.79	.25	.09	.31	-1.20	.11	2.72	291	.0068		1.2	1.1
Teacher 3 Post-Task Fulfilment	1069	1044.10	.23	.09	.30	-1.20	.11	2.60	291	.0098		1.3	1.1
Teacher 4 Post-Task Fulfilment	1070	1046.18	.21	.08	.29	-1.20	.11	2.50	291	.0129		1.2	1.1
Teacher 1 Post-Task Fulfilment	1069	1045.53	.22	.08	.28	-1.20	.11	2.46	291	.0145		1.2	1.1
Teacher 1 Pre-Structure & Coherence	1996	1934.49	.22	.11	.26	-.48	.07	3.88	583	.0001		1.3	.9
Teacher 1 Pre-Task Fulfilment	2080	2030.25	.22	.09	.25	-.89	.07	3.41	583	.0007		1.0	1.3
Teacher 4 Pre-Structure & Coherence	995	968.14	.21	.09	.22	-.48	.09	2.40	291	.0169		1.6	.9
Teacher 3 Pre-Structure & Coherence	992	965.27	.23	.09	.22	-.48	.09	2.38	291	.0179		1.0	.9
Teacher 2 Pre-Structure & Coherence	990	963.46	.25	.09	.22	-.48	.09	2.36	291	.0190		1.0	1.0
Teacher 3 Pre-Task Fulfilment	1034	1013.44	.23	.07	.20	-.89	.10	1.99	291	.0477		1.1	1.3
Teacher 2 Pre-Task Fulfilment	1030	1011.91	.25	.06	.18	-.89	.10	1.74	291	.0822		1.6	1.4

Teacher 4 Pre-Task Fulfilment	1033	1015.88	.21	.06	.17	-.89	.10	1.67	291	.0070	1.6	1.3
Self Post-Grammatical correctness	431	418.68	-.15	.08	.06	.69	.12	1.43	145	.1560	1.2	1.3
Teacher 2 Post-Structure & Coherence	907	898.48	.25	.03	.06	.00	.08	.72	291	.4750	1.2	1.2
Teacher 3 Post-Structure & Coherence	909	900.51	.23	.03	.05	.00	.08	.71	291	.4761	1.2	1.2
Teacher 4 Post-Structure & Coherence	911	903.76	.21	.02	.03	.00	.08	.61	291	.5424	1.2	1.2
Teacher 1 Post-Structure & Coherence	907	902.74	.22	.01	.00	.00	.06	.36	583	.7205	1.2	1.2
Teacher 1 Pre-Appropriate usage of vocabulary	1736	1735.38	.22	.00	.00	.25	.08	.04	291	.9709	.5	.5
Teacher 4 Post-Appropriate usage of vocabulary	850	850.59	.21	.00	-.01	.37	.08	-.05	291	.9615	.6	.6
Teacher 3 Post-Appropriate usage of vocabulary	846	847.20	.23	.00	-.01	.37	.08	-.10	291	.9215	.6	.6
Teacher 4 Pre-Appropriate usage of vocabulary	867	868.74	.21	-.01	-.01	.25	.08	-.14	291	.8859	.5	.5
Teacher 2 Post-Appropriate usage of vocabulary	843	845.08	.25	-.01	-.02	.37	.08	-.17	291	.8640	.6	.6
Teacher 2 Pre-Appropriate usage of vocabulary	860	863.28	.25	-.01	-.04	.25	.08	-.27	291	.7864	.5	.5
Teacher 3 Pre-Appropriate usage of vocabulary	860	865.38	.23	-.02	-.04	.25	.08	-.44	291	.6569	.5	.5
Teacher 1 Post-Appropriate usage of vocabulary	843	849.53	.22	-.02	-.05	.37	.08	-.54	291	.5912	.6	.6

Peers Pre-Task Fulfilment	550	551.42	-.77	-.01	-.08	-.89	.19	-.27	145	.7861	1.2	.9
Teacher 2 Post-Grammatical accuracy	786	797.75	.25	-.04	-.09	.69	.08	-.97	291	.3346	.6	.7
Teacher 4 Post-Grammatical accuracy	790	803.27	.21	-.05	-.09	.69	.08	-1.09	291	.2764	.6	.7
Teacher 3 Post-Grammatical accuracy	786	799.87	.23	-.05	-.11	.69	.08	-1.14	291	.2511	.7	.7
Teacher 1 Post-Grammatical accuracy	786	802.21	.22	-.06	-.23	.69	.12	-1.33	291	.1839	.6	.7
Self Post-Structure & Coherence	452	468.14	-.15	-.11	-.25	.00	.14	-1.96	145	.0521	1.1	1.1
Peers Post-Structure & Coherence	506	518.34	-.77	-.08	-.27	.00	.12	-1.79	145	.0757	1.4	1.4
Self Post-Appropriate usage of vocabulary	422	442.27	-.15	-.14	-.29	.37	.09	-2.37	145	.0190	1.1	1.1
Peers Pre-Appropriate usage of vocabulary	964	998.44	-.77	-.12	-.36	.25	.09	-3.23	291	.0014	1.2	1.2
Self Pre-Structure & Coherence	968	1009.90	-.15	-.14	-.43	-.48	.06	-4.01	291	.0001	1.1	1.1
Teacher 1 Pre-Grammatical accuracy	1324	1440.80	.22	-.20	-.44	1.26	.09	-6.97	583	.0000	.7	.7
Teacher 4 Pre-Grammatical accuracy	662	721.40	.21	-.20	-.45	1.26	.09	-5.01	291	.0000	.7	.7
Teacher 3 Pre-Grammatical accuracy	658	718.19	.23	-.21	-.47	1.26	.09	-5.09	291	.0000	.7	.7
Teacher 2 Pre-Grammatical accuracy	654	716.19	.25	-.21	-.66	1.26	.09	-5.26	291	.0000	.7	.7
Self Pre-Task Fulfilment	1463	1567.06	-.15	-.24	-.47	-.89	.07	-8.85	436	.0000	1.0	1.0
Peers Post-Task Fulfilment	530	558.97	-.77	-.20	-.94	-1.20	.16	-6.03	145	.0000	1.6	1.5
Peers Pre-Structure & Coherence	968	1067.69	-.77	-.34	-.98	-.48	.09	-10.96	291	.0000	1.1	1.1

Self Fulfilment	Post-Task	464	533.41	-.15	-.48	-1.27	-1.20	.12	10.50	145	.0000	1.0	.0
Mean		914.2	914.22	292.0	.00	-.01		.10	-.16			1.0	1.0
SD (Population)		346.9	341.91	107.4	.18	.44		.03	4.48			.3	.3
SD (Sample)		350.5	345.52	108.6	.18	.45		.03	4.53			.4	.3
Fixed (all = 0) chi-square:		965.6	d.f.:	48		significance	(probability): .00						

Appendix I: Task Measurement Report about Four Analytic Rating Scales of Self- & Peer Assessment Group

Analytical Rating Scales	Observed Average	Measure logit	Model *SE*	Infit *M* Sq	*Z* Std	Outfit *M* Sq	*Z* Std
Self-Pre-Grammatical	2.2	2.08	.07	.60	-7.2	.61	-6.9
Self-Post-Grammatical	2.3	1.76	.07	.86	-2.4	.88	-2.0
Peer-Post-Vocabulary	2.59	.90	.06	.76	-4.9	.80	-4.0
Peer-Pre-Vocabulary	2.80	.63	.06	.64	-8.5	.65	-8.2
Peer-Pre-Grammatical	2.80	.63	.06	.62	-9.1	.63	-8.7
Self-Post-Structure	2.92	.37	.06	.52	-9.9	.54	-9.9
Peer-Post-Grammar	2.98	.24	.06	1.40	7.5	1.42	7.8
Self-Pre-Structure	2.99	.23	.06	.68	-7.6	.69	-7.2
Peer-Post-Structure	3.00	.19	.06	.48	-9.9	.49	-9.9
Peer-Post-Grammar	3.25	-.36	.06	1.30	5.6	1.28	5.1
Self-Pre-Vocabulary	3.38	-.68	.07	1.35	6.1	1.27	4.7
Self-Post-Vocabulary	3.42	-.79	.07	1.08	1.4	1.05	.9
Self-Pre-Task	3.49	-.99	.07	1.99	9.9	1.71	9.9
Self-Post-Task	3.58	-1.27	.08	1.74	9.9	1.62	7.9
Peer-Post-Task	3.64	-1.46	.08	1.45	6.1	1.35	4.5
Peer-Pre-Task	3.69	-1.67	.08	1.41	5.2	1.21	2.5
Mean		.00	.07	1.06	-.50	1.01	-.80
SD (population)		1.07	.01	.46	7.4	.39	6.8

Note. Task: Task Fulfillment; Structure: Structure & Coherence; Vocabulary: Appropriate Usage of Vocabulary; Grammatical: Grammatical Accuracy

Person: Real Sep. .91; REL: .45 ITEM: Real Sep. 14.09: REL: .99