A TASK-BASED LANGUAGE TEACHING APPROACH TO GROUP DISCUSSIONS IN JAPANESE UNIVERSITY CLASSROOMS: AN EMPIRICAL STUDY OF GOAL-SETTING AND FEEDBACK

by

ROBERT STROUD

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language
and Applied Linguistics
College of Arts and Law
The University of Birmingham
March 2018

ABSTRACT

This thesis draws on a classroom-based empirical study to explore the actual effects that Task-Based Language Teaching (TBLT) has on students' performance, when applied to group discussions, and the impacts that different forms of Goal-Setting and Feedback (GSF) have on their learning. In doing so, it challenges the assumptions in the research literature that TBLT will necessarily improve multiple aspects of performance within group discussions with low-level students, and reveals that applying GSF can lead to very different outcomes.

A longitudinal mixed-method approach was adopted using surveys and peer-interviews with 10 teachers, and observations, surveys and peer-interviews with 132 low-level students in a Japanese university. Students used product or process GSF alongside TBLT group discussions across a semester. Findings showed improvements in fluency and accuracy, positive feelings towards learning, and larger improvements for lower performers. Furthermore, product and process goals influenced students' focus differently in terms of individual performance, collaboration and discussion outcome. These findings create a clearer picture of the impact of TBLT, when applied to group discussions, and show how students' focus within learning can be greatly influenced by task goals. Resultant recommendations for course design, student and teacher training, and implementation of TBLT and GSF are given.

DEDICATION

I dedicate this PhD to my wife, Yasuko, who gave me tremendous support to make it possible to complete. During its undertaking, we moved home, began new careers, had our first child together, and suffered family loss. The hard work, sacrifices and love which she showed me were the reasons I could finish it.

ACKNOWLEDGEMENTS

I would firstly like to acknowledge the continual support of Caroline Tagg, who went beyond the requirements of her role as my supervisor to give me countless rounds of helpful feedback and guidance to make the completion of this PhD possible.

I would also like to thank my second supervisor, Paul Thompson, for his guidance and continual support across the duration of the PhD when I faced additional challenges to completing it.

Finally, I would like to thank the teachers and students who participated in the study for their hard work and valuable opinions throughout the process.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

LIST OF ABBREVIATIONS

CAF - Complexity, Accuracy and Fluency

CLT - Communicative Language Teaching

FA - Formative Assessment

GSF - Goal-Setting and Feedback (task performance focused and self-regulated by students)

L2 - Second Language

LPs/HPs - Low Participators/High Participators (half of students who spoke the least/most in discussions at the start of the study)

ProcS - Process Students (used Process GSF sheet/diary during class across the semester)

ProdS - Product Students (used Product GSF sheet/diary during class across the semester)

SLA - Second Language Acquisition

SRL - Self-Regulated Learning

TBLT - Task-Based Language Teaching

CHAPTER 1. INTRODUCTION

1.1 Thesis focus and aims

This thesis is the third part of a Modular PhD investigating the use of Task Based Language Teaching (TBLT) oral group discussion tasks for language learning with low-level learners. The overall aim of the PhD is to investigate and report on ways to improve the learning for students. This was done by firstly determining key factors affecting low-level Japanese university students' oral participation within discussions in the first module (see Appendix A for a summary), and then by examining the short-term effects on participation of *pre-discussion planning* (a significantly reported factor in the first module) with low-level Japanese university students in the second module (see Appendix B for a summary). The main finding was that when the students undertook such additional planning, they would speak more and with more fluency during discussions immediately afterwards.

Three of the other task design factors reported to potentially improve participation in the first module were related to 1) *having a scoring system for performance*, 2) *getting feedback on performance*, and 3) *seeing measurable progress of performance over time*. As a result, I decided to focus this thesis on these three factors by investigating the effects on TBLT group discussion learning of self-regulated performance Goal-Setting and Feedback (GSF) via a semester-long classroom-based study. Data in this thesis considers observable changes in performance by Japanese university students due to the use of a TBLT approach to group discussions, changes observed with the use of two different types of GSF (task product versus process focused), and self-reported feelings of the students towards the learning undertaken. The findings contribute to TBLT and goal-related research by examining the suitability of TBLT group discussions as an

approach to improving language use with low-level learners and how GSF may support the learning or not.

1.2 Background and research motivation

Upon entering university, most Japanese students have studied English since an elementary school age, most recently with five years of mainly grammar-focused English instruction in Junior and Senior High School involving translating between Japanese and English, known as the *yakudoku* method (Gorsuch, 1998; Nishino, 2008; Nishino & Watanabe, 2008). Such classes have often not involved Communicative Language Teaching (CLT) approaches to second language learning, such as Task-Based Language Teaching (TBLT), resulting in limited chances for students to interact orally with each other in English. The Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) laid out plans in 2013 to enable students to hold conversations in English by the time they leave High School in preparation for the Olympic Games in Tokyo in 2020 (MEXT, 2013). If such ambitious goals are to be met, they require careful consideration with regards to the teaching of conversation skills during high school and into courses at the university level. However, because of the pressure placed on high school students to pass university entrance exams in Japan (Aspinall, 2005) classroom learning focuses mainly on the content of such tests via the yakudoku method. As a result, little time is left for orally interactive tasks, resulting in university students' oral English communicative competence being often limited to simple exchanges at best (King, 2012, 2013, p. 72).

I have been teaching English within Japan for ten years at the time of writing this thesis, having taught English communication skills at the elementary, high school, university and business-level. Of specific relevance to the focus of this thesis, I taught English communication courses at Kwansei Gakuin University in Kansai, Japan, between 2013 and 2016, and have been teaching similar courses at Hosei University in Tokyo since 2016. From my own experience of working within universities in Japan, students undertaking group discussion tasks have seldom experienced goal-setting for discussion performance, nor been provided with specific feedback to help focus their efforts on improving their performance related to such goals. However, a large amount of recent research, including some of my own, suggests that helping students focus on specific task performance goals and feedback can improve their motivation, efforts made, participation within classwork, and performance across time (Bargh, Gollwitzer, Lee-Chai, Barndollar, & Troetschel, 2001; Hart & Albarracín, 2009; Moskowitz & Grant, 2009; Stroud, 2017).

The number of choices available to teachers for implementing performance goals and feedback for oral tasks are vast (Lai, 2015; Leung, 1999; Leung & Lewkowicz, 2006; Norris, 2008) and are often subjective scale ratings of measures such as 'fluency', ‘accuracy’ and ‘complexity’ (such as in the TOEIC, TOEFL and IELTS speaking tests). A focus on such scoring can often leave students without an understanding of how to focus their efforts to improve in the future (Orsmond, Merry, & Reiling, 1997; Price, Handley, Millar, & O’Donovan, 2010). From what I have seen in Japan, feedback on classroom discussions also often comes in the form of such subjective, non-specific scale ratings from classmates or the teacher. I do not believe that this helps students understand their performance with enough detail, nor provide them with any

measurable progress on that performance over time to understand how to focus future efforts to improve. If Japanese students are expected to improve their performance across courses, they require specific and measurable goals to become motivated to take part in classwork (Moskowitz & Grant, 2009), as well as clear, specific, and ongoing feedback which provides them with what is called 'assessment for learning' (Dann, 2002) via a 'formative' style of feedback (Black & Wiliam, 2009; Harlen & James, 1997; Sadler, 1998; Tunstall & Gipps, 1996; Wiliam, 2018).

Several challenges exist for the implementation of goals and feedback into classroom group discussions. Firstly, it can be unclear for teachers and students how they should focus efforts within performance, such as goals related to individual speaking turns, interactions between speakers, or the outcome of the discussion itself. Secondly, there may be a lack of time for the use of goals or feedback within class. Such extra workload may take away from the time required for practising the use of the language. Also, English communication class sizes can sometimes be too large for the teacher to be able to spend time observing individual students across a course, in order to give them detailed individual feedback. Goals and feedback may need to be self-regulated by students themselves to avoid such issues with time. Thirdly, individual differences, such as learning preferences and English-speaking ability, can make the use of goals and feedback more difficult for some students than others. Lower-level students may already be struggling to perform within discussions alongside higher-level English speakers, and the additional workload of goals and feedback may actually have negative effects on their performance. Therefore, any goals or feedback used should be as quick and simple to use as possible. Lastly, any performance feedback provided to students needs to be clear and specific, but this may be difficult to do in a limited amount of time within classes.

Detailed research projects which investigate the development of student performance and feelings towards group discussions across time are scarce, even though this data would prove very helpful for teachers who are struggling to improve English oral interactions within their classes. Due to the extensive positive research which exists about the use of goals and feedback to improve classroom learning (see Chapter Four), as well as my own research and the findings in the first module (Appendix A), I decided to focus this study on how combined performance GSF might be self-regulated by students in typical English communication courses within Japanese universities to improve the learning with TBLT group discussions. I believe that such an approach is an important topic of future language learning research, as it can potentially help students understand their ability better (as determined by the goals and feedback used) and focus more on improving across time.

1.3 Research questions

The research questions within the study were selected to help improve the understanding of the potential effects of using a TBLT approach and GSF to support learning undertaken during classroom group discussions. With regards to the GSF used in the study, *goals* were those focused on discussion task performance which were set by students themselves within their electronic diaries (Appendices E and F) prior to each classroom discussion. *Feedback* for students referred to that which was provided by 1) audio recordings of group discussions, 2) notes which students took on their own discussion sheets and 3) the excel tables showing performance over time within the electronic diaries. The main overall RQ addressed was:

Main RQ: ***'What are the effects on learning of using Goal-Setting and Feedback (GSF) with TBLT group discussions across a semester?'***

I decided to approach this RQ by breaking it down into three separate RQs. The first RQ was used to specify what type of goals should be used with the students in the study. The second RQ then addressed observable changes in student performance over time with a TBLT approach, as well as with two different types of GSF used, including differences between students who spoke less (Low Participators) or more than others in discussions (High Participators) at the start of the study. The third RQ addressed self-reported student feelings towards undertaking TBLT discussions and the two types of GSF used (ProdS using Product GSF and ProcS using Process GSF). Within the study, Product GSF focused on goals related to the outcome at the *end* of discussions (the final group choice, reasons, examples and other possible choices and reasons), while Process GSF focused on goals related to the interactions which took place *during* the discussions (the number of opinions, reasons, examples, questions, answers, agreements and disagreements). The three RQs were:

RQ1: ***What are appropriate discussion performance goals for the Japanese university students in this study?***

RQ2: ***(a) How does observable discussion task performance change for the students across a semester using a TBLT approach (regardless of the type of GSF used)?***

(b) What different effects do Product and Process GSF have on observable performance across a semester?

(c) Are these effects the same for Low and High Participators?

RQ3: ***(a) How do ProdS and ProcS report feeling about performing in discussions across the semester?***

(b) How do they report feeling about the support the two types of GSF provided for their learning (or not)?

The findings for these RQs make important contributions to research by providing original data on the longitudinal effects which a TBLT approach to group discussions can have with low-level learners (in and out of Japan), as well as the impact which the addition of GSF has on learning. This is beneficial to both researchers and teachers currently using or wishing to apply such approaches to their own language courses.

1.4 Thesis outline

In line with the RQs above, the theoretical background discussed in this thesis addresses three main themes within Chapters Two, Three and Four. Firstly, a background to the current use of group discussions within language learning classrooms is examined, with particular attention given to the common use of a Task-Based Language Teaching approach. Secondly, a discussion of the literature connected to appropriately measuring group discussion performance for students is provided. Thirdly, the potential effects on performance and learning of the design of goal-setting and feedback for group discussions is discussed using current research and theories related to goal-setting, formative assessment and performance rubrics.

In Chapter Five, the methodology of the semester-long classroom study undertaken is explained. This includes a rationale for the mixed-methods approach taken for the data collection and analysis, details of the participants and procedures, specific details of the data collection for the three separate RQs, and the ethical considerations within the study.

In Chapter Six, the results, discussion and limitations for all three RQs are given. The first part discusses the observational data collected from a classroom pilot, as well as self-reported survey and interview data from teachers, which were used to create the two types of GSF in the study (RQ1). The second part summarizes the changes in student performance across the semester with TBLT groups discussions using the two different types of GSF via classroom observations (RQ2). The third part explores reasons for the changes seen in RQ2 by using data regarding student feelings towards their performance in discussions over time and the two types of GSF used with data from self-reported surveys, interviews and my own observations during classes and tests (RQ3).

In Chapter Seven, conclusions are reached about the use of TBLT group discussions as an approach to language learning with low-level learners and the effects of GSF. Based on these findings, the contributions made to research, recommendations for language teaching, overall limitations for the thesis, as well as recommended future research directions are explained.

CHAPTER 2. LANGUAGE LEARNING WITH TBLT DISCUSSIONS

2.1 Introduction

This chapter is made up of three main sections which gives an overview of TBLT group discussions as an approach to the learning and teaching of spoken English. The lack of classroom-based research to understand the actual effects of such an approach on student performance is highlighted and later analyzed using longitudinal data within the study. This chapter is focused mainly on how students and teachers may be benefiting or not from such a TBLT approach, and also how it may compare to a more traditional alternative approach called Present, Practise, Produce (PPP). The first section discusses relevant literature for understanding how students may acquire a second language through oral use of the language during tasks. The second section discusses the potential benefits and challenges to both learning and teaching with the use of group discussions within English communication courses. The third section gives an overview of the Task-Based Language Teaching (TBLT) approach to classroom group discussions, as well as the potential benefits and problems for both students and teachers with using it in general and specifically within Japan.

2.2 SLA in English communication classes

2.2.1 Oral communication and SLA

The processes through which students may acquire a second language needs careful consideration, so that courses can be designed to assist that acquisition. The two main, but contrasting, perspectives for this are the *nativist* and *interactionist* viewpoints, which will now be

discussed. Within this thesis, I do not make any conclusions as to which of these viewpoint is more likely to be correct for SLA. I explain how they both relate to orally interactive tasks, as well as highlight the lack of empirical data which currently exists to show how language use can develop across time within interactive tasks (such as group discussions), which the data in the study provides.

The *nativist* viewpoint within SLA is that it is the natural internal mechanisms of a student working on the language they hear and prepare to say which leads to SLA and resultant communicative competence. A specific example is Krashen's (1985) *Input Hypothesis*, which suggests that learning a language is more about acquiring it through input, rather than learning it through interacting and responding to others. In addition to this, the *Output Hypothesis* (Swain, 1985, 1995) states that it is important for students to orally produce language in order to improve at it, because it promotes noticing, experimenting, and becoming more structured and accurate at speaking through self-reflection of mistakes made and difficulties experienced. A more in-depth discussion of these theories is beyond the scope of this thesis. However, they both suggest that improving the oral communicative competence of students is mainly about having them practise listening to and understanding the speech of others, as well as producing their own speech by going through the internal processes of language production described as *conceptualizing, formulating* and *articulating* (De Bot, 1992; Levelt, 1989). In a second language communication course, this could involve a high focus on listening tasks and monologue speeches for example.

However, the *interactionist* viewpoint of SLA is that language acquisition occurs as a result of social interactions between speakers, rather than just the internal processing of input or output of speech. *Sociocultural Theory* (Vygotsky, 1978, 1986) highlights the need for speakers to use

the language to interact with others in a social context, in order for meaning-negotiation and hypothesis-testing to be present and drive SLA. Furthermore, Long (1989) stated that Krashen's input hypothesis (discussed above) is only practical for SLA if the input is comprehensible and if interactions between learners help clarify misunderstandings. The *Interaction Hypothesis* (Hatch, 1978; Long, 1996) states that interaction created by tasks helps students improve their language use, as opportunities are provided to attend to problems using the language within specific contexts. The modifications which take place in the negotiation of meaning and new utterances which are used to clarify meaning between speakers as a result, are believed to lead to SLA. This theory of learning has also been referred to as the *Interactionist Approach* (Gass & Mackey, 2007), proponents of which hold that learners who use meaningful and functional dialogue in an interactive way will be practising a more 'authentic' style of language use than within individual tasks and will become better at using the language (Larsen-Freeman & Anderson, 2013, p. 157; Savignon, 2002). Such language practice has been shown in studies to lead to better performance in future interactions (Gass & Varonis, 1994; Pica, 1994). In addition, the *Socio-Interactionist* view of SLA (Firth & Wagner, 2007; Sun, 2011) is very similar to the interaction hypothesis discussed above, as it suggests that interactions between speakers are key for SLA, not only because they practise using the language, but also because it is done within a social context and that the social interactions which students have are more supportive of the learning than their cognitive processes of producing sentences of speech. More classroom-based research is required at this time to see how/if students who practise second-language speaking skills through interactive tasks (such as group discussions) will acquire the language, as the above theories

suggest, as little empirical evidence exists to show this (Keck, Iberri-Shea, Tracy-Ventura & Wa-Mbaleka, 2006).

2.2.2 Orally interactive tasks and SLA considerations

Following on from the section above, and assuming that *interaction* plays an important role in SLA, teachers also need to consider some other important cognitive processes involved in the learning. Firstly, the *working memory* of students is limited and will determine how much pre-formulated language they can draw upon whilst interacting with others (Baddeley, 1986, 1993; Baddeley & Hitch, 1974). Skehan (1996, 1998, p. 97) suggests that students have a *limited capacity* for learning, and that tasks which require attention on certain elements in performance leave students without enough 'attention resources' to focus on other elements. However, Robinson (2001, 2003, 2007) opposes this belief, saying students can draw on different 'pools' of attention at the same time. For example, students trying to speak with higher accuracy, by making less errors in speech, may speak less fluently, by speaking more slowly (or vice-versa). Whether students can focus on and improve different aspects of performance at the same time or not requires further research, to better understand any cognitive limitations within learning for students.

Secondly, the *cognitive load* (Candlin, 1987) which orally interactive tasks put on students may cause problems with their learning. Asking students to interact in English requires consideration of what Skehan (1998, p. 99) describes as 'code complexity' (how difficult the language required for the task is), 'cognitive complexity' (how complex the task is to undertake), as well as 'cognitive processing' and 'communicative stress' (the amount of organizing and

processing of language required within the task time available). Teachers must ensure that these demands on students are not so high that they do not prevent interactions between students which are expected to lead to SLA.

Thirdly, there is a clear lack of current research data to link theories of interaction (Section 2.2.1) to SLA in *group* discussions across time. The studies mentioned by Gass and Varonis (1994) and Pica (1994) were only for pair interactions and only considered improvements in language use across very short periods of time. It cannot be assumed that the same effects of interactional tasks will occur within larger group sizes or in the longer-term within classes. As discussed above, some studies (such as Keck et al., 2006) suggest that there is no proven connection between oral interaction and SLA. This study provides more data related to this by making connections between the oral interactions which take place during group discussions and how performance within those discussions changes over time (Section 6.3). This helps teachers see more clearly if time invested in learning through group discussions is in fact leading to improvements in language use or not.

2.3 The group discussion approach to language learning

2.3.1 Potential learning and teaching benefits of group discussions

Orally interactive tasks, involving two or more students, are a commonly used approach to promote language learning, as they are believed to create the interactional setting necessary for SLA to occur (Section 2.2). The interactions which occur between multiple students are also believed to be of a higher 'quality' than within individual or whole-class tasks because of the variation in language use amongst speakers which will lead to comprehensible input necessary

for SLA (Long, 1990; Long & Porter, 1985), although more empirical data is needed to show this. Through working with other students to practise negotiating meaning by clarifying, questioning, responding to questions, disagreeing, and giving opinions, students are more likely to become communicatively competent than practising giving opinions through monologue-style speeches for instance (Lynch & Anderson, 1992; Rignall & Furneaux, 1997). Also, having two or more students discuss topics is believed to be more beneficial for learning than one-to-one with a teacher. This is because discussions between peers are more representative of authentic communication between speakers of a similar level, compared to discussions which are controlled and supported by a teacher (Johnson, 2001).

Other potential pedagogical benefits exist for the use of *groups* (involving three or more students) rather than *individual*, *pair* or *whole class* tasks. Discussions within groups are believed to offer a more positive affective climate than with a teacher or in front of a class for example (Long, 1985, 1990; Long & Porter, 1985). They can be a more intimate, private and supportive setting for students, where making mistakes and receiving feedback on errors creates less anxiety amongst students. Also, group discussions offer students more individualized speaking time and feedback compared to whole-class tasks (Foster, 1998; Long, 1977). They give students more individual freedom in their choice of speech content and language skills to focus on improving, as well as feedback from students within the same group. In addition, if students practise in groups of three or more, teachers will be able to watch a larger percentage of classroom discussions during classes, as there are fewer discussions taking place at the same time (as opposed to a higher number of pair or individual discussions for the same class). This more frequent *Teacher-Based Assessment* style of feedback during class is expected to lead to

better learning over time than feedback given on only test performance (Davison & Leung, 2009). One more pedagogical benefit of group discussions is that they encourage both *collaboration* and *cooperation* in the learning. *Collaborative learning* involves students working in tandem on shared goals and is believed to enhance social and cognitive skills, as each student is an accountable team member (Ahmadian & Tajabadi, 2017; Johnson, Johnson, & Holubec, 2013; Oxford, 1997). *Cooperative learning* requires each team member to work independently on their separate role and responsibilities for the group outcome, and encourages support between group members to do so (Donato, 1994; Lantolf, 1993). If a discussion task requires students to reach an agreed outcome together, such as an action plan related to the topic, both collaboration and cooperation would be expected to be present. Furthermore, leadership skills may develop among some students, due to the collaboration and cooperation which group discussions require (Forsyth, 2000, 2016, p. 255; Hackman & Johnson, 2013, p. 203). Such leadership is also viewed as an important part of completing tasks in groups and valuable as a learning experience for students (Ehrman & Dörnyei, 1998, p. 154).

2.3.2 Challenges for learning and teaching with group discussions

Despite the potential benefits of using group discussions, several potential restrictions to learning with them also exist. When students are asked to undertake work in groups, factors related to the relationships and interactions between speakers can have large influences on the learning. Many of these were reported by students as 'barriers to participation' related to group set-up within the first module of this PhD (Appendix A), such as the personal relationships and differences in language levels between group members. A recently published paper by Poupore

(2016) helps add to this discussion by explaining the importance of a positive and comfortable feeling amongst a group in order for participation and resultant learning to take place. Poupore explains how *Group Work Dynamics* (GWD) are important to consider if students are learning in groups, as the actions of each member can influence the actions of others depending on how collaborative or cooperative, and dominant or passive they are (Storch, 2002). For instance, if students decide to dominate the discussion in a non-cooperative way, then the others may not participate or improve their ability to perform in discussions as a result. The larger the group becomes, the more chance there could be of this happening. Also, even though larger groups may give a teacher more time to observe a lower number of groups with their available class time (as discussed above), larger group sizes may cause two problems. Firstly, the students may find it harder to take speaking turns within their groups, as speaking time will be shared between a larger number of group members. Secondly, the teacher will then have less chance of seeing all of the students speak, as there will be fewer students speaking at the same time within class compared to a when there are a larger number of groups. These points may be especially important to consider for students with lower speaking abilities within groups, as it is believed that having a lower level within a discussion may be enough to prevent students from speaking at all (Foster, 1998). If groups within classes contain a variety of speaking abilities, then this may be a problem for the use of group discussions. Thus, the learning undertaken and changes in performance over time by both low and high performers needs consideration. This is addressed in the study in this thesis by categorizing the participants as low and high performers, based on the words they say in discussions (Section 5.4.2), and analyzing differences and similarities between the performances of the two groups across time.

2.4 Task-Based Language Teaching (TBLT) group discussions

2.4.1 Communicative Language Teaching (CLT) and TBLT discussions

Communicative Language Teaching (CLT) is an approach to teaching which is focused on developing the communicative competence of students within a second language (Hymes, 1972, 1974) by helping them develop their ability to use the language both 'interactionally' (establishing and maintaining contact with others) and 'transactionally' (using language referentially to exchange information) (Brown & Yule, 1983; Ellis, 2003, p. 27). It can be used with what are described as a 'weak' or 'strong' approach (Howatt, 1984). The *weak approach* is focused on identifying and teaching specific components of communicative competence. This may mean having students learn about the notion and functionality of disagreements before actually trying to disagree with each other using a second language in discussion tasks for example. The *strong approach* is focused on the belief that language is acquired through communication. For discussions, this would mean that students would learn about disagreements within discussions through simply disagreeing with others when opportunities arise to do so.

Task-Based Language Teaching (TBLT) is an approach which adheres to the belief of the strong approach of CLT discussed above. In particular, it follows the belief that students will improve their communication within a second language if they practise completing what are called 'tasks'. Research in the last four decades has been unable to agree on precisely what a task should involve for students, but does agree on some fundamental points for the learning. In short, tasks should be used to promote meaning-focused language use in situations which students carry out in the same way as they would in the 'real-world' (Ellis, 2003, p.6; Long, 1985, p.89; Nunan, 1989, p. 6; Skehan, 1996). This means that teachers should not interfere with the

'authenticity' of a task by focusing on or planning specific forms of language (such as lexical items or sentence grammar) before or during students undertaking them, but only after the task has been completed. For group discussions, this would mean that students do not focus on any specific language points or discussion skills beforehand, but only after they have used them with others to communicate within a meaningful discussion involving processes, interactions and outcomes similar to a discussion held outside of a language classroom.

2.4.2 TBLT versus Present-Practise-Produce (PPP) for group discussions

Using a TBLT approach of focusing on negotiating through the 'meaning', rather than the 'form', of the language during tasks is promoted as an effective way to nurture SLA (Bygate, Skehan & Swain, 2013; Ellis, 2003; Littlewood, 2004; Long, 2014; Robinson 2007, 2011; Thomas & Reinders, 2015; Willis & Willis, 2008). However, more traditional communicative approaches to language teaching are often preferred by teachers in Asia (Iwashita & Li, 2012). The most common of these is the *Present-Practise-Produce* (PPP) approach. The important difference with PPP is that students begin practices by being introduced to specific language forms (Present), practise them in specific contexts chosen by the teacher (Practise) and are then asked to use them within a final practice stage designed by the teacher (Produce) (Gower & Walters, 1983). Such practices are believed to help students improve their ability to use specific language forms more than with TBLT, as the language which is practised can be controlled more by the teacher and matched up more closely with the assessment performed later on, compared to the more ‘open choice’ use of language with TBLT. This has been recognized as a useful routine for practising important parts of speech (Swan, 2005), but although the practising of such

language points may be expected to improve the ability of students to say them in class, no evidence exists to show that a PPP approach can improve language use during a *group discussion*.

Thus, TBLT may be a preferred approach to PPP for discussion tasks, because if students use language to undertake communication with much less external direction or restrictions on the language they must use within PPP, they will be able to *notice* the difference between their own ability and that of grammatical rules being used more (Batstone, 1996, p. 273). This is because the freedom of language use which TBLT gives students makes them more aware of what they can and cannot communicate, rather than just repeating language points already provided to them by PPP, and be able to focus more on improving those gaps in their abilities afterwards. This stage in learning is called *Consciousness-Raising* (CR) and helps students internalize grammar structures in language by drawing their attention to it (Ellis, 2002, p. 168). It is argued that other more form-focused approaches to learning, such as PPP, cannot raise the awareness of students' own language gaps which need improving in this way, but only the awareness of specific forms which have been selected by a teacher and practised beforehand (Ellis, 2003, p. 29). For classroom group discussions, the vast number of language points which may be required for students to exchange ideas and possibly reach decisions together may make PPP style practices impractical. It is my belief as a teacher that identifying gaps in ability and improving those weaknesses in language use for many different students within the time available in language courses requires the freedom of TBLT to do so.

Despite all of these potential benefits for learning of using a TBLT approach to group discussions, little evidence exists of the actual improvements in performance which it may result

in within classrooms. Thus, the study in this thesis addresses this using an analysis of discussion recordings across a semester to see how performance actually changes to help researchers and teachers better understand the effects which TBLT has.

2.4.3 Challenges for learning and teaching with TBLT group discussions

One of the potential problems with the freedom in language use discussed above in relation to TBLT is that it creates a large amount of possible focuses for students within discussion tasks. Although such freedom in language use may promote noticing, consciousness raising, and resultant SLA (Section 2.4.2), it may also leave students confused about what content, actions or language they should focus on within tasks (Burrows, 2008). It is clearly expressed in the literature that for improved language use to occur with a TBLT approach, students must reach an 'outcome' (Prabhu, 1987, Skehan 1998) or obtain an 'objective' (Bygate et al., 2013, p. 19) for the task. Ellis (2003, p. 8) helps clarify what this means for tasks by describing an *outcome* as something that students arrive at after completing a task, such as a story or a list of differences. He compares this to the *aim*, described as the pedagogical purpose of the task, which is eliciting receptive and productive meaning-focused language use. However, without any outside guidance or influence on what language students should be using or how they should interact with the language (which PPP can provide), teachers cannot be certain that students will undertake the processes which TBLT expects to lead to such SLA. For oral tasks, students can decide which performances to focus on or ignore, such as 'sacrificing' accuracy in oral tasks for fluency, (Cuestra, 1995; Foster, 1999), and can even choose to switch to their L1 to complete discussions away from the view of a teacher (Carless, 2007a). In such cases, students may still be able to

produce the same discussion output (a list or spoken summary of their agreed discussion outcome in the L2 for example) as those who did not. In addition, studies analyzing which of these focuses will actually result in improvements in oral language use over time are scarce. An important question for group discussion tasks is *how important is the focus on the outcome (what students decide upon) and the aim (the processes of language use they go through during the task) for promoting learning?* The study in this thesis addresses this question by examining how Japanese university students' focus on either the outcome (product goals) or aim (process goals) of their group discussions will improve their task performance over time (Section 2.4.3 and RQ2).

Another potential problem for TBLT group discussions is that some students may benefit less than others from this style of learning. Students with a higher proficiency in the L2 have been found to benefit more than those with a lower proficiency from a TBLT approach to oral tasks (Burrows, 2008; Tseng, 2006). This is believed to be due to the complexity of tasks and the freedom of language use required during task time. Therefore, it is important for teachers using group discussions with students of mixed levels to consider them on an individual basis for learning, performance and progress over time, as the study in this thesis does by looking at differences in performance change for low and high participators (Section 5.4.2). With a meaning-focused TBLT approach, where students do not receive specific guidance on or support with their language use prior to or during discussions, lower-level students may struggle to participate in groups with higher-level students. This may lead to motivational issues for those students who may require more guidance and support on an individual-basis to improve over time. A planning stage has been found to help students improve their overall performance in

tasks (Guará-Tavares, 2011, 2013, 2016) with improvements made within oral task participation and fluency (as was concluded in the second module, as shown in Appendix B), accuracy (Bygate, 1996, 2001; Bygate & Samuda, 2005; Lynch & McClean, 2000, 2001), and complexity (see Ellis, 2009 and Javad Ahmadian, Tavakoli, & Vahid Dastjerdi, 2015 for recent summaries of related research). By allowing students to plan for discussion tasks, the approach then becomes more of a 'Task-Supported Teaching' approach (Ellis, 2003, p. 28), where students can practise language items already practised before a task (more of a PPP approach). One concern with this is that students may not then interact as much in a meaningful way as the tasks become what Long calls 'synthetic', as opposed to 'analytic' in nature (2014, p. 7). This may lead to a reduction in 'real-world' interactions which occur, without a prior focus on language forms, which is described above as important for the pedagogy of TBLT.

TBLT also faces challenges with finding the required time to explain and train students and teachers, so that is becomes understood, accepted and used in the intended way for learning (Lai & Lin, 2015; Waters, 2009). TBLT has been reported as an inappropriate approach to learning on a world-wide scale, mainly because of the need for more time for tasks than other approaches (such as PPP), and the under-prepared feeling that some teachers report (Ogilvie & Dunn, 2010; Sparks, 2010; Van den Branden, 2006, p. 217). The requirement to 'facilitate' discussion tasks, by giving students control of the process of discussions, rather than guiding them in the language they should be practising, can cause confusion and power-struggles for some teachers (Ellis, 2003, p. 271; Stroud, 2013). PPP has often been chosen as an approach to teaching over TBLT as it offers teachers more control than TBLT, and PPP is believed to have an 'excellent relationship with teacher training and teachers' feelings of professionalism', and 'lends itself very neatly to

accountability' (Skehan, 1998, p. 94). TBLT often does not get used in classrooms, as the PPP approach is believed to create clearer language use goals which better match assessment criteria and exam results (Butler, 2011, Thornbury, 1999, p. 64). The study in this thesis uses the addition of GSF with discussions to see if it can help with these potential weaknesses of TBLT, compared to PPP, by providing more clarity and connections between learning and assessment for both students and teachers (Section 4.5).

A final question for a TBLT approach is *can students who practise discussions with such freedom of language use then transfer what they learn to new discussions afterwards?* In general, skills practised in one task (whether it be related to language learning or not) would be expected to be transferred to another, if the goals, method and approaches are similar (Blume, Ford, Baldwin, & Huang, 2010). Although much discussion of transfer is only theoretical (such as Barnett & Ceci, 2002; Haskell, 2001), some transfer has been shown for low-level speakers of English from one orally interactive language learning task to another (Benson, 2016). Students with lower-speaking abilities within a study group were able to use conversation skills learnt for everyday situations (such as giving directions) in subsequent practices. However, no data exists to show that such transfer can take place across time for *group* discussion tasks, involving three or more students, especially if the topics are changed each week. The study in this thesis addresses this by analyzing performance changes across time (using different weekly topics), which is crucial for teachers who use such an approach to understand what improvements may or may not be occurring in performance.

2.4.4 Challenges in Japan for TBLT group discussions

As discussed in the introduction, the overall intent of communication courses within Japan is to educate a generation of students who can exchange their thoughts and feelings through using the English language (MEXT, 2009, 2013). The use of a TBLT approach to improving the oral communicative competence of students can be argued as an appropriate approach, due to the expectation that students will 'notice' gaps in their language use and improve upon them with a higher understanding of their own ability (Section 2.4.2). However, for Asian students, TBLT faces not only the difficulties mentioned above (Section 2.4.3), but additional challenges related to expectations for learning from students, teachers and institutions (Carless, 2004; Lai, 2015; Littlewood, 2007).

The Japanese students within the first module of this PhD reported many 'barriers' to participating within TBLT-style discussions (Appendix A), including a 'fear of making a mistake' during speech which can cause students to feel anxious and remain silent (Chang, 2011; Tsui, 2001; Williams & Andrade, 2008; Woodrow, 2006). Such problems with participation are not uncommon among students in other Asian countries. Asian students often report feeling underprepared to talk within TBLT tasks and switching to using their first language to avoid making mistakes (Burrows, 2008; Carless, 2007a). Because TBLT is an unfamiliar approach for Asian students, studies have shown that they often prefer to practise the language forms before oral tasks (with a PPP-style approach), so that they will be more accurate in their language use (Lai, Zhao & Wang, 2011), rather than being concerned with learning through interactions, noticing and consciousness-raising, as TBLT is designed to promote. Therefore, an important challenge for the use of TBLT discussions within Japanese classrooms is to make students feel

relaxed and confident to undertake discussions through being aware of what is expected in terms of interactions and/or outcome of tasks (which may not always be recognized as the same by teacher and their students), as well as feeling prepared to interact with others in the L2 to do so. This was a benefit I expected to come from students using the GSF in the study and is analyzed for its specific effects on performance and learning (RQ2 and RQ3).

As discussed in the previous section, TBLT is often rejected as a teaching methodology, due to a lack of clear connections between the learning undertaken and the assessment criteria which most students are focused upon. This has also been reported as a problem in Asia, (Deng & Carless, 2009; Lai, 2015), where students, such as Japanese, are often asked to undertake knowledge-based, summative, vocabulary and grammar based practices which teachers feel will prepare them better for their exams. Such a PPP-style approach to learning is often preferred, as it directly addresses grammar points which are tested at the end of courses and teachers report feeling more confident that they can direct the learning of the students to pass such tests (Carless, 2007b). One other problem for TBLT is that teachers in Japan often report not being comfortable or clear about their own role within the learning, and become uncomfortable with the power-shift from a 'teacher', who may control the language being practised, to a 'facilitator', who leaves the contents and outcome of tasks up to the students (Stroud, 2013). Thus, if TBLT discussions are to be used within language courses in Japan, it is important that teachers can see clear connections between 1) the processes and outcomes of discussions, and 2) the assessment criteria for the course (Butler, 2011). This would help them understand how they should (or should not) direct student behavior and language use to nurture improvements in test-related performance over time. I believe that the GSF used in the study can help teachers in this way by providing

important feedback across time (Section 4.3) which can be used by teachers to support the learning (see Section 7.2 for recommendations).

2.5 Chapter summary

This chapter explored the potential benefits and challenges of a TBLT group discussion approach to language learning and teaching, and also highlighted the lack of current classroom-based research data within this area. According to the theories and research discussed, TBLT group discussions may be an appropriate approach to the learning of a second language for several reasons. Firstly, the interactions taking place between students may support more SLA than non-interactive tasks (such as monologue speeches) due to the need to contextualize and use the language to react to others. Secondly, learning as a group was discussed as preferred to learning as individuals, in pairs or as a whole-class, because it provides the greatest opportunity for a variety of language use, individualized speaking time and feedback from others, positive and motivating environments, as well as the practising of collaboration, cooperation and leadership skills. Thirdly, a TBLT approach was described as being more appropriate than PPP, as the negotiation of meaning with authentic language use among students would be expected to encourage higher levels of noticing and consciousness-raising and resultant SLA. However, none of these benefits have been proven with research for group discussion tasks. In particular, more data is needed to see whether TBLT classroom discussions can actually result in improved performance across a course. Also, more data is needed on the effects of having students focus on the *outcome* (final group decisions) versus the *aim* (language use during the task) of a discussion. This chapter highlighted several contradictions between the beliefs of *TBLT* (that task

outcome is essential) and *interactionalist* beliefs (that language use is more important than the task outcome). The study addresses this by analyzing the performance progress and learning of two separate groups of students for each of these two focuses. Furthermore, it is important to analyze performance for individual students, as this chapter also discussed how group work can complicate learning with the added potential barriers of limited speaking time between larger numbers of group members, differences in language levels and Group Work Dynamics (GWD) making it difficult for some students to speak. The study in this thesis also addresses these important points by analyzing changes in group discussion performance across a semester for low and high participating Japanese university students with a low-level of English.

CHAPTER 3. DETERMINING ORAL GROUP DISCUSSION PERFORMANCE

3.1 Introduction

This chapter discusses how group discussion performance may be best judged, using a variety of measures, which can then be used to help guide and assess learning (using GSF in the study). It specifically addresses observable oral performances (rather than non-verbal performance, such as listening skills or body language), which were used to decide measures for use within the study in this thesis (Section 5.6.2). The first half of the chapter discusses the use of commonly adopted participation and CAF measures for oral group discussions. The second half explains other important measures of performance which also need consideration within group discussion settings. A discussion of the contradictions between the measures discussed in the first and second half of the chapter is then given. The chapter concludes with a summary of the use of these measures to evaluate the second language ability of different students within a group discussion, as well as the challenges associated with creating a well-balanced picture of performance.

3.2 Participation and CAF measures

Determining student oral performance within language tasks, based on what the teacher can observe, is of high importance, but also very difficult (Bonk & Ockey, 2003; Fulcher, 2003; Fulcher & Davidson, 2007, p. 24). In addition to student participation measurements, the most common three constructs of oral linguistic performance analyzed are *Complexity*, *Accuracy* and *Fluency* (collectively known as CAF). Many choices exist for measures to represent these

constructs, which creates inconsistencies in both the approaches to measurement and findings for research (Housen & Kuiken, 2009). For group discussion tasks, it is important to select appropriate measures to represent performance which connect to past research (which have been mainly focused on monologue oral tasks), as well as any additional conditions which a group discussion set-up may require. These are now explained.

3.2.1 Participation

The number of *words spoken* is a quick and easy way to assess the level of participation which a student is showing in a discussion which can be counted using recordings and transcripts. The number of *turns taken* is a useful second measure to complement words spoken, as it reveals more about the number of times a student decided to participate in a group discussion by speaking. These two measures were discussed and selected as appropriate for use with low-level Japanese students in Module Two of this PhD (see Appendix B) and will not be discussed in any great detail here. However, they reveal nothing about the CAF or content of those turns, and so further measures are required to better understand performance.

3.2.2 Fluency

Oral fluency is said to represent the 'ease' and 'smoothness' of speech (Guillot, 1999, p. 14; Koponen & Riggenbach, 2000, p. 8; Lennon, 1990) and is commonly split up into dimensions of *speed*, *breakdown* and *repair* within applied linguistic research (Skehan, 2009; Tavakoli & Skehan, 2005). Firstly, *speed* is usually measured using speech rates (syllables spoken per minute) and shows how quickly a student is able to articulate their speech when they

have a speaking turn in a discussion. Measures used for this are usually either Speech Rate A (syllables spoken per minute, known as SRA), or Speech Rate B (SRB) after removing the repetitions and reformulations to provide perhaps a truer picture of fluency (see Sangarun, 2005; Tavakoli & Skehan, 2005; Yuan & Ellis, 2003). Secondly, even if a student can deliver speech at a fast rate, *breakdowns* which occur in speech during speaking turns need to be looked at as indicators or fluency. Pauses (more than one second in length) and L1/L2 fillers (such as 'um') are such breakdowns, as they show problems with articulation when explaining ideas or responding to others (see Mehnert, 1998; Skehan & Foster, 2005). Finally, *repairs* in speech, such as repetitions and reformulations, have generally been perceived negatively in research, as signs of a student struggling to deliver speech (see Bygate, 1996; Elder & Iwashita, 2005; Foster & Skehan, 1996; Gilabert, 2007; Kawauchi, 2005; Skehan & Foster, 2005; Tavokoli & Skehan, 2005). As these measures of fluency were already discussed and determined as appropriate for low-level Japanese student discussions within Module Two of this PhD (see Appendix B), they will not be elaborated on anymore here.

3.2.3 Accuracy

Accuracy within oral language use refers to how much a speaker deviates from the normal usage of that language (Wolfe-Quintero, Inagaki & Kim, 1998, p. 62). This is considered important for second language tasks, as it shows how 'accurate' the language of a speaker is when compared to that of a person performing the same task in their first language. This type of analysis seems straight forward, but one large challenge for such assessment is the definition of an 'error' for discussions. Within second language use, *linguistic* errors in speech, such as

grammatical ones, are perhaps simple to identify within an analysis of a discussion transcript. Linguistic accuracy is often quantified using measures such as *errors per 100 words* spoken (see Mehnert, 1998) or Percentages of *Error-Free Clauses* (PEFC) within sentences (see Foster & Skehan, 1996). However, if group members use the target language in a way which may deviate from use by first language speakers, but is understood as normal usage by the group, it is up to the teacher to decide if the error is in fact an error (Housen, Kuiken & Vedder, 2012, p. 4; Pallotti, 2009). For instance, the use of "don't mind" in Japan is intended to mean "please don't worry about that" in English. Although this would be perceived as a type of error in spoken English (semantic or pragmatic perhaps), some teachers may decide to not count it as an error, as Japanese students would be expected to understand its meaning.

In addition, spoken errors can also include 'non-appropriate' or 'unacceptable' uses of language (Housen & Kuiken, 2009). Within discussions, identifying such cases is of course much more difficult, as appropriacy and acceptability of language use require more consideration of additional skills beyond linguistic errors. This may refer to 'sociocultural' errors, where students do not express messages in the appropriate way to match the social or cultural setting (Hymes, 1972, 1974; Leung, 2005a). More specifically, this could include errors with *social* (related to age, gender, status, and social relationship), *stylistic* (politeness, genre and spoken register), and *cultural* (perhaps against cultural norms within communication between Japanese) factors which may be seen to cause communication problems between group members (Celce-Murcia, 2007, p. 45; Celce-Murcia, Dörnyei & Thurrell, 1995). However, not only are these factors very difficult to analyze and assess in terms of 'accuracy', but a non-Japanese teacher may not even be aware of these social norms between students.

Considering the difficulties discussed here, perhaps the most appropriate way to assess accuracy for low-level learners within discussions would be by counting linguistic errors (grammatical in nature with measures such as errors per 100 words or error-free clauses), as there would be less analysis needed of social, stylistic and cultural appropriateness. This may not be an entirely objective approach to assessment, but is appropriate for low-level learners (such as those in this study), who often exhibit a high number of grammatical errors in speech, which should serve as a starting point for assessing performance in terms of accuracy. Teachers could decide to move beyond this simple focus on grammar and assess social, cultural and stylistic accuracy if they wished to, but this would require more time, leaving less time to analyze other performance measures which may be more fundamental for assessing low-level students.

3.2.4 Complexity

Complexity is perhaps the most difficult of the CAF measures to understand and measure. It is commonly described as the 'size, elaborateness, richness and diversity' of a student's L2 system (Housen & Kuiken, 2009; Housen et al., 2012). It is important to see students trying to improve such complexity in target language use, as it shows they are taking risks in order to improve their ability to use it in a more varied and enriched way (Skehan, 2009). Spoken complexity is often classified as 'syntactical' or 'lexical' in nature. *Syntactical complexity* refers to the complexity of grammar usage within 'T-units', which are the shortest grammatically allowable sentences during speech (Hunt, 1965). The most common variables used to assess syntactic complexity within monologue oral tasks include *clauses per t-unit*, *words per t-unit* and *words per clause* (see Norris & Ortega, 2009 for an overview of relevant studies). However,

these measures do not give a clear overall picture of how complex speaking turns in discussions are.

For *group discussions*, where turn-taking has to take place, an analysis of *turn complexity*, using measures such as *words per turn* (Philp, Oliver & Mackey, 2006), may be an important addition to better understand overall syntactical complexity for individual students. Students who say more words within speaking turns should be judged to be using the language in a more complex way than those who use shorter turns, as they are demonstrating an ability to form longer turns to explain their thoughts. However, within group interactions, long speaking turns may cause other problems for assessing performance, which are explained later in Section 3.3.1.

Lexical complexity refers to the diversity and frequency of use of specific words within speech (Norris & Ortega, 2009). Common methods of assessing this type of complexity include the use of a *type-token ratio* (the number of different word types divided by total words) (Skehan, 2009; Yu, 2010) and the counting of occurrences of less frequently used words in English speech, according to corpus databases (Skehan, 2009). Changes in these measures over time show improved/worsened complexity by speakers, as this demonstrates an ability to use a larger/smaller variety of words within speech. However, the reliability and validity of this measure is questionable for discussions, as the language use of the students depends on the learning goals and discussion topic at hand. Students who improve their frequency and variety of word usage may not necessary be able to demonstrate this if the discussion topics at hand do not require or encourage such words to be used. This needs consideration when using lexical complexity to assess discussion performance, and a simpler analysis (such as syntactic

complexity) may be adequate and more reliable for representing performance. In the study reported in this thesis, an analysis of lexical complexity was not undertaken due to the limitation in word count and my belief that the influence of using different topics each week, with different vocabulary needs, would distort the data. However, this missing analysis is discussed as one possible future research direction in Section 7.3.

3.3 Additional performance considerations

When determining the performance of students within group discussions, additional considerations, which go beyond the CAF measures discussed above, are needed in order to determine overall communicative ability (Clapham, 2000; Fengying, 2003). Fulcher (2003, p. 39) states that speakers of a language might be able to 'use the grammar of a language, pronounce the sounds and speak fluently, but this may not mean they can communicate well.' This is certainly true, as students who prepare and deliver a speech, for example, are not showing their ability to interact with others. Therefore, in order to assess the ability of students to take part in discussions, a broader analysis is required. Oral communicative competence has been described as a demonstration of *grammatical* knowledge (linguistic elements discussed above), *social* context knowledge (knowing how to interact appropriately with others in order to complete tasks), and *discourse* functions (how to combine utterances and communicative functions with respect to discourse principles) (Canale, 1983; Canale & Swain, 1980). This means that in order to understand how well a student can perform in a second language discussion, teachers must also look more closely at the content and functions of the turns they take. This would mean an analysis of the types of interactions between group members, how

students make their messages clear to others, as well as efforts made to complete the task (using what are called 'outcome-promoting turns' later on in the study). A discussion of these is given in the next three sections with important questions summarized for teachers at the end of each one.

3.3.1 Group interactions

Within orally interactive language tasks, a communicatively competent speaker is described as someone who can communicate through *questioning, responding to questions, disagreeing,* and *giving their own opinions* during interaction with others (Lynch & Anderson, 1992; Rignall & Furneaux, 1997). Such interactions have been described as a part of 'real world' communication, in which speakers demonstrate their ability within classrooms to use the language in similar ways to which first-language speakers do in their daily lives for similar tasks (Bachman, 2000; Bachman & Palmer, 1996; Clapham, 2000; Ghaith, 2002). Responding to others is viewed as an important demonstration of communicative competence, often called *Interactional Competence* (Celce-Murcia, 2007; Celce-Murcia et al., 1995). This is the belief that students who are able to *backchannel*, *ask questions* and *agree or disagree*, show a higher level of communication than those who do not. This is because such actions show the ability of students to listen to, understand and respond to others in the language, which goes beyond just listening to or producing speech as unrelated acts. It demonstrates a higher level of communicative competence, as such actions would be required when using the language in a classroom discussion or when speaking with others in English away from the classroom.

Also, a sub-category of interactional competence is called *Conversational Competence* (Sacks, Schegloff & Jefferson, 1974), which refers to the ability of students to work together in a group effectively through *turn-taking* to achieve a conversational goal. In other words, groups who are able to work collaboratively to offer and take turns together are more likely to be recognized by an observer as communicatively competent than those who lack such 'smoothness' and cooperation in changing speakers. An appropriate assessment of ability within group discussions should also consider these interactional factors. By only taking CAF measures into account when determining performance, it is entirely possible that students could be perceived to be performing fluently, accurately and with high complexity, but may not actually be working together as a group on the task.

A significant challenge for determining overall performance within a group discussion is the balance between measures of linguistic complexity and interactional competence discussed above. Complexity in oral tasks is partly measured by looking at how long the speaking turns are by using measures such as words per turn (Section 3.2.4). However, if students speak in very long turns to explain their opinions, this leaves less time in a discussion for responses from others on those opinions, which as discussed earlier, is also considered important. On the other hand, if students show a higher level of interactional competence and turn-taking ability via more frequent, but shorter, speaking turns (with high levels of participation represented by lots of very short questions and answers for example), the complexity of language use will be viewed as very low according to words per turn. This poses an important question for teachers; *should complex individual speaking turns, or shorter and more frequent interactions between speakers, be considered better performance within a group discussion?* By answering this question before

communication courses are designed and begun, the focus of students towards complexity of speech and/or frequent interactions within discussions can be more appropriately directed using support (as addressed by the GSF in the next chapter).

3.3.2 Clarity of communication

Another important way in which students can demonstrate their ability to hold discussions is by selecting and sequencing words and utterances within speech in order to make clear messages which others can understand. This is often referred to as *Discourse Competence* (Celce-Murcia, 2007; Celce-Murcia et al., 1995). A similarly focused measure of performance, is another sub-category of interactional competence (Section 3.3.1) called *Actional Competence* (Celce-Murcia, 2007), which describes the ability of students to express and elaborate on their opinions with *reasons* and *examples* in an understandable way. It is important within discussions that such clarity of opinions exists or other group members may not understand or be able to react to those turns with their own thoughts. However, if student opinions are not understood, then further efforts made by the same speaker to clarify meaning can also be considered signs of communicative competence. This could come in the form of *repeating* or *reformulating* speech within the same turn, or repeating or *paraphrasing* a turn to help clarify meaning (Hedge, 1993; Hymes, 1972, 1974; Tarone, 1980). Also, other group members could demonstrate communicative competence at that point by helping the speaker with language difficulties to aid understanding between group members.

However, several contradictions between CAF measures and the ability to communicative clearly within group discussions can be seen in the research literature. With

regards to fluency, repairs within speech, such as repetitions and reformulations, are often seen as signs of poor fluency (Section 3.2.2). However, these actions, as well as paraphrasing and students helping each other in clarifying meaning, were discussed above as important signs of discourse competence within group discussions. Also, it is entirely possible that students can demonstrate high levels of accuracy and complexity, but with turns which are very difficult for other to comprehend. A clear example of this is given by Pallotti (2009, pp. 11-12) using two utterances with the same intended meaning. The first utterance is "colorless green ideas sleep furiously on the justification where phonemes like to plead vessels for diminishing our temperature." This utterance could be considered to have high levels of complexity and accuracy, as well as the potential to be said fluently by a speaker, but it is very hard to understand its meaning. The second utterance is "No put green thing near bottle. Put under table." Although the CAF measures for this utterance are clearly lower than for the first, the intended message is arguably much clearer to a listener. Hence, an important question for teachers is *should turns with high CAF, but problems with clarity in meaning, be considered indicators of higher or lower discussion performance than turns with low CAF, but a very clear meaning for listeners?* Again, by answering this question prior to courses, the focus of students' performance towards linguistic and/or clarity in communication can be more appropriately directed using support (such as with GSF in the next chapter).

3.3.3 Discussion outcome

Another important performance consideration is how the speaking turns which students undertake in discussions might promote the desired outcome (often an agreed choice or action plan related to the topic for example). If a communicative task has a required outcome, then it is important that students are not only explaining their opinions on the topic and reacting to each other, described as the learning 'aim', but that they use such interactions to achieve the desired task 'outcome' (Ellis, 2003, p. 8). Otherwise, a group discussion may be judged to lose an important part of its overall communicative purpose, which is achievement of a goal through collaboration using the language, although this has not been demonstrated by any research findings (Section 2.4.3). Also, it is very hard to assess individual performance based on the outcome of an oral task, as it tells us very little about what students said during the task itself (Long, 2014, pp. 332-334). However, an analysis of the efforts made by individual students within groups to reach the outcome during the discussion time may provide a clearer picture of performance. If an outcome is required from a group task, then students who take turns to promote that outcome (by checking the final opinions of others or checking a final group decision for example) should be considered to have performed better than others in that respect. In contrast, students who discuss things completely unrelated to the topic at hand should be viewed as working counter-productively to the overall group goal and performing at a lower level than others.

A challenge with assessing performance in this way is the valuing of action related to what Ellis (2003, p. 8) called the 'aim' (students expressing and elaborating on opinions, as well as responding to each other) and the 'outcome' of the task (final decisions made on the topic via

collaborative and cooperative behavior during the task time) (Section 2.4.3). Both can be argued to demonstrate ‘good’ performance, but determining how important they both are depends on the belief of the teacher. Therefore, an important question is *should students demonstrate their performance in discussion by interacting with others about individual opinions on the topic, or by working collaboratively and cooperatively with others to reach a final decision together (or both)?* As mentioned within the previous two sections, teachers who answer this important question early on, can guide student efforts and focus towards either (or both) individual-style or group collaborative-style discussions using external support (such as GSF in the next chapter).

3.4 Chapter summary

This chapter discussed what the most appropriate methods may be for determining the performance of students within classroom group discussions. Firstly, it drew on a variety of relevant research to discuss how this should be done using measures of participation and CAF, such as words spoken, turns taken, speech rates, breakdowns, repairs, errors, lexical complexity, sentence complexity and turn complexity. Secondly, it drew on other studies to explain how assessing performance should go beyond these linguistic measures by also analyzing the specific interactions between group members (asking questions and disagreeing for instance), how clearly students can make their turns understood by other group members, and the degree of focus of students to reach a discussion 'outcome' or not. Contradictions in the research literature between these linguistic and discussion measures were then highlighted, as well as the resultant difficulties which this creates for teachers when trying to categorize what low-level students may need to say in a discussion to be considered ‘good’ performers. The main challenge discussed

that students who score highly in terms of CAF measures may not necessarily be able to demonstrate the ability to interact frequently and clearly with others (and vice versa) because of conflicting performance measures (see the final question posed at the end of each section). Therefore, it is important to decide how to grade performance using a selection of measures which will accurately determine the level of students within group discussions. The study addresses this in RQ1 by analyzing teacher opinions of appropriate measures for low-level Japanese university students (Section 5.5). The next chapter discusses how measures which are selected by teachers may then be incorporated as task performance goals and goal-setting and feedback across time to support the learning.

CHAPTER 4. GOAL-SETTING AND FEEDBACK (GSF) FOR GROUP DISCUSSIONS

4.1 Introduction

This chapter is made up of four main sections which discuss the possible impacts on task performance and learning of integrating Goal-Setting and Feedback (GSF) into classroom group discussions, as explored within the study. The first section discusses how goals related to language use may or not support learning within classroom oral tasks, such as discussions. The second section discusses how feedback on performance across time may be expected to support learning and improve oral task performance across a language learning course. The third section discusses how task performance scoring rubrics can assist learning and how they might be applied to classroom oral tasks. The fourth section uses the research literature from the previous sections to discuss how a GSF system for group discussion tasks might be best designed, in terms of goal focus and performance scoring, to support learning over time.

4.2 Goal-setting and learning

4.2.1 Task goal-setting, motivation and engagement

The setting of goals within learning is an effective way to motivate students to undertake classwork (Dörnyei, 2001, p. 29). Goals can provide a framework for students to understand how they should interpret and react to tasks, and clearly defined goals should therefore lead to both higher levels of *motivation* (Maehr, 1984) and higher levels of resultant *engagement* (Klem & Connell, 2004). For oral tasks, this refers to students becoming more *behaviorally* engaged (taking actions such as speaking more and taking more turns), *emotionally* engaged (having more

positive feelings towards undertaking the task), and *cognitively* engaged (investing more mental effort to undertake the task, especially by overcoming difficulties negotiating meaning in a second language) (Fredericks, Blumenfeld, & Paris, 2004; Philp & Duchesne, 2016).

The introduction of performance goals has been connected to better task performance in recent classroom-based studies. Gardner, Diesen, Hogg and Huerta (2016) found that introducing specific goals related to task performance within medical skills training, especially those focused on developing strategies and processes for completing the task, resulted in better task performance than a task with the same 'do your best' condition. Furthermore, studies have shown that the use of goals with just one task can 'prime' students to keep focusing on performing towards the same goal in successive tasks across time, consciously or sub-consciously (Hart & Albarracín, 2009; Parks-Stamm, Oettingen, & Gollwitzer, 2010; Vohs & Baumeister, 2016).

For this thesis, the effects of goal-setting on performance needs to be connected more specifically to group discussions. Martin, McNally and Taggar (2016) showed that the oral behavioral engagement of students within group discussions can be immediately increased by introducing a clear task goal. By simply asking for a minimum of 50 items for a list of uses for a knife within discussions, rather than a 'do your best' effort, the lists created by groups were much longer. Such research shows that the presence of performance goals can push students to perform better within oral discussions. However, the study was not within a language learning setting, did not analyze the specific interactions which took place within the tasks, and did not consider the feelings of the students towards the learning undertaken.

Although the studies discussed above did not provide any longitudinal data, Goal-Setting Theory (GST) suggests that longer-term positive effects of setting task goals on learning are also

probable (Locke & Latham, 2002). It states that the presence of a goal not only determines the immediate reaction and behavior within tasks (such as higher levels of motivation and engagement), but by having students focus on performance goals above their current level across time, they will see discrepancies in their ability and continue to be motivated and engaged to try and improve. Thus, teachers need to nurture such goals within tasks when possible to keep students aiming for better performance from task to task.

4.2.2 Interpersonal and intrapersonal task goals

Achievement Goal Theory (AGT) helps explain why students may become motivated and engaged within classwork because of goals related to their performance (Covington, 2000; Elliot, 1999). AGT divides goals into 'performance' goals and 'mastery' goals. *Performance goals* are focused on the comparison of scores between different student performances within the same task and encouraging students to outperform each other. They encourage students to see success as *interpersonal*, by doing better than others, rather than improving their own performance over time. Such goals have been found to result in more effort to reach a higher level of achievement in learning (Elliot & McGregor, 2001). However, competition between students does not always result in a positive effect on engagement. When directly comparing individual task performances, some studies have shown students to avoid taking part in tasks and disrupt them (Ryan & Patrick, 2001) and avoid asking for help (Ryan & Pintrich, 1997). Also, such goals are also believed to increase anxiety levels among different students, as they feel under more pressure to perform in front of others (Covington, 2000). Having to aim for goals, and let others around you see you doing so, may have a negative impact on learning, as it can create 'feelings of

tension, anxiety or frustration' (Latham, Seijts & Slocum, 2016, p.3). Being under such pressure to perform has the potential to make students focus too much on strategies to complete the task, and save face, rather than focus on the actual learning intended (Seijts & Latham, 2001).

Therefore, it is important for teachers to consider these potential learning problems with comparative goal-setting among students. Some may enjoy focusing on performing better than others, but goals which are set privately and focused more on improvements in individual performance over time may sometimes be more appropriate. *Mastery goals* within AGT are those which focus students more on getting better at the necessary skills for improving their overall performance from one task to the next. They are believed to encourage students to view success as doing better than they did the last time they attempted a task, rather than competing with others around them. This longer-term, more *intrapersonal* view of success has been shown to lead to higher achievement (Bong, 2009), increased efforts to undertake classwork (Miller, Greene, Montalvo, Ravindran & Nichols, 1996), more attempts to ask for help when it is needed (Ryan & Pintrich, 1997), and higher levels of cognitively engagement (Wolters, 2004). Unlike with performance goals within AGT, no negative effects on learning have been found in research for mastery goals. A longer-term intrapersonal approach to goals is also suggested as important for student engagement in learning by *Self-Determination Theory* (SDT) which states that all human beings possess the inherent desire and need for personal growth and wish to have a feeling of competence and autonomy to do so (Deci & Ryan, 1985; Reeve, 2012; Reeve, Deci & Ryan, 2004; Ryan & Deci, 2000). Thus, students who feel that they are getting better at tasks across time, by reaching goals related to mastering the skills necessary to do so, might be expected to remain positive about the learning, as well as motivated and engaged. This is

examined within this study by seeing whether the use of performance goals for discussions results in improved performance across time (RQ2), as well as student reporting of positive feelings, improved motivation or engagement as a result of using those goals (RQ3).

To conclude, what is needed within language learning classes are intrapersonal task goals which can meet the individual motivational needs of students to promote positive affect towards tasks, as well as higher levels of motivation and engagement in their learning. According to the research discussed above, this may require the encouragement of continuous self-setting of increasingly difficult, but attainable, goals focused on language use skills by individual students over time which lead to improved task performance (addressed by the study, as discussed above). Additionally, competition between student scores, through interpersonal goals, should be allowed within classes (as knowing the scores of others or directly comparing scores can improve performance for some students), but not encouraged by the teacher (and is not within the study), as it may lead to anxiety, and/or lower levels of motivation and engagement among some students.

4.3 Formative Assessment (FA)

An important factor for helping students improve their performance within tasks is the feedback which they receive on that performance. Feedback which is provided at the end of courses, with the purpose of assessing the ability of students to use a language, is called *Summative Assessment* (SA) and is usually focused around what is called an 'exam culture' (Butler, 2011; Hamp-Lyons, 2007). This may satisfy testing needs for educational institutes, as well as providing students with an understanding of their ability to take tests, but performance

feedback which is provided continuously across time, and focused on supporting the learning of students, is more likely to help them significantly improve their performance. This type of feedback has been referred to as *Formative Assessment* (Black & Wiliam, 2009; Sadler, 1998; Tunstall & Gipps, 1996; Wiliam 2018), *Assessment for Learning* (Dann, 2002; Harlen & James, 1997), *Teacher-Based Assessment* (Davison & Leung, 2009), and *Diagnostic Assessment* (Huhta, 2008). All these approaches to learning agree that in order for students to improve at tasks across time, they require direct and continuous feedback on their performance. The term *Formative Assessment* (FA) is used from this point on to represent the collective research and beliefs of all of these terms to avoid confusion.

However, within language learning courses, FA is often not utilized as a learning tool and student are simply categorized at the end of course using Summative Assessment (SA), usually via tests (Leung, 2005b; Weir, 2005, p. 39). This type of feedback alone may not give students the support they require to focus on improving their language skills across time. Teachers need to ensure that what they are teaching across time within a course matches up well with what they are assessing at the end of it, so that students are developing the skills intended (Anderson & Krathwohl, 2001). The use of FA is believed to support such learning in three main ways. Firstly, it explains more about the current performance of students at different stages across a course (Atkin, Black, & Coffey, 2001, p. 26). By knowing this, teachers can give more immediate and constructive feedback across time, rather than having to wait until the end of a course to be able to do so (Davison & Leung, 2009). This makes students more aware of their abilities to perform tasks across time, and gives them the chance to focus their efforts and learning appropriately to improve (Stiggins, Arter, Chappuis, & Chappuis, 2004, p. 44). Secondly, FA also shows students

the expected levels they should reach at different times in comparison to their current levels. This may be a comparison of final test scoring systems for courses with the current feedback on performance within similar tests. This comparison helps students see and aim for the performance standards expected of them every time they receive feedback on performance across a course (Assessment Reform Group, 1999, p. 7). Lastly, FA helps students close the gap between their current performance and that expected of them by helping them to understand their own strengths and weaknesses within test performance (Assessment Reform Group, 1999, p. 7; Chappuis, 2005). The study in this thesis examines how relevant these three benefits may be for TBLT group discussion tasks (discussed as potential problems with TBLT in Section 2.4.3) by analyzing changes in performance across time with the use of FA (RQ2), and by connecting those changes to student reports about how FA was useful or not for improving their performance (RQ3).

4.4 Task performance scoring rubrics

To help teachers integrate FA into classes, what is needed is a specific guideline for language task performance feedback across time. This will not only help motivate students to undertake tasks, but also specify what 'good performance' within tasks is perceived to be within that course (Harlen & James, 1997, p. 377). The use of scoring rubrics within feedback is expected to help with this for several reasons. Firstly, scoring rubrics can grade specific types of performance within a task and highlight what areas students are expected to focus on (Andrade and Du, 2005; Black & Wiliam, 1998a; Reynolds- Keefer, 2010; Schamber & Mahoney, 2006). For instance, by assigning a higher score for fluency than accuracy within an oral task rubric,

students will understand the greater need to focus on speaking fluently, rather than worrying about errors in speech, to get a good score. Secondly, students can reflect on their current performance with specific scores provided by a rubric and are able to plan specific areas to improve within their performance for next time (Andrade and Du, 2005; Panadero, Alonso-Tapia & Huertas, 2012). For example, if a student has a much lower score for fluency than accuracy, they would be expected to realize the importance of focusing more on improving their fluency for their overall score. Thirdly, the improvements in clarity of performance and expected focus in learning provided by rubrics should lead to reduced levels of anxiety and stress (Kuhl, 2000; Panadero, 2011; Panadero & Jönsson, 2013; Wolters, 2003), as well as improved confidence to succeed within tasks (Andrade, Wang, Du & Akawi, 2009; Panadero et al., 2012), which are of course desirable for students.

Despite the benefits discussed, scoring rubrics for oral task performance still bring several other considerations. Firstly, most of the research into the effects of scoring rubrics are not related to orally interactive language tasks, but more to written tasks or oral monologues (such as presentations). Secondly, performance rubrics often use *subjective* scoring systems (Section 4.5.2.1), such as the opinion of a teacher about how 'fluent' a student is on a scale of one to five, and may not make it clear how a student can focus on improving. Thirdly, research on the effects of scoring rubrics have been mainly focused on test scores and student feelings towards learning, rather than how they can affect student behavior and interactions within tasks. Fourthly, teachers need to ensure that students do not receive too much feedback at once from a rubric, which will confuse them (Wiliam, 2018, p. 131), and that they have enough time to understand and apply it to improve future efforts to make the rubric effective (Black, Harrison,

Lee, Marshall & Wiliam, 2002, p. 9). Finally, the effects of scoring rubrics on different levels of performers within the same classes need to be considered. Lower performers within tasks have been found to benefit more than higher performers (Balan, 2012, p. 130; Black & Wiliam, 1998b), although this has not been proven in a language learning context. This is important for teachers to think about, as some rubrics may not be as beneficial for some (perhaps lower-level students) within classes.

A classroom study which I undertook (Stroud, 2017) addressed many of the above problems by investigating the effects of self-regulated performance scoring and tracking across five classes on student engagement within language use during classroom pair discussions. Students used a card game to score and track how often they gave opinions and reasons on discussion topics with their partner, receiving specific points for each action which they took (one point for one opinion or a reason for example). Students who used the cards and tracked their own scores over time with the rubric significantly increased their number of spoken words, opinions and reasons in discussions more than those who did not. The study demonstrated the ability of a scoring rubric to encourage students to orally participate more within discussions. However, it did not consider performance goals other than giving opinions and reasons, effects of the rubric on other task performance measures (such as those discussed in Chapter Three), student feelings towards the use of the rubric, differences in effects for low and high performers, or the potential longer-term effects on learning (such as across a typical 15-week semester). These are all important considerations to better understand the effects of task performance scoring rubrics on learning within typical language learning courses and are analyzed in this thesis.

4.5 Group discussion GSF design

This next section explores support in the research literature for combining goal-setting, FA and task performance rubrics, to potentially create an autonomous system for students to use to support their second language learning called *Goal-Setting and Feedback* (GSF). With this system, students set specific performance goals before each discussion, based on a scoring rubric, and reflect on feedback for that performance after the discussion is completed. The process is performed autonomously by students in a cycle across time, with new performance scores driving the goals of the next discussion, depending on which scores within the rubric students wish to improve at. This type of *Self-Regulated Learning* (SRL), as it has been known, ensures that students autonomously monitor and guide their own learning towards goals (Sitzmann & Ely, 2011: Zimmerman & Campillo, 2003; Zimmerman & Moylan, 2009; Zimmerman & Schunk, 2011). This kind of GSF cycle is an original approach proposed within this thesis to supporting second language learning with oral group discussions and assumes that such a system would incorporate all of the benefits discussed above for goal-setting, FA, task performance scoring rubrics and SRL. However, the use of GSF would require additional time and effort by students within communication classes to ensure that it can be effective for learning (Black et al., 2002, p. 9). Also, the focus of the goals and how performance is scored will determine how the GSF may affect the learning. These are important points which were considered for the GSF used within the study, to make it as supportive as possible for the learning, and is now discussed.

4.5.1 GSF focus

A good starting point for integrating GSF into discussions is to decide the focus of the goals. Careful consideration is needed for the 'type of information' that assessment and feedback on task performance provides to students, so that it can be used in a way to improve the learning as intended (Salvisberg, 2011, p. 15). As the goals selected will influence the efforts of students to improve at discussions, it is important to choose those which are aligned with what is defined as performance within assessment at the end of a course. Furthermore, the goals and feedback provided need to cover a variety of measures, as a higher level of detail about performance will help students understand it more, rather than a few measures showing them a very narrow representation of it (Fengying, 2003). However, teachers also need to be selective about the total number of feedback measures and detail that GSF requires, so that it can be easily interpreted within the time allowed for feedback, without confusing students (Wiliam, 2018, p. 131). This careful balance of focus for GSF for group discussions is now discussed with regards to focusing on individual process and/or group product goals.

4.5.1.1 Individual Process GSF

One way to focus GSF for group discussions is by considering the language use of each individual student during discussions. This is referred to as *Process GSF* within this thesis, which is used across the semester by one of the two groups of students. The use of individual participation and CAF measures (such as those discussion in Section 3.2) for each student within discussions is perhaps the most obvious choice for individual GSF. Such measures match up with current individual student oral testing (such as those used in TOEFL, TOEIC and IELTS

speaking tests) and teachers and students may already be familiar with and have a general understanding of terms such as complexity, accuracy and fluency. However, there are many possible choices for GSF focus for individual speech and students may have a limited capacity for what they can focus on (Section 2.2.2). For instance, Wang (2014) found that by focusing on accuracy within feedback for oral tasks, Chinese students became very nervous about speaking, as they were mainly trying to avoid making errors in speech (hence, possibly 'trading off' participation, fluency and complexity measures to focus more on accuracy). GSF focused on one area may result in less attention for others, such as perhaps less participation and fluency when accuracy and complexity are assessed in feedback, and vice versa. Also, 'risk-taking', without fear of losing face in front of others or losing scores because of CAF-related weaknesses in speech, is an important part of learning to speak in a second language (Brown, 2007, p. 149). Thus, too much focus on linguistic CAF measures within GSF may actually prohibit such risk-taking if it makes students too anxious to perform.

Process GSF for discussions could also focus on more *interactional* measures for students. These would be more representative of interactional competence (Section 3.3.1) and discourse competence (Section 3.3.2) and help focus students more on improving their interactions with others, and not only the CAF of their individual turns. This could include a combination of associated measures, discussed in Chapter Three, such as turn-taking skills, asking and answering questions, paraphrasing, agreeing and disagreeing, back-channeling, asking for and giving each other help with the language, and promoting a group outcome with others. However, which of these would be most 'appropriate' to set goals for and use feedback to improve across time is a difficult question, as no research exists to show the long-term effects of

such individual process-focused GSF for group discussions (RQ1 addresses this by using teacher opinions about appropriate performance measures). In addition, how many and what combination of these measures are important to support appropriate learning is unknown. Too many may lead to trade-off between performance measures (as mentioned above for CAF), as well as too much workload for students, in addition to them trying to listen to, comprehend, and react to the turns of others during a discussion (analyzed within RQ2 and RQ3).

In summary, using a cycle of setting individual performance goals and analyzing feedback on performance within a discussion may be expected to lead to better clarity and focus in discussions, as well as higher levels of motivation, engagement and achievement in learning (Sections 4.2, 4.3 and 4.4). However, making students accountable for their own performance, especially by performing in front of others, was also discussed as a possible cause of anxiety, demotivation and disengagement with individual GSF (Section 4.2.2) (analyzed using student reported feelings in RQ3). These effects are analyzed using data from one group of students who undertook this type of GSF within the study (ProcS). Another approach, adopted by the other half of students in the study, is undertaking GSF as groups (ProdS), which is now discussed.

4.5.1.2 Group Product GSF

A different approach to GSF for discussions is considering the shared outcome of groups as task performance, called *product* goals, which were used by one of the two group of students in the study. This involves groups focusing on collective goals for the outcome of a discussion, rather than the actual language use and interactions during the discussion to achieve this (i.e. Process GSF as discussed in the previous section). The two important focuses of *product* goals

for groups are the 'content' (the focus of students on a group result for the task, rather than individual performance) and 'specificity' (what exactly that group result should be) of the goals (Nahrgang, DeRue, Hollenbeck, Spitzmuller, Jundt, & Ilgen, 2013, p. 13). For group discussions, this could mean reaching an agreement on the discussion topic (the 'content') and having a list of specific points to present to a teacher post-discussion (the 'specificity').

One way in which *Product GSF* may be more beneficial for performance and learning than Process GSF is that groups working together on the same goal can support each other by distributing attention, searching for information separately, and by putting together knowledge by exchanging ideas (Latham & Locke, 2007). The effects of a group product goal on performance for oral tasks has been tested, such as with the group goal of a list of 50 uses for a knife discussed earlier for the study by Martin et al. (2016) (Section 4.2.1). The use of the goal was shown to result in a better output by groups, in the form of longer lists than a 'do your best' condition. This collaboration between students may also lead to better learning (and perhaps task performance) for individuals within groups, especially for lower-level students who feel stressed when performing in front of others with Process GSF (Section 4.5.1.1). According to the learning theories behind TBLT discussions, focusing on the outcome of a discussion as a group may be enough to promote learning and improvements in performance, without the need for direct feedback on interactions during task time (Sections 2.4.1 and 2.4.2). However, studies such as Martin et al. (2016) have not analyzed improvements made in individual second language use as a result of group goals.

On the other hand, the additional collaborative and cooperative communication which product goals require between group members, compared to process goals, may also cause

problems for learning. Shared goals require students to agree on an outcome and the differing motivations and abilities of group members to work collaboratively and cooperatively to reach that goal may affect that task outcome (Chen & Kanfer, 2006; Latham & Locke, 2007). This needs to be considered on an individual basis for students working in groups. Although students with higher-efficacy than others have reported preferring *Convergent Assessment* for task performance (based on agreed group outcomes), *Divergent Assessment* (based more on individual performance than combined group performance) has been reported to be more motivating and preferred by lower performers within groups (Huang, 2011). The present study discusses the differences in performance changes over time for lower and higher-level group members (categorized as LPs/HPs) with both product and process goals, to help better understand how they may support the learning of students at different levels (Section 6.3.6.3).

In summary, focuses on individual Process GSF (such as participation, CAF and interactions measures) and group Product GSF (such as agreed outcomes between group members) for group discussion tasks are not understood at this time for their effects on student performance and learning. Deciding a balance between these focuses to support learning is challenging and requires further investigation for classroom tasks (Long, 2014, pp. 332-334). An important question addressed and analyzed using the data from two different groups of students in the study (ProdS and ProcS) is *to what extent can group product-focused and individual process-focused GSF lead to improvements in individual speech (CAF), interactions between speakers and the ability of groups to reach an outcome together?*

4.5.2 Performance self-scoring method

In addition to deciding what focus GSF should have for group discussions, it is also important to decide how students should score themselves using the chosen performance measures, so that the GSF cycle is clear and effective for helping them improve over time, without the need for continuous teacher-feedback. This section discusses two possible approaches to this; *rating scales* and *counting systems.*

4.5.2.1 Performance rating scales

The most common method for rating oral task performance in language learning courses is with *rating scales*, such as with those used in the TOEIC, TOEFL and IELTS speaking tests. In these tests, student performances such as topic development, language use (especially vocabulary, grammar and accuracy), and delivery (such as fluency, intonation and pronunciation) are usually scored with numbers on scales ranging from 0 to 4 for the TOEFL test, up to 0 to 9 for the IELTS test for example. The total number awarded to students on these scales represents their current performance and can be used as feedback to assist learning. At present, there is no standardized assessment guideline for group discussions within Japanese universities. Therefore, within classrooms, TBLT group discussion performance goals, assessment and feedback are often determined by the teacher using their own choice of rating scales. Adopting existing performance rating scales, such as those used within the tests mentioned above, could be beneficial for group discussion GSF, as it would keep the focus of learning on improving performances which match up to international speaking tests.

However, there are several potential drawbacks to using rating scales for self-scoring of performance within group discussions. Firstly, because rating scales are vague and subjective forms of assessment, they make it difficult for students to know how to set clear language performance goals which can then be easily interpreted within feedback and applied to improve future attempts (Orsmond et al., 1997; Price et al., 2010). For instance, if a student scores themselves with a fluency rating of three, and sets a goal of four for the next discussion, it is very difficult for them to know how they should focus their efforts to improve that score, or if they have attained that goal, other than by making a subjective decision about their own performance. Secondly, raters of oral task performance can vary greatly in the scoring they award (Bachman, 2000; Bachman & Palmer, 1996; Leung, 1999; Leung & Lewkowicz, 2006), making it impossible to create a standardized scoring system for classes which can help categorize student levels. Such variation in scoring may also create a sense of unfairness and demotivation among classmates, as some students may award themselves higher scores than others for similar performances. Thirdly, the use of rating scales requires a high level of comprehension of the language, as well as experience or training for assessing performance through scoring measures such as CAF on a scale (Bonk & Ockey, 2003, Sook, 2003). Learning how to apply such scoring to oral tasks can take a long time for students and they may require time for training and practice if it is to be effective (Carless, 2009a). Some language courses may not have time for such training and therefore rating scale scoring may not be an effective approach for GSF. Lastly, and perhaps most importantly, students are unlikely to be able to rate their own performance with rating scales, as they may not be aware of their own problems and errors in language use. It is not practical to have students score themselves with linguistic measures (CAF) if they do not

know how to improve them by themselves. An alternative approach adopted within the study is now discussed.

4.5.2.2 Performance counting systems

Following on from the above discussion, an important question is *what type of performance scoring can students use themselves for GSF in discussions which is quick to learn, clear in terms of scoring, feedback and future goals, and can avoid inter-rater issues by being standardized across a class?* A new approach proposed and analyzed within this thesis is to count specific performance-related actions by individuals during the discussion time (within a process approach), or to count final decisions made as a task outcome (within a product approach). What this means is that students count how often they take performance-related actions within discussions (those discussed in Chapter Three). With a process approach, this could include how many turns, opinions, reasons, examples, agreements, disagreements, and questions they ask. With a product approach, it could include how many things groups can agree together before the discussion is finished, such as a list of their possible topic choices, reasons, and examples. This approach to scoring may make performance feedback and goals more specific and measurable compared to subjective rating scales, which should result in higher levels of motivation and engagement (Moskowitz & Grant, 2009). In addition, much less training and time for analyzing performance and applying feedback would be expected with this simple counting system compared to rating scales. For instance, it would be easier for students to know how many questions they asked (process), or how many reasons they agreed on for their group decision (product), rather than rating how complex, accurate or fluent their own language use

was on a scale. If students can assess their own performance during the discussion time using simple counting, it would reduce their work load and give them more time to focus on practising the actual discussion.

However, some potential problems exist with this approach. Firstly, this form of scoring is not easily connected to formal oral tests, such as TOEIC, TOEFL and IELTS. As discussed in the previous section, such tests use rating scales of measures such as CAF for performance. By using a counting system for GSF, students (and teachers) may struggle to see the connection between the learning undertaken and testing of speaking skills which may be important in the future. This may create resistance for its usage and even result in demotivation. Secondly, scoring through counting actions and results for discussions can be considered a very 'mechanistic' approach to assessing performance which overshadows important opinions about overall performance (Davison, 2004, p. 324). Students may be able to set and monitor goals and performance within discussions based on counting actions/output, but these numbers do not specify important details about the quality of language use, interactions or decisions about the task which they make as individuals or as a group. More specific elements of performance, such as details about CAF of language use, are clearly more difficult using counting, as they require a more subjective and qualitative analysis of performance, rather than a simple quantitative-style counting one. This shows that *counting systems* for performance are limited to actions within turns which can be quickly counted, but give little information about the ‘deeper quality’ of language being used (such as topic development, language use and delivery possible with rating scales).

In summary, *performance counting systems* can make group discussion GSF more easily self-regulated, standardized, speedy, simple and easier to use for students compared to *rating scales,* and was therefore adopted within the study. However, counting systems are clearly less connected to formal oral testing rubrics and are a more mechanical approach which reveal less detail within GSF about specific oral performance measures (such as CAF). No research currently exists on the effects of such scoring on learning and performance over time. Therefore, an important question for teachers (addressed by the study) is *can student GSF, based on counting task actions/outcomes, significantly improve discussion performance and learning over time?*

4.6 Chapter summary

This chapter discussed the potential use of group discussion performance-focused GSF to support learning for students within language learning classrooms, as explored within the study. It reviewed literature on how goal-setting, Formative Assessment (FA) and task performance scoring rubrics could be used over time to motivate students and how this can be expected to result in higher levels of task engagement and achievement in learning. The main problems of choosing appropriate focuses for measuring performance within GSF was also discussed. For group discussions, it may be that the most appropriate and realistic focus for GSF is the self-counting of individual spoken actions taken during discussions (such as stating opinions, asking questions, and agreeing or disagreeing) and/or group outcomes for the discussion (such as a list of ideas and reasons), and was therefore the method chosen for the current study. Self-counting provides a quick, simple, self-regulated and standardized system for GSF, but one which is likely

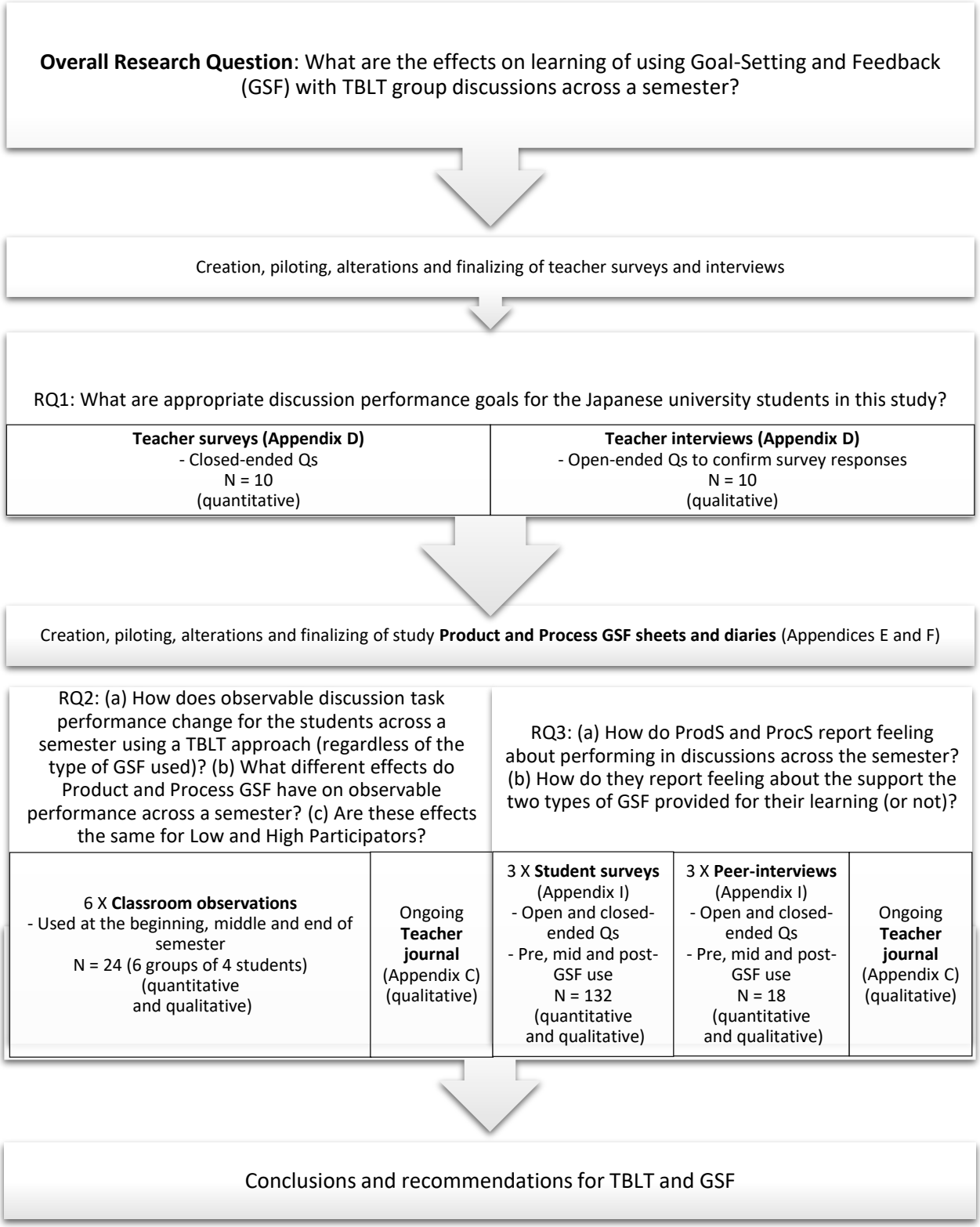
quite limited in its detail for feedback for learning to students, the implications of which are explored as part of this study. Firstly, however, the methodology of the study in this thesis is now explained to show how the effects of discussion counting systems with both Product and Process GSF were analyzed for Japanese university students across a semester.

CHAPTER 5. METHODOLOGY

5.1 Aims of the study

The main aims of the study were to (1) investigate how group discussion performance changes over time for first-year Japanese university students using a TBLT approach and (2) how classroom-based Goal-Setting and Feedback (GSF) might be used to support the learning. Figure 5.1 shows an overview of the procedure undertaken in the study to answer the research questions, which are addressed using data in the next chapter. I did not intend to find out whether one single approach to GSF would be better than another, but rather to investigate how two very different approaches to GSF may affect students' feelings towards and performance within group discussions across time. This was done by using task product orientated GSF (all goals and feedback relating to the task outcome) with three classes of students in the study and task process orientated GSF (all goals and feedback relating to the content of oral turns taken within the discussion time) with three other classes. To triangulate data regarding potential benefits of GSF, I used a mixed-methods approach to data collection and analysis within a longitudinal study. By doing so, I intended to produce useful data and recommendations for teachers and researchers on the combined use of TBLT and GSF to improve English group discussion skills across a semester-long communication course. This chapter explains the rationale for the data collection methods adopted, details of the study participants, the study procedure, the data collection and analysis, and the ethical considerations made for the participants.

Figure 5.1. Research summary



5.2 Rationale for the research methods

This section explains the choices made for collecting data within the study in order to address the research questions (Section 1.4). It explains the decision to use a mixed-method approach, clarification of how this was done, as well as details of each method used.

5.2.1 Mixed-method approach

Mixed Methods Research (MMR) is often used as a very generic term which requires further clarification in terms of the reasons and procedures for triangulating quantitative and qualitative data (Angouri, 2010; Creswell, 2009; Greene, 2008; Tashakkori & Teddlie, 2003). For this thesis, three different approaches to collecting mixed-method data were used, called *data triangulation, investigator triangulation* and *methodology triangulation* (Denzin, 1970). By using data from different sources of information, different investigators and via different methods, findings for the study could be stated with more confidence in order to answer the RQs.

Firstly, *data triangulation* and *methodological triangulation* (Denzin, 1970) was undertaken through the collection of qualitative and quantitative data from surveys, observations and interviews in order to address all three RQs. This was done because the use of observational and self-reported data can *complement* and *expand* on each other to produce a clearer overall picture of not only what participants are doing, but perhaps why they are doing it (Creswell & Plano Clark, 2011; Dörnyei, 2007; Greene, Caracelli & Graham, 1989; Ivankova & Greer; 2015). Teachers who can see how students are performing in discussions also need to understand more about student feelings if positive alterations are to be made in the future to improve the learning.

Secondly, data from teachers within the department was used to guide the *development* (Greene et al., 1989) of GSF undertaken within the study (RQ1) and student data was used to see how the GSF used might aid learning or not (RQ2). By doing so, recommendations with regards to the use of GSF, which would be relevant for both teachers and students, could be provided.

Thirdly, data was collected and analyzed by three different researchers (two student interviewers and myself). This *investigator triangulation* (Denzin, 1970), was used to gain a wider understanding of student perspectives of the GSF used (RQ3), rather than rely on only what students would report only to their teacher, or a single researcher within a study.

The following sections explore the different elements of the MMR in more depth.

5.2.2 Use of classroom observations

Observations of students undertaking discussions were used as the primary source or data within this thesis to identify the effects of TBLT/GSF on learning over time (RQ2). This was done because the most valuable assessment of the effects of TBLT/GSF on discussion performance for university teachers are those which they can see themselves and assess within their own courses. By better understanding changes in observational performance due to TBLT/GSF, teachers could apply such data to their own courses to improve the discussion feedback and assessment they provide and see changes in order to see similar performance changes among their own students.

Also, by using observations of the students during class and tests, the data could show what they actually did, rather than relying on only what they said they did, which may not be the same thing (Dörnyei, 2007, p. 185). Thus, the observational data used in the study was very

important, as it could help prevent reliability issues commonly associated with the use of only elicited data (see the next section).

While observing student discussions during the piloting of GSF and across the semester, I took notes within an informal journal with regards to any points which I considered to be important, but perhaps not captured by the other research methods I used (Appendix C). I used these notes to add to the discussion within this thesis where I felt that the observations and related opinions of the teacher of the course would be relevant. These points referred mainly to student actions during discussions, negative actions or comments by students regarding the GSF, the use of GSF within class from a teaching perspective, and additional differences between learners which I observed. I did not rely on the observations in the journal notes as much for drawing conclusions as my other research methods, but felt that it was an important addition to the data which incorporated the perspective of the actual teacher of the courses (myself).

5.2.3 Use of surveys

Closed-ended surveys were used with the teachers to determine their overall ratings of different variables to use for GSF with the students (RQ1). With the students, open- and closed-ended survey data was used to *complement* and *expand on* (Greene et al., 1989) the observational data discussed above. By using self-reported data to assess changes in feelings towards the learning and GSF (RQ3), possible reasons for changes in discussion performance which were observed across the semester could be established. This triangulated data was used to form the recommendations for the use of TBLT/GSF for teachers within the conclusion section (Chapter Seven).

The surveys were a time-efficient way to get direct feedback from 10 teachers before the semester began, as well as from 132 students at three different times across the semester. However, survey data is often criticized for what Dörnyei (2003, pp. 10-14) lists as reliability and validity issues. The first challenge with this type of data is the potential for *superficiality* and *unmotivated* responses due to the *fatigue effect*. This was less of an issue for the teachers (who only completed one open-ended survey), but by having students complete three similar open- and closed-ended surveys, it is possible that their motivation to explain detailed thoughts about their learning diminished towards the end of each survey and from one survey to the next. To counteract this, the fewest number of survey items possible (to collect adequate data to address the RQs) were used, and students were provided with surveys and interviews in their first language (Japanese). This was done as it made responses easier to give and surveys seem less of a chore each time they were done (compared to doing them in English).

Another challenge with the data is *social desirability bias and self-deception* (Dörnyei, 2003, pp. 12-13). It is possible that participants may have responded in the way which they believed I, or anyone else viewing their responses, may expect a student or teacher to do. For the teacher surveys, this was perhaps the biggest threat to the reliability of the data, as no observational data was collected to confirm any of their self-reported data with their actual assessment practices. Items were also worded as neutrally as possible in both student and teacher surveys to counter any bias in responses (without any leading questions), and a verbal explanation was given to all participants (and in written form for the teachers) that there were no 'right' answers.

5.2.4 Use of peer-interviews

Interviews were used to collect data from both students (RQ3) and teachers (RQ1), but only to clarify survey responses for the teachers. As with the survey data discussed above, this was done to build upon the observational data to find reasons for the observed effects of the GSF on learning, as well as any effects not captured by the observations undertaken.

All of the interviewers were performed by what I refer to as 'peers' in this thesis, although student interviewers were one year older than the student interviewees and whom I trained to do the interviews myself (as explained in Section 5.4.1). I decided to user peer-interviews because the relationship between the interviewer and interviewee can often have a significant effect on the empathy, interactions and co-construction of the interview content and the responses given (Briggs, 1986; Mann, 2011; Watson 2009). By having myself interview my colleagues, and students interview other students, I assumed that interviewees would be more comfortable sharing their personal opinions in these situations, as the person they are speaking to is in a similar learning or teaching situation to themselves. As a result, reliability issues related to this relationship, called the *interviewer effect* (Denscombe, 2014, p. 189) could be reduced.

In addition, interviewers sat next to the interviewees, and added an informal introduction section to the interviews to try and build rapport and create a relaxed environment before starting the interviews (as suggested by Denscombe, 2014, p. 194). It was assumed that this would also increase the chances of respondents reacting more informally and giving more open and honest responses.

The use of peer-interviews to collect more reliable data from the participants was also based on the beliefs of Community Research, such as those outlined by Goodson and Phillimore

(2010, 2012). In their study, researchers with refugee backgrounds were trained to gather data from within refugee communities about issues around mental health, education and employment. By using people from within a community to gather feedback from that community, it is believed that the data collected can be more reliable and revealing about genuine feelings than when someone outside of the community collects the data without being able to empathize with the community (as advised by Goodson & Phillimore, 2010, 2012). For the student interviews, I assumed myself to be the outsider (a non-Japanese, older person, who was the teacher of the courses) and student interviewers as members of the 'same community' as the interviewees (both Japanese, of similar ages and studying at the same university). Although my research was short-term, and involved less training of researchers and was less ethnographic in nature than Goodson and Phillimore's approach, I hoped to use it to gather fruitful data from within, rather than from outside the student 'community' in a similar way.

5.3 Participants

5.3.1 Teachers

The participants consisted of ten teachers (eight male and two female), excluding myself as the eleventh, teaching English communication courses to first-year students at a satellite campus of Kwansei Gakuin University. They were all either American, British or Australian (with English as their first language) and had between three and ten years of university teaching experience in Japan. The teachers and I had all been working together within the same department for at least one year and had good rapport and comfortable working relationships.

5.3.2 Students

132 first-year students (89 male and 43 female) from the six English communication classes which I was teaching at Kwansei Gakuin University in 2015 agreed to take part in the study. These students had all entered the university after passing an entrance exam and were all majoring in Science and Technology related fields. They were all studying on mandatory English reading, writing and communication courses, with different teachers teaching the three different courses to them. At the start of the study, the students generally reported low levels of self-perceived overall ability to use English in general and to use English within discussions, low levels of confidence and high levels of anxiety when speaking in English (Table 6.12).

Their overall proficiency levels were difficult to assess, as 61 of the students had either not yet taken an English proficiency test (such as TOEIC or TOEFL) or did not wish to disclose that information. However, 71 students submitted a recent TOEIC score at the start of the study, with an average score of 403 (Elementary Plus level). From my own teaching observations and direct interactions with the students across the first semester of the year, the majority of them found discussing topics in English very difficult and could not maintain a conversation for more than one or two simple exchanges. This was typical of classes which I taught and I felt that the students in the study were a good representation of the common English oral communication ability of first year students entering the university.

5.4 Research procedure

5.4.1 Classroom-based study preparation

Table 5.1 below shows the preparation undertaken before the start of the classroom-based semester-long study to address RQ1 (Section 5.5). Firstly, a pilot for the teacher survey and interviews was run with one part-time English teacher from within the same department. The teacher had several questions about the exact purpose of the study and clarification of the meaning of 'importance' for the factors. The teacher did not feel these points were clear in the survey and so I decided to add a face-to-face verbal, as well as a brief written introduction to the survey, to better explain them (Appendix D).

Table 5.1 Study preparation

May 2015	- Teacher survey and interview piloted, adjusted (verbal and written context clarification added) and finalized (Appendix D)
June 2015	- Teacher surveys completed (within one week) - Teacher survey-clarification interviews completed and reviewed (within one week) - Survey/Interview data reviewed and follow-up teacher interviews held to clarify uncertain responses
July 2015	- Interview data reviewed, analyzed and final discussion goal list created - GSF sheets and diaries created (using finalized goal list)
August 2015	- GSF sheets and diaries piloted, adjusted and finalized (Appendices E and F) - Interviewer schedules agreed and interviewers trained using a summer course class

The ten teachers completed their surveys and interviews within one week. I used an empty classroom on the campus for the interviews, which generally lasted between 25 and 30 minutes each, but did not have a specific time limit. I also conducted further follow-up interviews with the teachers a week later (about five-ten minutes in length each) to clarify responses and allow them to change any responses.

The survey data was then analyzed (Section 5.5) and used to create a list of countable goals for incorporation into the first version of the Process GSF sheet and diary. This was piloted across five different classes (within a two-week period) with 28 first-year students undertaking an English intensive course with me during the summer break at the campus. Small adjustments were made to the sheet and diary following this (Section 6.2.3) and final Product and Process GSF versions created for use in the study (Appendices E and F).

In preparation for the peer-interview data collection, the two student interviewers were informed of the study purposes, the learning which the students were going to undertake, and were given a chance to use the sheets and diaries themselves to better understand their function. I also used the English intensive course discussed above to train the interviewers with three different volunteer students each. I gave them training on non-persuasive interviewing skills (such as not using leading questions and how to elicit with questions such as 'can you tell me more about that?'), rapport-building body language (sitting next to the interviewee rather than across from them for example) and note-taking skills (not writing things in full but in short-hand to be able to focus on the interview itself rather than writing). I gave them feedback, after observing their interviews myself, in terms of interviewing skills, note-taking and translating of the recorded interview afterwards. After three rounds of feedback with the interviewers I felt that they had learnt a great deal about undertaking interviews and that I had answered all of their questions about how to interview in accordance with the goals of the study.

5.4.2 Semester-long classroom-based study

Table 5.2 shows the details of the 15-week classroom-based study undertaken in the second semester of 2015 to address RQ2 (Section 5.6) and RQ3 (Section 5.7).

Table 5.2 Study procedure

Week 1 (Sept 28-Oct 2, 2015)	- Course content introduction - Assessment criteria explanation (30% total for discussion tests, 50% for attendance and participation, and 20% for TOEIC listening test) - Recording equipment introduction - Ethical consent form explanation and signing (Section 5.8.1 and Appendices G and H) - Group formation (student choice) and discussion practices
Week 2 (Oct 5-9)	- Group finalizing (student choice) - 1st recorded classroom group discussion - Students volunteered for interviews - Initial recording analysis to determine Low (LPs) and High Participators (HPs) to be interviewed (Section 5.4.2)
Week 3 (Oct 12-16)	- 1st discussion test - 1st student attitudinal survey (open and closed-ended) (Appendix I) (N=132) - 1st peer-interviews (using survey questions in Appendix I) (N=18)
Weeks 4-6 (Oct 19-Nov 13) (no classes held Nov 2-6)	- 1st meeting with student interviewers (Week 4) - Introduction for students to GSF sheets and diaries (Appendices E and F) - Setting up of student diary passwords and access - Weekly discussion practices with GSF
Week 7 (Nov 16-20)	- 2nd recorded classroom group discussion
Week 8 (Nov 23-27)	- 2nd discussion test - 2nd student attitudinal survey (closed-ended only) (Appendix I) (N=132) - 2nd student peer-interviews (closed-ended only)
Weeks 9-11 (Nov 30-Dec 4)	- Weekly discussion practices with GSF
Week 12 (Dec 7-11)	- 3rd recorded classroom group discussion
Week 13 (Dec 14-18)	- 3rd discussion test - 3rd student attitudinal survey (open and closed-ended) (Appendix I) (N=132) - 3rd peer-interviews (N=18)
Week 14 (Dec 21-22)	- 2nd meeting with student interviewers

In the first week of classes, I explained the outline of the course and my study to the students in a verbal and written form in both English and Japanese (Appendix G). This included how they would practise discussions within groups each week and their assessment for the course. This scoring system was decided and agreed by the English teachers and the head of the department at a meeting before the start of the semester. The same scoring guideline and number of tests were used for every English communication course within the department.

The students were then introduced to the classroom recording system and given time to use them for a practice discussion. For the final 20 minutes of the first class, students were made aware of the study and completed the ethical consent form (Appendix H), and formed groups of their choosing for discussions.

In Week 2, students were given time to change groups if they wished to, so as to reduce any problems which group set-up may have for participation for students (such as those shown in the findings of Module One in Appendix A). Students then undertook a group discussion which I recorded and used for later analysis to determine which half of the students within each class spoke the least and which half the most during a discussion as an indication of oral proficiency (in a similar way to Wigglesworth, 1997). I then labelled these students as **'Low' (LPs) and 'High Participators' (HPs).** Table 5.3 shows how the six observed groups consisted of a total of 12 LPs and 12 HPs, with a matching distribution of LPs and HPs among the Product and Process groups.

Table 5.3 LP and HP group distribution

Product GSF		
Group 1	Group 2	Group 3
LP, LP, LP, HP	LP, LP, HP, HP	LP, HP, HP, HP
Process GSF		
Group 1	Group 2	Group 3
LP, LP, LP, HP	LP, LP, HP, HP	LP, HP, HP, HP

I decided to use the words which students spoke in the Week 2 discussion as a measure of proficiency (as I did in the second module of this PhD, shown in Appendix B), rather than proficiency test scores (such as TOEIC speaking tests), as such tests have no proven correlation to group discussion performance. Also, test score data was not available for all of the students at the start of the study. Judging student performance based on words spoken alone (and from only one discussion) was a limited and perhaps misleading view of proficiency, but the best choice which I felt I had with the time available to analyze the performance of the students between Weeks 2 and 3. In future, the use of several pre-study discussions to create a clearer picture of performance would be preferable (and perhaps additional measures for assessing proficiency) if time is available to do so.

At the end of the Week 2 classes, students who were willing to undertake interviews in Weeks 3 and 13 (in exchange for chocolate bars each time) submitted their names to me. I then used this list to choose between two and four interviewees from each class, which came to 18 students in total (seven from Product GSF classes and 11 from Process GSF classes). These uneven group sizes are addressed as a limitation within the RQ2 analysis in Section 6.4.8. Only one student from each group was chosen and never more than four volunteers from the same class (and so all volunteers for interviewers were accepted and decisions did not need to be made about how to choose them).

In Weeks 3, 8 and 13, groups undertook eight-minute **discussion tests**, one at a time, in the classroom while other groups from the same class waited within a different classroom. Another teacher was always present and monitoring the waiting students within that classroom, so that they could not communicate with groups who had completed the test to discuss the test topic prior to their tests. At the start of each test, groups were presented with the topic, but were not given any preparation time for the discussion. I used the same GSF sheet which groups had used for themselves during classroom practices to watch and score groups (ProdS with a shared group score and ProcS with individual scores). I then used those scores to award a test score based on my judgement of their performance compared to the rest of the class (to maintain a 'bell-curve' for all of the scores in each class, as I was instructed to do so by the university department). This lack of a precise connection between classroom and test scores might have caused motivation problems, as the students may have felt that the assessment of performance by the GSF should match the assessment in tests more closely (although I did not hear any students mentioned this during the semester). This test scoring system could be improved upon in future studies by creating a system which can quickly generate specific grades based on observed performance within tests (but this discussion goes beyond the scope of this thesis).

After the tests, students completed either a **survey** on a computer in a separate room, or had an **interview** with one of the two student interviewers, while other groups took their test. The interviewers transcribed and translated their open-ended interviews in English (in Weeks 3 and 13) and met with me to discuss them in the following week.

At the start of Week 4, three classes were introduced to the Product GSF and the other three to the Process GSF for use within **classroom discussions**. Classroom practices were carried

out at the same time by all groups within a class and, unlike tests, allowed for a rehearsal stage (see the next section). I randomly assigned the six classes to one of the two types of GSF and felt confident that I had divided them into two as equal as possible groups in terms of size, gender, ability and motivation (although I did have any data regarding group discussions to confirm these final two points, but only my intuition about the students at that time). Across Weeks 4-7 and 9-12, the classes then used one of the two GSF types (consisting of the sheets and diaries in either Appendix E or F) to support their learning during discussion practices. The GSF diaries were saved and password protected on my online google account, but each student had access to their own diary at any time by accessing the system and entering the individual passwords I had assigned them.

5.4.3 GSF class procedure (Weeks 4-7 and 9-12)

Table 5.4 shows the class procedure undertaken by the students when using GSF.

Table 5.4 Class procedure

5m	- Warm up activity (weekend-related questions in pairs)
5m	- Discussion topic introduction (Appendix J) and GSF diary scores/goals review
10m	- Pair discussion rehearsal (same weekly partner from a different group)
15m	- Discussion group formation - Goal-setting undertaken in GSF diary - Eight-minute group discussion (recorded) using GSF sheet
30m	- Discussion recording/sheet review with GSF diary note taking - Teacher feedback and peer-review of language goals (face-to-face and written on the system) - GSF diary scores/goals reflection and updating
20m	- TOEIC test listening practice (unrelated to the study)

For the first five minutes of class, students undertook a **warm up activity**, during which they would find a randomly assigned partner in the room and ask each other questions about their weekend in English. I walked around during this time and greeted the students while collecting their homework (related to TOEIC listening practices).

After that, students returned to their groups (the same each week) and were introduced to the discussion topic for that class. Students then opened their electronic GSF diaries on the computers on their desks and **reviewed goals and points** from the previous week together. During this time, I encouraged the students to try to improve their overall score as much as possible in that class by considering those points but with no specific guidelines how to do so.

Next, students sat with the same partner each week (someone outside of their discussion group chosen by the students themselves in Week 2) to undertake a **ten-minute rehearsal-style practice** of the discussion. This was done before each discussion following the conclusions from Module Two of this PhD. It was found that the addition of such a rehearsal stage improved student participation and fluency within a group discussion on the same topic afterwards (see Appendix B) and I decided to include it as a way of improving participation within discussions across the semester.

After the rehearsal stage, students returned to their groups and undertook a **group discussion** of the same topic while using their GSF sheet to make notes on content, but not language problems to encourage students to focus more on meaning than form within the discussions (as expected with a TBLT approach). I recorded and collected the discussions using microphones on desks, and sent copies of the recordings of their own discussions to the students' computers after each discussion using a classroom management system.

Once the discussions were finished, students used the audio recordings and **GSF sheets** to review the content of their discussions, make notes on the language problems encountered and update their scores and notes in their **GSF diaries** (all for the first time in Week 4 without prior training). During the 30 minutes of class time allowed to do this, students raised their hands and checked language points with me and practised saying the English sentence corrections I gave them with their group. I encouraged students to ask me for as much help as possible during this time and many of the students asked me questions about English grammar and vocabulary which they were unsure of. I also gave students the option for me to help correct their English notes by accessing their diary online if they did not wish to speak to me directly about it in class. Some students chose this option and it was a good way for me to give feedback to students indirectly or outside of class time if they did not wish to communicate with me during class.

Students were then given five minutes to discuss their scores for the discussion with their group and asked to set goals for the discussion next week (points for each section in the diaries shown in Appendices E and F). During this time, I encouraged students to try and perform better by getting a higher overall score within the next practice, but did not specify how to do that. Students then closed their diaries and were free to access them whenever they wished to outside of class. However, data from my google account showed that none of the students in the study accessed their diaries at any point outside of class time.

5.5 RQ1 data collection and analysis

The first research question was ***'What are appropriate discussion performance goals for the Japanese university students in this study?'*** In this section, I give details of the data collection and analysis methods used to address this.

5.5.1 Teacher surveys

The survey given to the teachers was designed to gather and analyze the teacher beliefs about the 'importance' of different factors for assessing the individual performance of first-year students in the department during group discussion tests (Appendix D). Survey items were categorized under (1) when speaking, (2) when taking speaking turns, (3) when reacting to speaking turns or (4) when problems occur. I chose 21 observable measures in total for the survey, based on my experience as a teacher and assessor of group discussion tests, as well as the research literature discussed in Chapter Three related to assessing communicative competence within interactive settings (Celce-Murcia, 2007; Celce-Murcia et al., 1995; Clapham, 2000; Fengying, 2003; Lynch & Anderson, 1992; Rignall & Furneaux, 1997).

5.5.2 Teacher follow-up interviews

The interviews were done using a semi-structured open-ended approach (Silverman, 2001) to allow the teachers to clarify any unclear information in the survey and expand upon reasons for their answers for each item. However, no survey responses were changed by the teachers following these interviews.

Upon completion of all ten interviews, I reviewed my notes and audio recordings again to help summarize common reasons given for the overall preferences of the teachers to use items in the list for assessing students in discussions or not. These were used to further clarify the overall choice of items to include in my summary of goals for discussions.

5.6 RQ2 data collection and analysis

The second research question was ***'(a) How does observable discussion task performance change for the students across a semester using a TBLT approach (regardless of the type of GSF used)?, (b) What different effects do Product and Process GSF have on observable performance across a semester? and (c) Are these effects the same for Low and High Participators?'*** Details of the methods used to collect and analyze data for this are discussed in this section.

5.6.1 Classroom observation data collection

The recordings of classroom discussions and tests (Section 5.2.2) were used to analyze the changes in discussion performance across time. Audio files using microphones on the desks and video files using cameras were collected. Recordings were collected simultaneously for multiple groups during classroom practices and for one group at a time during tests. After each class/test was completed, the matching sound and video files for each group were combined using software and a single video then created to use for analysis of performance across time using SPSS version 24.

5.6.2 Discussion performance measures selection

Using the literature discussed in Chapter Three, a wide range of measures were selected to assess group discussion performance (Table 6.3). Firstly, various measures which have been commonly used to assess oral participation, as well as complexity, accuracy and fluency (CAF) within oral tasks were chosen. Within Section 3.2, participation within discussions was determined to be appropriately measured using words spoken and turns taken and is the approach taken in the study. Measuring fluency was described as best represented by measures of speed (speech rates), breakdowns (pauses) and repairs (repetitions and reformulations) within speech and were therefore also selected. The measurement of accuracy is based on the recommended approach in Chapter Three of using errors in speech (via PEFC and errors per 100 words), and using an analysis of syntax (clauses per t-unit and words per t-unit) and speaking turns (words per turn) for complexity. All of these measures are the same as those used within Module Two of this PhD (see Appendix B), which involved the same discussion tasks as this thesis in terms of group size, discussion topics, length and goals. The consistency in measures between these two modules helps build on the previous findings within this PhD, as Module Two took a cross-sectional approach to analyzing discussions, which were analyzed longitudinally (across a semester) in this thesis.

Secondly, discussion performance was analyzed using the measures within the Process GSF sheet and diary (Table 6.2). These measures were decided upon using the literature about group interactions as a measure of discussion performance within Section 3.3.1 and the opinions of the teachers about those measures within RQ1. This analysis was used to see how the ProcS, who directly addressed these process-related measures, might be able to improve at them across

time. In addition, this could reveal if the ProdS, whom did not address them directly, might improve at them by only focusing on the outcome of the task with their GSF.

A third approach to the analysis was more qualitative in nature and what I called 'product-focused' performance, based on the ‘outcome’ of a task being considered an important aspect of performance within Section 3.3.3. Some differences were found in the data in the way in which ProdS and ProcS had used their spoken turns to finalize the group decision (the output goal of the discussions). These differences were demonstrated by classifying student turns as either 1) *on-task* (related to the discussion topic, but not viewed as promoting a discussion outcome), 2) *off-task* (unrelated to the discussion topic or group outcome) or 3) *outcome-promoting* (related to the topic and group outcome). The results of RQ2 (Section 6.3.4.1) provide more details and examples of these turns. Differences in the occurrences of these types of turn between ProdS and ProcS revealed the difference in focus of discussions. For instance, a discussion without any outcome-promoting turns and many off-task turns can be said to be less focused on an outcome than a discussion with only on-task turns and many outcome-promoting turns.

Finally, an analysis of other performance considerations was undertaken to go beyond the linguistic, process and product measures above. This was discussed in Section 3.3 as an important way to establish a clearer picture of communicative competence (Clapham, 2000; Fengying, 2003) and was also used to look at changes in language use not captured by the other measures. This included differences in the use of clarifications, turn-taking strategies and possessive pronouns by the ProdS and ProcS (Section 6.3.4). This additional data may help reveal more about how the two types of GSF may have aided SLA on the whole (Fulcher, 2003;

Ghaith, 2002; Norris, 2008). However, the focus of this thesis was to look at changes to measures which both students and teacher could realistically assess using feedback within classrooms and tests. I saw this data as more important within the context of the participation issues that many teachers find for their TBLT group discussions (Carless, 2009b; Littlewood, 2007, Thomas & Reinders, 2015), as it attempted to give feedback which was expected to directly aid future performance with a formative-style of assessment for learning (as discussed in Section 4.3 and recommended by researchers such as Black & Wiliam, 2009; Chappuis, 2005; Stiggins et al., 2004). Therefore, I used the limited space of this thesis to focus mainly on this.

5.6.3 Discussion transcript coding and analysis

Discussion recordings for one group from each of the six classes were selected for analysis (three using Product GSF and three using Process GSF). Each group was in a different class to reduce the threat of other factors within a class, such as class dynamics or the time of day of classes, affecting the perceived impact of the GSF on learning over time. The six groups which I chose had all four members present every week of the semester (to avoid smaller group sizes in some weeks being unaccounted for in the learning, feedback and assessment). They also represented all different possible combinations of LP and HP member mixes (Table 5.3) to try and account for various combinations teachers may experience in their classes.

For the six groups, the three classroom discussion videos and three test videos were then watched, transcribed, and analyzed with various performance measures (see Appendices K, L, M and N for examples). This included the words spoken (including repetitions, reformulations added in two types of brackets), pauses more than a second in length, sentence breaks (where I

considered sentences to start and stop), and turn times for calculating speech rates. I also made comments about paralinguistic information and body language which I used to help confirm what type of turns were being taken by students if it was not clear (such as head movements or hand gestures suggesting that a turn is intended to be a disagreement, an opinion or a question). This was a very time-consuming process and a limitation for replication of the study by other teachers. Not only is this work load perhaps not practical for some teachers, but another problem was that I could only begin answering the second research question a long time after I had completed the analysis of the data. Many interesting points arose with regards to the student performances across time, but by the time these points were revealed, I was already teaching all new students. This gave me little chance to follow up on any points within the findings with additional surveys, interviews or discussion data for example. Table 5.5 shows how the transcripts were coded in preparation for calculating performance measures with Excel formulae.

Table 5.5 Discussion coding and analysis process

Transcript version	Process	Measures calculated
1. Unpruned participation and fluency analysis	Initial transcribing of discussion videos using formulated Excel templates, followed by coding of repetitions, reformulations, pauses>1s, timing of t-unit turns (using PRATT software to calculate SRA)	Syllables and words spoken, repetitions per minute, reformulations per minute, pauses>1s per minute, SRA
2. Pruned fluency, accuracy and complexity analysis	Pruning of transcript 2 by removing coded repetitions and reformulations. Coding added for (non) t-units, (non) t-unit turns, errors, and error (free) clauses	T-unit turns per minute, SRB, errors per 100 words, PEFC, clauses per t-unit, words per t-unit, words per turn
3. Unpruned discussion process analysis	Coding of process measures (opinion, supporting reason, example, agreement, disagreement), outcome-promoting turns and off-task turns	Opinions, reasons, examples, questions, opinions and (dis)agreements, total outcome-promoting turns and total off-task turns

5.7 RQ3 data collection and analysis

The third research question was ***(a) How do ProdS and ProcS report feeling about performing in discussions across the semester? and (b) How do they report feeling about the support the two types of GSF provided for their learning (or not)?*** In this section, I give details of the data collection and analysis methods used to address this.

5.7.1 Attitudinal surveys about classroom discussions, tests, goal-setting and feedback

The first part of the attitudinal survey (Appendix I) contained closed-ended questions about student feelings towards undertaking English discussions. Students rated (along a seven-point Likert scale) their overall enjoyment, motivation, confidence, anxiety, and self-perceived ability within discussions and their use of English in general, as well as how much time they felt they spent studying for discussions outside of class time. These questions were answered again by the students in Weeks 8 and 13 by rating how much each of these variables was perceived to have changed since the previous survey (also with a seven-point Likert scale). By doing so, changes in self-reported feelings towards studying and doing classroom discussion tasks could be measured across time for the two different groups (product and process). Differences between changes to these variables over time could then be attributed to which of the two GSF styles students undertook during learning.

The second part of the survey in Weeks 3 and 13 comprised an open-ended question about the difficulties students experienced in undertaking the test. Differences between their Week 3 and Week 13 answers were compared to see if their opinions about what difficulties they reported for performing in discussions had changed across the semester. Also, commonly

reported difficulties for tests could help explain any changes (or lack of) in certain performance measures within RQ2 (Section 6.3.5). These would be important findings, as they could be used to help design less difficult tasks and syllabi for Japanese universities.

A third part to the survey in Week 13 consisted of questions about how useful the students found the (1) GSF sheet and (2) GSF diary to improve their discussion performance across time (open-ended), and whether they wished to use them again in future classes (closed-ended). The purpose of this data was to show how the GSF had been received both positively and negatively by the students, how it might be improved in future attempts to support learning, and so that connections might be made between reported test difficulties, GSF benefits/problems and observed performance changes over time.

After collecting and transcribing all of the open-ended responses myself, a Japanese teacher in the department (who was only shown written responses without student details) checked my interpretation of any comments which I was not fully confident about. Due to the large amount of responses collected, it was not possible for the teacher to check every response and as the surveys were done anonymously, I could not check the intended meaning again with the students themselves. However, I felt that I had interpreted and translated them well enough to get an overview of the students' perceptions.

5.7.2 Student peer-interviews

The peer-interviews were conducted in Japanese immediately after discussion tests in an adjacent classroom with no time limit, but usually took about 25-30 minutes each. The interviewers sat side by side with the students and spoke in their first language, Japanese. After

each class was complete, the interviewers transcribed and translated the interviews (into English) in Excel files using the recordings while alone in a classroom on the campus. Within three days they met me and we read the transcripts together while highlighting key points mentioned by students. I then used these sheets to code points made by students and look for patterns by myself (Section 5.7.3). One reliability issue with this was the trust put into the ability of the interviewers to transcribe and translate the interviews by themselves. I did not have enough time to check the written transcripts against the audio recordings of the interviews while I was running my classes and collecting and processing observational, survey and interview data across the semester. Any errors made by the interviewers would have gone unseen, but I made it clear to them to check their transcripts carefully at least three times before presenting them to me.

5.7.3 Student response coding and analysis

For all of the (1) attitudinal survey responses and (2) peer-interview transcripts, the procedure shown in Table 5.6 was followed, which was similar to that recommended by Ellis and Barkhuizen (2005). This approach was also similar to Thematic Analysis (Braun & Clarke, 2006), in which all of the data gets coded before themes are decided for categorizing responses.

Table 5.6 Survey and interview responses coding and analysis procedure

Step 1	Read the responses within Excel sheets and highlighted and made notes on any points mentioned by the student
Step 2	Created and refined a list of points after reading the data several times very closely
Step 3	Created a coding system for each point
Step 4	Affixed the codes to the responses within the data (sometimes more than one point within a response)
Step 5	Used Excel formulae to count the occurrences of each point with the files
Step 6	Looked for patterns among the responses, merged repeated points, and created summary tables of the responses

I chose this procedure because after reading the responses and making notes on the points made by students (Step 1 in Table 5.6), most of the comments made in both surveys and interviews were short in nature and simple enough in meaning to immediately list and code (Steps 2-4). After that, I counted the occurrences of each coding, and then grouped them together in themes as a final stage to the analysis (Steps 5-6). This seemed more appropriate for the kind of data I had than to start by creating themes and sub-themes and categorizing each response under them, such as with Framework Analysis (Ritchie, Spencer & O'Connor, 2003) and Interpretive Phenomenological Analysis (Smith & Osborn, 2008). I felt that creating themes/categories at an early stage of the analysis generalized the responses too early on and would require too many alterations to those themes/categories as the analysis continued (which I could just create as a final stage to the coding procedure anyway).

Despite these choices for the analysis, some problems were encountered with coding some responses (see Section 6.4.8 and Appendices O, P and Q for some specific examples). In surveys, students sometimes did not clarify the difficulties they had with tests, and did not give very clear reasons why they found the GSF useful or not. For the interviews, even though many student opinions were explained well, the interviewers did not always succeed at eliciting clear reasons for stated opinions. This made some student opinions difficult to confidently interpret. To counter any potential debate about the codes I chose to assign such responses, broadly-worded categories were used to summarize the overall opinions of the students, reducing the chance of responses being under the correct category (see the tables in Section 6.4).

5.8 Ethical issues

The guidelines of the British Educational Research Association (BERA, 2011) were used to ensure that the interests of the students were protected during the study in terms of what are described as 'voluntary informed consent', 'openness and disclosure', 'right to withdraw', 'incentives', 'detriment arising from participation', 'privacy' and 'disclosure' (pp. 5-8).

5.8.1 Data collection

Informed consent was given in written form by all teacher and student participants before any data was collected (Appendix H). All participants were provided with both a written (in English, as shown in Appendix G) and verbal (in both English and Japanese for students) explanation of the purpose and potential benefits of the study, the way in which data would be stored securely and confidentially, how all data would be anonymized, information about the right to not take part in the study, and the right to withdraw at any time (Dörnyei, 2007). Participants were given the opportunity to ask questions in either Japanese or English to clarify any information they wished to during and after class, or by making an appointment to see me. I also explained to the students that none of the data collected would affect their treatment on their course in any negative way. While it is not possible to ascertain whether some students may have felt an undue obligation to participate, I felt that I had done all I could to clarify the situation and was also confident that taking part would not be harmful to their studies in any way.

While completing surveys, students were not observed or pushed to answer more than they wished to in open-ended questions. Interviewees had volunteered to undertake interviews and it was made clear that they would need to do so three times during the semester (at the start,

middle and end). During the collection of survey, interview and observational data, I did not apply any pressure to students to give more detail or speak more if they did not wish to. The students were free to give as much or as little data as they wished to and were free to choose to not give any if they did not want to. To further ensure this, surveys were completed immediately after tests in a separate room while I continued to observe tests for other groups. Students were free to write as much as they chose to and leave that room without me observing them or knowing how much time they spent completing their survey.

Observational data was initially stored on my cameras (video recordings) and the main classroom computer (password protected audio recordings). This was immediately deleted from the cameras and classroom computer once transferred to my own computer and was password protected for the duration of the study. Once this PhD has been completed, all video and audio files are schedule for deletion. Survey and interview responses were also stored and password protected on my home computer and are scheduled for deletion after completion of the PhD. All hand-written interview notes, audio recordings and transcription files created by the interviewers and myself were stored in a locked cabinet within my office when not being analyzed and are also to be destroyed upon completion of this PhD.

5.8.2 Teaching considerations

Consideration was given to the need to ensure that all students in the study received a fair and equal education. Regardless of which group they were in for the study, all of the students undertook similar classroom practices each week. All of the students also experienced one of two kinds of GSF, so no students were used as a 'control' group without this kind of support for

learning. Neither GSF approach used (Product or Process) was believed to be 'better' than the other and so no clear advantage for learning was given to any students.

To further ensure equality of learning for all of the students, I made my own feedback and support available across the study and took time to help any student who requested support from me during class time or outside of class at the campus (by requesting meetings with me via email or during class). Although some students may not have had the confidence to utilize this support, I felt it was an important message of openness and support from my side, as their teacher.

One possible difference which did arise between the two groups was the self-reported anxiety when undertaking discussions. The ProcS reported less of a decrease in anxiety to undertake discussions across time than the ProdS. The ProcS also reported more issues in surveys and interviews related to trying to gain their individual scores within the discussion time allowed, which was not reported as often by the ProdS. The individual accountability of scoring placed on students with the Process GSF did appear to be more stressful than for students receiving points as a group total (Product GSF). I became aware of this added pressure for the ProcS across time and attempted to reduce this pressure by making extra efforts as a teacher to give positive feedback on individual performances after tests. It was perhaps unfair that the ProdS, unlike the ProcS, did not necessarily have to be an active group member to get a score, as their performance was assessed as a combined group effort. However, I decided not to add pressure to ProdS participating less than others in their group, as they had the right to participate as much as they wanted to.

Finally, I designed the peer-interviewing process to be as academically beneficial for the two student interviewers as possible, in addition to the pay they received. The benefits I felt that they received were learning how to collect research data via interviews, how to design a university syllabus, and the overall considerations to be made when undertaking a research project. A debriefing session with both interviewers at the end of the semester helped them discuss and reflect upon these points with me.

CHAPTER 6. RESULTS AND DISCUSSION

6.1 Chapter Introduction

The main RQ to be answered in the study was ***'What are the effects on learning of using Goal-Setting and Feedback (GSF) with TBLT group discussions across a semester?'*** The first part of this chapter explains how process performance measures were decided for the students (RQ1), to create the GSF sheets and diaries. This is followed by a discussion of the potential problems and limitations to the approach used. The second part discusses the effects of Product versus Process GSF on learning for the students across a semester using observations of classroom practices and tests (RQ2). Implications for learning are then discussed, as well as limitations to the approach used to collect and analyze the data. In the third part, survey and interview data from students across the semester are used to better understand student feelings towards the learning and GSF with discussions (RQ3), followed by a discussion of implications for learning and limitations to the data collection and analysis. Finally, a summary of the results gives an overview of the data with regards to the potential effects of the GSF used and overall implications for learning.

6.2 RQ1: Appropriate discussion performance goals

6.2.1 Introduction

The starting point of data collection was with regards to the first research question *'**What are appropriate discussion performance goals for the Japanese university students in this study?'*** The purpose of RQ1 was to establish both process (based on individual spoken

performance and interaction with others during discussions) and product-focused (based on agreed group outcomes for discussions) frameworks for GSF for students in the study. I expected both types of GSF rubrics to increase transparency of what the teachers wanted students in the department to do in class (Reynolds- Keefer, 2010; Schamber & Mahoney, 2006) as well as help the students self-regulate their efforts by focusing on improving their performance related to those goals in their learning (Andrade & Du, 2005; Panadero et al., 2012).

I elicited opinions from ten teachers within the department via surveys and interviews, as well as drawing on my own opinions filtered through a pilot of the Process GSF and teacher journal notes during that time. Many of the teachers in the study reported that they were very happy that I was doing this research, as they did not have a set guideline to follow for their required assessment of group discussion performance. Results and a discussion of each of these are given below. Possible alternatives to the selection of goals made in this section are not discussed in great depth, as the focus of the overall study was to analyze the overall effects of Process and Product GSF on learning over time, rather than an in-depth analysis of different types of goals. However, general recommendations for adapting the GSF used for different educational settings are given in Section 7.2.

6.2.2 Teacher survey and interview results

Table 6.1 shows the results from the teacher survey (Appendix D) regarding opinions of the ‘importance’ of different variables to assess student performance within discussion tests (see Section 5.5.1 for justifications for this chosen list).

Table 6.1. Teacher ratings of individual process measures for individual assessment in discussions
(not at all important = 0, not very important = 1, slightly unimportant = 2, slightly important = 3, important = 4, very important = 5)

Category	Process performance measure	Mean score	Standard deviation	Teachers scoring 4+
When speaking	**Stating a clear opinion**	**4.40**	**0.84**	**8/10**
	Giving a reason to support an opinion	**4.60**	**0.52**	**10/10**
	Giving an example to support an opinion	**3.80**	**0.79**	**6/10**
	Giving several reasons/examples to support an opinion	2.90	1.37	4/10
	Speaking in long sentences	2.10	1.52	2/10
	Speaking in long turns	1.60	1.58	2/10
When taking speaking turns	Asking for a turn	1.60	1.58	2/10
	Taking a turn when someone else is speaking	2.10	1.60	3/10
	Keeping a turn when being interrupted	2.20	1.93	4/10
	Giving a turn to someone else	**3.40**	**1.35**	**5/10**
When reacting to speaking turns	**Answering follow-up questions**	**3.80**	**1.14**	**8/10**
	Agreeing with at least one reason	**3.10**	**1.60**	**5/10**
	Disagreeing with at least one reason	**3.50**	**1.43**	**6/10**
	Using several reasons or examples to support (dis)agreements	2.50	1.78	3/10
	Back-channeling	2.90	1.45	4/10
	Asking appropriate follow-up questions	**3.70**	**1.16**	**7/10**
When problems occur	Using English fillers	2.20	1.75	3/10
	Asking for English help with English from other students	**3.40**	**0.97**	**6/10**
	Giving help with English to other students	**3.40**	**0.70**	**5/10**
	Clarifying information well	**3.80**	**0.63**	**7/10**
	Paraphrasing well	**3.40**	**0.97**	**6/10**

Variables with a mean score of three or more (being rated on average as 'slightly important' or more by the teachers), and rated either 4 or 5 by at least half of the teachers, are shown in bold. I used this system to assess the perceived importance of the different process measures based on the fact that students should only focus on a limited number of performance measures during feedback, starting with those most highly rated. It is important to be selective about the feedback a teacher gives students during learning, so that students have enough time to understand and use that feedback to improve without a teacher providing too much feedback and confusing students within the time available (Wiliam, 2018, p. 131). By having the teachers rate

the measures for importance, I was able to reduce the list from 22 to 12 appropriate measures for assessment, which could also be used to provide continuous feedback to students on their performance across the semester. These measures were all related to meaning-based interactions within discussions, with performance being viewed by the teachers as the ability to give, explain and react to each other's opinions, rather than the ability to improve one's own turn fluency and complexity by taking and speaking in long turns, using several reasons and examples within the same turn, or using fillers or back-channeling.

6.2.2.1 Giving opinions

The performance variables within the survey rated most highly were stating clear opinions (M = 4.40, SD = 0.84) and giving a reason to support opinions (M = 4.60, SD = 0.52). In addition, being able to explain opinions by giving examples was reported to be of importance (M = 3.80, SD = 0.79). However, there was a lower scoring among the teachers for the students to give several reasons to explain their opinions (M = 2.90, SD = 1.37), although the higher standard deviation for this measure than for the other two measures suggests a greater difference in opinions. Also, speaking in long sentences (M = 2.10, SD = 1.52) and long turns (M = 1.60, SD = 1.58) were not rated as highly, or consistently, as the other variables related to speaking turns.

On the whole, the teachers reported in surveys that giving opinions with only one reason, as well as perhaps an example, would be an appropriate focus within speaking turns for discussion tests for the students at hand. There was a feeling among most of the teachers that the students did not need to expand on or increase the complexity of their contributions by trying to

give several reasons or examples for opinions, or by using long sentences or turns. The interviews which followed the surveys confirmed these beliefs among the teachers. Statements included "*Just getting them to state a clear opinion with good grammar in their sentences is enough for them to focus on. Complete sentences are important", "They should not be worried about long sentences or turns or giving lots of reasons. They should just show us they can give an opinion on the topic",* and *"Speaking in long turns and sentences isn't needed as it doesn't improve the quality of what they are saying".*

6.2.2.2 Taking speaking turns

Giving turns to others within discussions was rated as appropriate for performance (M = 3.40, SD = 1.35), with half of the teachers reporting it as important or very important (4 or 5). However, more direct turn taking skills were not viewed as significant for performance. Asking for a turn (M = 1.60, SD = 1.58), taking a turn from someone who is speaking (M = 2.10, SD =1.60) and keeping a turn when being interrupted (M = 2.20, SD =1.93) all had a mean score of below three in the surveys and were reported to be important or very important by less than half of the teachers.

Although some teachers did report these variables as important for performance (scoring it 4 or 5 in the survey), the interviews revealed further doubts among teachers for the use of turn-taking measures for assessment. Firstly, demonstrating the use of English to take turns was viewed as too different from the way in which students would communicate with each other as Japanese students. Comments included "*Letting them take turns without interrupting each other is fine, such as communication is in Japan"* and *"Taking and offering turns with English is too*

business like with their friends and not normal for Japanese students". Secondly, turn-taking language was reported to be too difficult to teach to the students at hand and too time consuming and confusing as a focus within the time-frame given to improve discussion skills (fifteen ninety-minute classes in this case). This was shown by interview comments such as "*It's not a good focus for a course for the students and shouldn't be assessed. It takes too much time to teach and isn't needed"* and *"It's too difficult for the students to ask for or give turns or interrupt. They should wait for each other's turns without interrupting, as they could lose their chain of thought".* Finally, the use of English to specifically take turns was described as irrelevant and 'unnatural' with comments such as *"That's not at all important as a focus, as they will take turns naturally without words", "The students don't sound real when they speak like that and they shouldn't practise doing it"* and *"Expressing their feelings is important and not the negotiation of turns they make in English".*

6.2.2.3 Reacting to speaking turns

In terms of reacting to other speakers, measures rated most highly were asking appropriate follow-up questions (M = 3.70, SD = 1.16), answering follow-up questions (M = 3.80, SD = 1.14), agreeing with at least one reason (M = 3.10, SD = 1.60) and disagreeing with at least one reason (M = 3.50, SD = 1.43). All four of these variables had high standard deviations, indicating some differences in opinions of the importance of each of these spoken acts for performance. Despite this, the fact that so many variables related to reacting to other speakers were scored with a mean value of three or more, as well as four or more by at least half of the teachers, suggests their perceived importance for assessment. However, giving several

reasons or examples to support agreements or disagreements were not rated so highly in the surveys (M = 2.50, SD = 1.78). This was similar to the teacher opinions about the need for students to only give single reasons and examples for opinions (Section 6.2.2.1).

Within the follow-up teacher interviews, asking and answering questions was reported as relevant because "*students should all follow through on opinion giving by asking questions of each other. This is 'natural' discussion"*. This was also described as *"important as it shows they are actively listening to each other".* Agreeing and disagreeing with reasons was also explained as important, as it shows a teacher that students are considering the points of other group members and actively reacting to them with their own opinions. Teachers made statement such as *"They are not just waiting for others to speak"* and *"are giving reasons to support what they think about what others are saying which shows they are really having a discussion and not just speaking separately".*

Back-channeling had a lower rating as a performance factor in the surveys (M = 2.90, SD = 1.45). This was further explained in the interviews, as it was described as being used 'unnaturally' and 'unnecessarily' within interaction. One teacher said *"The students will learn just one type of back-channel, like 'I see' and use it unnaturally and too much, so shouldn't get a score in the test. Just nodding is ok".* Non-verbal back-channeling, such as nodding, was reported in the interviews to be enough from the students during interactions, without the need for them to verbally back-channel to demonstrate performance.

6.2.2.4 Clarifying turns

Finally, there were several variables related to dealing with discussion problems when they occur which were scored with a mean of over three. These were asking for help with English (M = 3.40, SD = 0.97), giving help with English (M = 3.40, SD = 0.70), clarifying information (M = 3.80, SD = 0.63) and paraphrasing (M = 3.40, SD = 0.97). However, the use of fillers to deal with problems in a discussion was rated less highly and more inconsistently (M = 2.20, SD = 1.75). Thus, 'problem' handling which merits assessment within discussions seemed to be viewed as mainly clarifying spoken turns during interaction with others, rather than addressing fluency within individual speech, such as using fillers.

However, the interview responses revealed less certainty about the importance of these measures for learning and assessment. Comments included "*Helping each other understand is more about attitude than ability. Students can solve these kind of problems themselves without being made to by the teacher"* and *"This is too much for covering with the students during the course and they need more time to focus on simple exchanges and sentence formulation instead".* Fillers were also described as being *"too complex" and "unimportant and not worth worrying about for performance".*

6.2.3 GSF pilot and teacher journal results

I ran a pilot for the Process GSF (Section 5.4.1) using the most highly scored measures of performance within the teacher surveys (bold in Table 6.1). It is also important to note that none of the teachers suggested any additional measures of task performance within their interviews,

having been given one week (between the survey and interviews) to consider this. Therefore, no additional measures were added.

After the pilot, I decided to remove the 'giving a turn to someone else' measure. Even though it was reported as significant for performance in the teacher survey (M = 3.40, SD = 1.35), I observed some potential problems with it within my teacher journal (Appendix C, Part 1). During the pilot, students were recording scores for themselves for saying turns as short as 'you?' or 'you're next. Go'. I felt this was too easy for students to use to get points which were unrelated to the discussion content. Also, the negative comments made about using turn-taking for assessment within the teacher interviews (Section 6.2.2.2) further suggested a need for me to delete it.

Secondly, I removed all of the variables related to 'when problems occur'. During the pilot, I overheard many of the students commenting to each other that they were too 'confusing' and 'distracting' (Appendix C, Part 1). When using the Process GSF sheets and diaries, students often asked me how and when to ask for help, give help, or paraphrase. From my own observations, most of these acts were done unnecessarily for clarifying interactions, as the teachers had also suggested from experience (Section 6.2.2.4). For instance, students would ask for help with English, give each other help with English or paraphrase, in order to show 'good' performance, even when other group members understood what they had just said.

6.2.4 RQ1 results summary

Table 6.2 summarizes the performance variables which I decided to adopt as the most appropriate measures for process-focused GSF for individual students within discussions. They

were the occurrences of (1) giving opinions, reasons and examples, (2) asking and answering questions, and (3) agreeing or disagreeing with others with at least one reason. These were incorporated into the Process GSF sheets and diaries used by the students in the study (Appendix F).

Table 6.2. Finalized individual process measures for students during discussions

Category	Process GSF Measure
Giving opinions	Clear opinions
	Clear reasons (following opinions / (dis)agreements)
	Clear examples (following opinions / reasons)
Understanding	Follow-up questions
	Follow-up question answers
Agreeing/Disagreeing	Agreements with a turn (with at least one reason)
	Disagreements with a turn (with at least one reason)

For the Product-focused group, I decided to use only the measures in Table 6.2 which I felt could be applied to an 'outcome' of a discussion decided by the group. This was done with the intention of using the GSF to focus those students on the product (as opposed to the process) of the discussion within their learning. However, the measures in Table 6.2 were created for the Process GSF and it is important to note that the Product GSF was created using only my own opinions and was not piloted (discussed as a limitation within Section 6.2.6). The product-related measures which I chose were 1) the final choice by the group, 2) reasons for that choice, 3) examples for that choice, as well as 4) reasons for other ideas which had been suggested during the discussion. The first of these measures was unrelated to those in Table 6.2, but the other three were created using the reasons and examples measures for the Process GSF. For the Product

GSF, there was no direct focus placed on the process of interactions which should occur (with no feedback on whether they asked/answered questions or agreed/disagreed with each other). To summarize, the Product GSF (Appendix E) was only focused on reaching an outcome together for groups, whilst the Process GSF (Appendix F) was only focused on how the group members performed as individuals during the discussion time.

6.2.5 RQ1 discussion

This section discusses the suitability and possible implications for learning of the two different types of GSF created (product and process-related) for use in the study.

6.2.5.1 Performance rubric considerations

The teacher surveys and interviews in this section, as well as the piloting of the Process GSF, were **effective at creating a clear rubric for discussion performance which represented the beliefs of the teachers within the department**. Table 6.2 shows how the exchanging, explaining, expanding and commenting on opinions by each other during discussions was viewed as the most appropriate for individual assessment. The original survey measures which were not included in the final list (Table 6.1) related to students speaking in long sentences and turns, giving several reasons or examples for opinions, as well as using English for turn-taking, asking for help, giving help, clarifying, paraphrasing, and back-channeling. The removal of these measures, as well as the lack of suggestions by the teachers for any additional measures, indicated that the teachers felt less need to see students focusing on the Complexity (within sentences and turns), Accuracy or Fluency (CAF) of their speech. Thus, the focus of individual

assessment for process GSF in this study was decided as participating orally as much as possible by communicating and responding to others on topics using opinions, reasons, questions, agreements and disagreements (Section 3.3.1).

Despite this, the process measures in Table 6.2 **provided quite a limited view of performance**, and the analysis of how the students communicated needs to go beyond this list to get a fuller perspective of oral communicative competence (Clapham, 2000; Fengying, 2003; Fulcher, 2003). Although the students did not receive feedback on it, I ran a much broader analysis of the observational data within RQ2 (Table 6.3). This data was a central part of my study which would provide teachers with specific details of how the Product and Process GSF used may affect performance beyond the goals addressed directly within the GSF, as well as what they can observe in real time during classes or tests. This included the use of several CAF measures, as well as a more qualitative analysis of the discussions (Tables 6.9 and 6.10).

Another point to consider is that the rubrics created **do not directly connect to the official language tests** which students typically undertake in their time at university in Japan (such as TOEIC and TOEFL interview speaking tests). Although this was not the focus of the study, connecting the scores which students receive with the two rubrics to such tests is important to consider as a factor for student motivation and is recommended in the conclusions (Section 7.2).

6.2.5.2 GSF and learning considerations

As a teacher within the department myself, **I believed that the finalized process GSF measures (Table 6.2) were suitable for the low-level students in the study** (Section 5.3.2). My

observations of discussions within the pilot (Appendix C), revealed that these goals focused student efforts more on negotiation of meaning, rather than specific linguistic measures (such as errors being made with CAF). As a result, the students would be expected to be more confident and willing to speak, with less fear of making mistakes, which was discussed in more depth in Module One as a reported 'barrier' to discussion participation among university students in Japan (See Appendix A; Chang, 2011; Tsui, 2001; Williams & Andrade, 2008; Woodrow, 2006). A common problem for students in the university (which was often discussed by the teachers and myself during work), and perhaps in many others in Japan, is the lack of motivation to speak within oral discussions. By providing feedback to the students on how they could perform and/or complete discussion tasks, rather than pointing out errors in their speech for example, I expected both the process and product GSF created in RQ1 to help improve this motivation and increase the amount they would speak (Section 4.3) which is analyzed in RQ2 (Section 6.3.3.1).

However, using only the seven measures chosen for the Process GSF had the **risk of making students focus on specific interactional elements of discussions too much** (such as asking a lot of questions to get points) and focus less in their learning on other important areas of performance which students without these goals may do (Nahrgang et al., 2013). It may be the case that students who are not provided with the individual Process GSF, based on Table 6.2, and are only focused on the group outcome, are just as able to develop their oral language use across a semester in a similar way and to a similar degree. Task-Based Language Teaching (TBLT) assumes that groups having an 'output' focus for an oral task will encourage SLA among students (Bygate et al., 2013; Ellis, 2003, p. 8; Skehan; 1998). Observational data in RQ2 and

students' self-reported data in RQ3 analyzed how true this may be for group oral discussions for the students in the study.

Also, in terms of student motivation, it is **important to analyze any potential negative effects** of this very interactional approach to GSF created for ProcS. This may include the added pressure of reaching individual goals within a time limit, as opposed to making a group effort (as with ProdS). This has been shown to have the negative effect of making individual students focus more on finding ways to reach a goal rather than improving their language use (Seijts & Latham, 2001). This is discussed using data in the sections for RQ2 and RQ3 in this chapter.

6.2.6 Limitations

Four main limitations existed for the data collection and analysis in relation to RQ1. Firstly, I had to assume that the teachers gave honest responses about their assessment beliefs, but problems with *social desirability* (telling me the answers they thought a teacher should give, rather than their own opinion) may have affected reliability (Dörnyei, 2003) (Sections 5.2.3 and 5.2.4). Secondly, the opinions came from only ten teachers within one Japanese University. A larger variation of teacher data from different universities across Japan would be needed to see whether the list in Table 6.2 is considered suitable for other university classes. Thirdly, the final Process GSF sheet and diary were created using my own interpretation of the teacher data and my own observations during the Process GSF pilot, and the Product GSF sheet and diary using my own judgement of how an outcome should be assessed. Although the teachers appeared pleased with the list when I presented it to them on an unofficial basis, including them in the final selection of the measures for both the Process and Product GSF may have produced

different versions. The teachers were unfortunately unavailable to help me in such a way because of their very busy schedules at that time, which is why I interpreted the data by myself. Finally, the interviews did not go far beyond confirming the teacher survey responses, such as learning more about what they often described as 'natural' English use (Section 6.2.2), or how they believed the performance measures might be best implemented into learning. More detailed and in-depth discussion may have yielded data that would have been useful for further clarifying what 'good' discussion performance was perceived as and justifying the selection and application of the GSF. However, the lack of time I had to interview the teachers made these additional explanations difficult to obtain. Also, even with these more detailed opinions about performance, I do not believe that I would have made any further alterations to the measures selected, as those which I selected had all been reported to be important by at least half of the teachers.

6.3 RQ2: Changes in observable discussion performance

6.3.1 Introduction

After process goals were decided in RQ1, I created the GSF sheets and diaries for both ProdS and ProcS (Appendices E and F). These were used across a semester to answer RQ2: ***'(a) How does observable discussion task performance change for the students across a semester using a TBLT approach (regardless of the type of GSF used)? (b) What different effects do Product and Process GSF have on observable performance across a semester? and (c) Are these effects the same for Low and High Participators?'*** The following sections summarize the performance changes for selected variables within classroom discussions and tests for students

using the Product or Process GSF. Repeated measures ANOVA tests were run with SPSS version 24 to see whether performance measures altered with statistical significance between the beginning, middle and end of the semester. Follow-up pair-sampled t-tests were then run for all of the performance measures between the different classroom tasks and tests to isolate where any statistically significant changes had occurred for the two groups between those different time periods. Also, results for LPs and HPs within both groups are discussed, although t-tests were not performed for these sub-groups due to the lack of reliability for such small group sizes (only six LP and six HP students categorized under each group). In summary, I discuss the implications for the overall performance measure changes for ProdS and ProcS, other potentially important performance changes which I observed, and the differences between LPs and HPs among the ProdS and ProcS.

6.3.2 Overview of performance measure changes

The overall repeated measures ANOVA results are shown below in Table 6.3. Statistically significant results ($p<0.05$) are shown in bold and are used as a point of reference for the sections which follow. The analysis revealed significant changes over time within every category in the table except for complexity of language use. Using Cohen's (1988) recommendation, the effect size of the GSF on the variables over time was determined using partial eta squared (η_p^2) (small at .01, medium at .09 and large at .25). If measures had no significant ANOVA result and/or increased and decreased significantly without a clear pattern across time within Tables 6.4 - 6.8 in the next section (such as faster speech rates by the middle

of the semester, but then slower again by the end of it), I decided not to try and interpret these changes. This is discussed later as a challenge I faced in the analysis (Section 6.3.8).

I made two alterations to the original Process category of measures in Table 6.3, due to problems I had with interpreting the data. Firstly, 'examples' given by students were very few in number (usually only once or so in a discussion, as evident in Appendices K-N) and did not change with any significance across time anyway. Therefore, I decided to not include them in Table 6.3, as I did not want to try and draw conclusions from such unclear and insignificant changes. Secondly, due to the low number and high standard deviations of turns with opinions, agreements and disagreements in them, I combined them in a measure called 'opinion / (dis)agreement turns'. These represented a total of turns which I considered to be very similar in content and structure, and allowed me to see clearer and more reliable statistical significance (or not) for changes within the data over time. However, it can be argued that giving an opinion, agreeing and disagreeing are different skills within discussions, which is a possible limitation to the data.

Table 6.3. Repeated measures ANOVA results

			CLASSROOM TASKS					TESTS				
Category	Variable	Group	Type III Sum of Squares	Mean Square	F (2, 22)	Sig.	η_p^2	Type III Sum of Squares	Mean Square	F (2, 22)	Sig.	η_p^2
Participation	Words spoken	Product	15.714	7.857	0.660	0.527	0.057	6.318	3.159	0.201	0.819	0.018
		Process	2736.500	1368.250	0.913	0.416	0.077	5764.389	2882.194	2.356	0.118	0.176
	Turns per min	Product	0.072	0.036	0.198	0.822	0.018	0.022	0.011	0.193	0.826	0.017
		Process	1.070	0.535	2.800	0.083	0.203	2.154	1.077	7.540	**0.003**	**0.407****
Fluency	SRA	Product	1104.892	552.446	0.454	0.641	0.040	3678.967	1839.483	7.182	**0.004**	**0.395****
		Process	11763.775	5881.888	5.384	**0.012**	**0.329****	7297.662	3648.831	2.511	0.104	0.186
	SRB	Product	800.296	400.148	0.303	0.742	0.027	3478.497	1739.249	3.885	**0.036**	**0.261****
		Process	11359.226	5679.613	6.067	**0.008**	**0.355****	6361.039	3180.520	1.990	0.161	0.153
	Pauses per min	Product	10.983	5.492	0.420	0.662	0.037	55.695	27.847	5.695	**0.010**	**0.341****
		Process	121.824	60.912	15.979	**0.000**	**0.592****	37.201	18.600	2.168	**0.038**	**0.165****
	Repetitions per min	Product	22.867	11.434	2.778	0.084	0.202	12.635	6.318	0.761	0.479	0.065
		Process	21.591	10.796	1.389	0.270	0.112	59.548	29.774	5.362	**0.013**	**0.328****
	Reforms per min	Product	5.303	2.651	1.035	0.372	0.086	31.474	15.737	7.904	**0.003**	**0.418****
		Process	0.766	0.383	0.161	0.853	0.014	4.262	2.131	0.914	0.415	0.077
Accuracy	Errors per 100 words	Product	98.515	49.257	3.948	**0.034**	**0.264****	53.864	26.932	2.973	0.072	0.213
		Process	106.208	53.104	4.213	**0.028**	**0.277****	25.440	12.720	0.328	0.724	0.029
	PEFC	Product	698.158	349.079	1.726	0.201	0.136	155.793	77.896	0.875	0.431	0.074
		Process	978.874	489.437	3.533	**0.047**	**0.243***	38.936	19.468	0.064	0.938	0.006
Complexity	Clauses per t-unit	Product	0.023	0.011	0.190	0.828	0.017	0.003	0.001	0.034	0.967	0.003
		Process	0.054	0.027	0.778	0.472	0.066	0.073	0.037	1.228	0.312	0.100
	Words per t-unit	Product	16.961	8.481	1.698	0.206	0.134	9.378	4.689	1.784	0.191	0.140
		Process	1.657	0.829	0.491	0.619	0.043	11.821	5.911	2.196	0.135	0.166
	Words per turn	Product	72.336	36.168	2.380	0.116	0.178	6.712	3.356	0.466	0.634	0.041
		Process	7.491	3.745	1.613	0.222	0.128	124.370	62.185	2.514	0.104	0.186
Process	Opinions	Product	10.500	5.250	1.262	0.303	0.103	7.389	3.694	1.850	0.181	0.144
		Process	24.889	12.444	4.968	**0.017**	**0.311****	28.222	14.111	4.325	**0.026**	**0.282****
	Reasons	Product	6.222	3.111	0.623	0.545	0.054	11.056	5.528	0.989	0.388	0.083
		Process	9.500	4.750	0.934	0.408	0.078	37.556	18.778	4.779	**0.019**	**0.303****
	Questions	Product	2.889	1.444	0.087	0.917	0.008	1.167	0.583	0.064	0.938	0.006
		Process	27.056	13.528	1.389	0.270	0.112	87.167	43.583	3.374	**0.050**	**0.235***
	Opinion / (dis)agreement turns	Product	31.056	15.528	2.888	0.077	0.208	0.056	0.028	0.009	0.991	0.001
		Process	7.167	3.583	0.622	0.546	0.053	28.389	14.194	2.957	0.073	0.212

Notes. ***Bold text*** *= significant finding, * = sig<0.05 and medium effect size, ** = sig<0.05 and large effect size, Product N = 12, Process N = 12*

Within class practices, ProdS showed significant changes for accuracy (errors per 100 words with a large effect size), while ProcS showed changes for fluency (large effect sizes for SRA, SRB and pauses per minute), accuracy (a large effect size for errors per 100 words and medium effect size for PEFC) and process measures (large effect size for opinions). With tests, ProdS showed significant changes for fluency (large effect sizes for SRA, SRB, pauses per minute and reformulations per minute), while ProcS showed significant changes for participation (large effect size for turns per minute), fluency (large effect size for repetitions per minute) and process measures (large effect sizes for opinions and reasons, as well as a medium effect size for questions). These findings are further analyzed in the following sections and possible reasons for these changes are discussed later (Section 6.3.6).

6.3.3 Specific performance measure changes

Although the highlighted measures in Table 6.3 were found to change with significance, the ANOVA results do not tell us exactly how performance changed for the students, or between which weeks of the semester the changes were most significant. Follow-up paired-sample t-test results for each variable within Table 6.3 are now discussed in order to clarify this. Changes found to be statistically significant are shown and explained below (using t, p, CI and effect size values) and any unclear pattern of changes over time and/or those with no ANOVA test significance are shown in red. Effect sizes for significant changes are described using Cohen's (1988) description for paired-sample t-tests (small = 0.2, medium = 0.5 and large = 0.8). Each section also discusses differences in performance for LPs and HPs (the students who

spoke the least and most words in discussions, as defined in Section 5.4.2) among the ProdS and ProcS.

6.3.3.1 Participation

The overall follow-up paired sample t-test results for words spoken and turns taken are summarized in Table 6.4. **Within classroom practices and tests, changes to words spoken by both ProdS and ProcS were statistically insignificant** (Table 6.4). This did not match my prediction of increases in words spoken across time with the use of GSF (Section 6.2.5.2), but it is worth noting that ProcS did start with a higher mean number of words within classroom practices (M = 143.75, SD = 77.39) and tests (M = 141.33, SD = 55.13) than ProdS within classroom practices (M = 107.50, SD = 72.69) and tests (M = 116.75, SD = 73.67). This shows some non-equivalence between the two groups at the start of the study which is discussed later as a limitation to the data (Section 6.3.8).

Although there were no statistically significant changes to word count across the semester for the students, there were clear differences between LP and HP mean words spoken (in eight minutes). **Within classroom practices, LP ProdS and ProcS increased their mean words spoken much more than HPs.** LP ProcS increased their mean words spoken by 66.00% from 82.83 (SD = 41.85) in Week 2 to 137.50 (SD = 33.33) in Week 12, and by 68.33% from 53.17 (SD = 13.21) in Week 3 to 89.50 (SD = 33.12) in Week 12 for ProdS. Thus, all of the classroom group discussion learning undertaken appeared to result in similar increases in words spoken for the quieter students in discussions (LPs), but small overall decreases in mean words spoken were seen for the students who spoke more at the start of the study (HPs).

Table 6.4. Participation paired-sample t-test results

			CLASSROOM TASKS						TESTS					
			Week 2		Week 7		Week 12		Week 3		Week 8		Week 13	
Variable	Group	Sub-Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Words spoken (8m)		OVERALL	107.50	72.69	117.58	63.43	119.58	51.74	116.75	73.67	124.83	71.74	133.08	88.94
	PRODUCT	LP	53.17	13.21	78.33	37.87	89.50	33.12	72.33	36.18	72.17	26.58	72.83	28.93
		HP	161.83	66.07	156.83	60.99	149.67	51.18	161.17	76.79	177.50	62.91	193.33	88.63
		OVERALL	143.75	77.39	149.75	55.09	164.50	43.73	141.33	55.13	170.75	79.93	164.50	58.93
	PROCESS	LP	82.83	41.85	123.83	38.23	137.50	33.33	100.50	29.47	124.17	45.34	136.50	40.00
		HP	204.67	50.19	175.67	60.03	191.50	36.69	182.17	42.61	217.33	82.42	192.50	64.48
T-unit turns / minute		OVERALL	1.25	0.83	1.17	0.70	1.15	0.49	0.98	0.49	1.03	0.47	0.98	0.61
	PRODUCT	LP	0.81	0.49	0.98	0.64	0.90	0.40	0.79	0.43	0.81	0.30	0.67	0.36
		HP	1.69	0.89	1.35	0.76	1.40	0.47	1.17	0.51	1.25	0.52	1.29	0.68
		OVERALL	1.20	0.59	**1.30**	**0.49**	**1.60****	**0.45**	**0.88**	**0.45**	**1.28****	**0.60**	**1.46***	**0.33**
	PROCESS	LP	0.79	0.45	1.27	0.41	1.44	0.45	0.56	0.27	1.02	0.29	1.44	0.32
		HP	1.60	0.41	1.33	0.60	1.77	0.41	1.19	0.36	1.54	0.74	1.48	0.37

Notes. ***Bold text*** *= significant finding,* ***Red text*** *= unclear pattern of changes over time and/or no ANOVA test significance, * = 10-week sig<0.05, ** = 5-week sig<0.05, OVERALL N = 12, Low Participator (LP) N = 6, High Participator (HP) N = 6*

In tests, HPs among the ProdS increased their words spoken much more than LPs from Week 3 (M = 161.17, SD = 76.79) to Week 8 (M = 177.50, SD = 62.91) and again by Week 13 (M = 193.33, SD = 88.63). However, **the LPs among the ProcS accounted for most of the increases in words spoken in tests**. LPs increased mean words spoken from the Week 3 test (M = 100.50, SD = 29.47) to the Week 8 test (M = 124.17, SD = 45.34) and to the final Week 13 test (M = 136.50, SD = 40.00). However, HPs had less of an increase than LPs between the first (M = 182.17, SD = 42.61) and final test (M = 192.50, SD = 64.48).

The only statistically significant change found for participation across time for either ProdS or ProcS (Table 6.4) was that **within classroom practices and tests, ProcS significantly increased their number of turns per minute.** They increased turns between Week 7 (M = 1.30, SD = 0.49) and Week 12 (M = 1.60, SD = 0.45), t (11) = -2.52, p = 0.03, CI = -0.56 to -0.04, d = 0.64 (medium effect size), as well as between Week 3 (M = 0.88, SD = 0.45) and Week 8 (M = 1.28, SD = 0.60), t (11) = -2.58, p = 0.02, CI = -0.75 to -0.06, d = 0.75 (medium effect size). However, ProdS did not experience any such changes.

Also, within classroom practices and tests for ProcS, the increases in turns for LPs was significantly larger than for HPs (Table 6.4). In the tests for example, LPs almost tripled their mean turns taken between the Week 3 test (M = 0.56, SD = 0.27) and the final Week 13 test (M = 1.44, SD = 0.32), while HPs increased their number of speaking turns with much less significance (from M = 1.19, SD = 0.36 to M = 1.48, SD = 0.37).

In summary, if we compare the test participation changes for LPs and HPs among the ProcS with that among the ProdS, there are large differences. ProdS tests generally became more dominated by HPs in terms of words spoken and turns taken. An example of this can be seen

with student 3 in Appendix L, who did a larger proportion of the speaking for the group by the final test, but not in Week 3 (Appendix K). However, tests with ProcS became much more equal in terms of who took a speaking turn and how much each group member said. Appendix N shows an example of this with the quite equal spread of speech between the columns for the four speakers by the Week 13 test. The significant implications which these two changes over time may have for learning are discussed later (Section 6.3.6.3).

6.3.3.2 Fluency

Changes in the fluency performance measures across the semester are shown by the follow-up paired sample t-test results in Table 6.5. **Within classroom practices, ProcS delivered speech with ever increasing speed across time, but ProdS did not.** ProcS significantly increased their SRA from Week 2 (M = 139.86, SD = 47.31) to Week 12 (M = 183.60, SD = 23.53), t (11) = -2.92, p = 0.04, CI = -76.73 to -10.75, d = 1.17 (large effect size), as well as their SRB from Week 2 (M = 107.17, SD = 35.52) to Week 12 (M = 149.39, SD = 27.82), t (11) = -3.26, p = 0.01, CI = -70.74 to -13.70, d = 1.32 (large effect size). However, ProdS spoke with almost exactly the same mean SRA in Week 2 (M = 137.94, SD = 25.28) and Week 12 (M = 138.64, SD = 38.29) and SRB in Week 2 (M = 114.41, SD = 25.25) and Week 12 (M = 115.07, SD = 34.27).

Table 6.5. Fluency paired-sample t-test results

			CLASSROOM TASKS						TESTS					
			Week 2		Week 7		Week 12		Week 3		Week 8		Week 13	
Variable	Group	Sub-Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Speech Rate A (SRA)		OVERALL	137.94	25.28	150.03	36.66	138.64	38.29	**119.88**	**21.08**	**144.57****	**21.53**	**130.58****	**27.54**
	PRODUCT	LP	141.15	32.74	141.87	45.92	131.22	39.87	104.71	13.02	141.01	21.44	117.83	15.62
		HP	134.73	17.57	158.18	26.23	146.06	38.79	135.04	16.00	148.12	23.03	143.33	32.15
	PROCESS	OVERALL	**139.86**	**47.31**	167.69	37.80	**183.60***	**23.53**	**154.19**	**44.19**	161.82	38.00	**187.48***	**39.19**
		LP	122.11	46.05	165.96	48.27	177.49	31.24	131.47	36.38	146.33	36.46	194.94	39.74
		HP	157.60	45.25	169.42	28.40	189.70	12.36	176.91	41.64	177.31	35.65	180.01	40.83
Speech Rate B (SRB)		OVERALL	114.41	25.25	124.72	47.18	115.07	34.27	**97.83**	**18.96**	**120.80****	**19.90**	**103.04****	**31.11**
	PRODUCT	LP	114.92	35.26	116.32	62.12	108.94	35.45	82.95	8.35	121.66	23.73	91.30	20.07
		HP	113.90	12.62	133.13	29.48	121.19	35.16	112.71	13.76	119.93	17.51	114.78	37.36
	PROCESS	OVERALL	**107.17**	**35.52**	137.39	35.43	**149.39***	**27.82**	123.57	43.14	128.16	35.33	153.78	46.32
		LP	95.53	37.52	139.70	43.64	141.58	31.13	103.71	39.81	116.67	35.94	163.28	50.81
		HP	118.81	32.30	135.08	29.05	157.20	24.24	143.43	39.54	139.65	33.73	144.29	43.84
Pauses / min of speech		OVERALL	13.63	2.88	12.55	5.30	13.01	2.57	**15.94**	**3.10**	14.03	3.37	**12.92***	**3.23**
	PRODUCT	LP	13.86	4.13	12.51	7.48	13.57	2.94	18.18	2.77	15.86	2.10	13.84	3.86
		HP	13.39	1.01	12.59	2.41	12.45	2.28	13.69	1.17	12.20	3.53	12.01	2.47
	PROCESS	OVERALL	13.27	2.38	**12.25**	**2.26**	**8.96****	**1.87**	**13.06**	**3.49**	**10.57****	**1.75**	11.68	3.75
		LP	13.77	2.89	12.17	3.19	10.14	1.34	14.49	3.41	10.80	1.31	12.98	4.54
		HP	12.77	1.87	12.33	0.99	7.78	1.60	11.63	3.21	10.34	2.21	10.38	2.52
Repetitions / min of speech		OVERALL	5.65	3.50	**4.74**	**3.06**	**6.69****	**4.11**	6.10	6.62	5.15	3.35	4.68	3.11
	PRODUCT	LP	6.36	3.60	5.06	3.59	7.95	3.75	7.69	6.87	5.60	3.69	4.33	2.95
		HP	4.94	3.58	4.42	2.72	5.43	4.40	4.52	6.57	4.70	3.24	5.03	3.51
	PROCESS	OVERALL	7.24	4.74	6.79	3.38	8.61	4.46	**4.99**	**3.11**	**7.74****	**3.97**	**7.69***	**4.95**
		LP	6.52	2.51	5.98	2.38	7.04	4.09	4.11	1.89	8.12	3.62	7.30	4.98
		HP	7.97	6.47	7.60	4.22	10.18	4.59	5.86	3.98	7.36	4.61	8.08	5.35
Reformulations / min of speech		OVERALL	1.90	2.10	1.99	1.57	2.75	1.31	2.02	1.62	**1.99**	**1.48**	**3.99****	**2.05**
	PRODUCT	LP	1.13	1.29	1.03	1.13	2.65	1.15	1.12	1.38	0.96	1.25	4.01	2.30
		HP	2.67	2.56	2.95	1.39	2.86	1.56	2.93	1.37	3.02	0.83	3.97	1.99
	PROCESS	OVERALL	4.03	2.37	3.68	1.17	3.80	2.09	3.67	1.83	4.28	3.23	3.47	1.67
		LP	2.78	1.03	3.35	0.72	3.59	1.62	3.26	1.74	4.20	2.27	2.65	0.71
		HP	5.29	2.74	4.01	1.50	4.01	2.63	4.07	1.99	4.36	4.22	4.29	2.01

Notes. ***Bold text*** *= significant finding,* ***Red text*** *= unclear pattern of changes over time and/or no ANOVA test significance, * = 10-week sig<0.05, ** = 5-week sig<0.05, OVERALL N = 12, Low Participator (LP) N = 6, High Participator (HP) N = 6*

Also within classroom practices, both LPs and HPs among the ProcS increased their speech rates (by between 30-50%)**, but neither LPs nor HPs among the ProdS did**. This shows that all of the students who undertook Process GSF had the potential to improve their fluency in classroom discussions across time, via faster speech, regardless of how much they usually spoke in discussions.

In addition, **LPs among the ProcS, but not ProdS, increased their speech rates during tests more than HPs**. In the Week 3 test, ProdS LPs spoke with a mean SRA of 131.47 (SD = 36.38) and HPs with a mean of 176.91 (SD = 41.64). However, by the Week 13 test, LPs increased their mean SRA to 194.94 (SD = 39.74) and HPs to 180.01 (SD = 40.83). Even though LPs started with slower speech in tests than HPs (which may explain why they were categorized as LPs, who spoke less during discussions) they accounted for a large amount of the increases in speech rates across time for the group. This suggests that Process GSF in discussions may encourage students who speak less than others in tests to begin to speak more quickly in successive tests. However, it may also mean that students who already speak the most within group discussions may not speak any more quickly in tests within that same group.

In classroom practices, ProcS began to pause less often, but ProdS did not. Their number of pauses decreased between Week 7 (M = 12.26, SD = 2.26) and Week 12 (M = 8.96, SD = 1.87), $t(11) = 4.075$, $p < 0.01$, CI = 1.51 to 4.08, $d = 1.59$ (large effect size). However, this improvement in fluency was not seen for ProdS. Their mean pauses per minute in classroom discussions in Week 2 (M = 13.63, SD = 2.88) was almost identical to that in Week 12 (M = 13.01, SD = 2.57). **Also in classroom practices, both LPs and HPs among the ProcS decreased pauses.** LPs decreased mean pauses per minute from 13.77 (SD = 2.89) in Week 2 to

10.14 (SD = 1.34) in Week 12, and HPs from 12.77 (SD = 1.87) in Week 2 to 7.78 (SD = 1.60) in Week 12. This suggests that the Process GSF can result in improvements in classroom fluency (via less pausing) for both LPs and HPs.

Within tests, both ProdS and ProcS decreased their pauses, especially LPs. ProdS decreased pauses between Week 3 (M = 15.94, SD = 3.10) and Week 13 (M = 12.92, SD = 3.23), t (11) = 4.14, $p < 0.01$, CI = 0.02 to 0.08, d = 0.95 (large effect size). ProcS also decreased pauses between Week 3 (M = 13.06, SD = 3.49) and Week 8 (M = 10.57, SD = 1.75), t (11) = 2.19, p = 0.05, CI = -0.02 to 4.99, d = 0.90 (large effect size).

Within tests, ProdS increased reformulations and ProcS increased repetitions, but neither changed within classroom practices. Specifically, ProdS increased their reformulations between Week 8 (M = 1.99, SD = 1.48) and Week 13 (M = 3.99, SD = 2.05), t (11) = -3.35, p = 0.01, CI = -3.31 to -0.68, d = 1.12 (large effect size) and ProcS increased their repetitions between Week 3 (M = 4.99, SD = 3.11) and Week 8 (M = 7.74, SD = 3.97), t (11) = -3.37, p = 0.01, CI = -4.55 to -0.95, d = 0.77 (medium effect size), as well as across the whole semester between Week 3 (M = 4.99, SD = 3.11) and Week 13 (M = 7.69, SD = 4.95), t (11) = -2.94, p = 0.01, CI = -4.73 to -0.68, d = 0.65 (medium effect size).

In summary, within classroom practices, the Process GSF, but not Product GSF, encouraged both LPs and HPs to speak faster and with fewer pauses. However, within tests, both the Process and Product GSF helped reduce pauses (especially LPs), but increased repetitions and reformulations, and had very uncertain results for speech rates. These points are discussed more in terms of their significance later (Sections 6.3.6.2 and 6.3.6.3).

6.3.3.3 Accuracy

Changes in accuracy across time for both ProdS and ProcS are shown in Table 6.6. **Within classroom practices, both ProdS and ProcS decreased errors per 100 words between the middle and end of the semester.** ProdS decreased errors from Week 7 ($M = 16.63$, $SD = 5.20$) to Week 12 ($M = 12.63$, $SD = 4.62$), $t(11) = 3.20$, $p < 0.01$, CI = 1.25 to 6.75, $d = 0.81$ (large effect size). ProcS also decreased errors from Week 7 ($M = 15.06$, $SD = 4.90$) to Week 12 ($M = 10.90$, $SD = 4.64$), $t(11) = 3.74$, $p < 0.01$, CI = 1.71 to 6.60, $d = 0.87$ (large effect size). **No significant difference was found between LPs and HPs among the ProdS or ProcS for the errors made.** Therefore, there was no clear connection between how many words students say in discussions and how many errors they make across different discussions.

Table 6.6. Accuracy paired-sample t-test results

			CLASSROOM TASKS						TESTS					
			Week 2		Week 7		Week 12		Week 3		Week 8		Week 13	
Variable	Group	Sub-Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Error / 100 words	PRODUCT	OVERALL	15.18	5.20	**16.63**	**5.20**	**12.63****	**4.62**	**15.16**	**4.39**	15.60	3.76	**12.82***	**4.34**
		LP	16.86	6.54	17.28	5.71	13.60	2.27	16.03	5.58	16.18	5.05	12.88	5.40
		HP	13.50	3.13	15.99	5.09	11.67	6.28	14.30	3.07	15.02	2.17	12.76	3.52
	PROCESS	OVERALL	13.56	5.57	**15.06**	**4.90**	**10.90****	**4.64**	14.92	7.04	13.47	4.62	12.93	5.19
		LP	13.45	6.02	13.92	2.21	11.14	4.07	13.93	4.31	15.47	4.31	14.05	6.55
		HP	13.67	5.65	16.20	6.69	10.66	5.53	15.90	9.39	11.47	4.33	11.81	3.67
Percentage of Error-Free Clauses (PEFC)	PRODUCT	OVERALL	**43.16**	**12.92**	**32.66****	**16.45**	40.04	26.05	38.13	11.18	39.87	10.24	34.85	15.57
		LP	42.76	16.63	31.30	20.37	29.17	20.24	34.28	8.91	41.89	12.94	31.37	16.86
		HP	43.55	9.51	34.02	13.26	50.91	28.27	41.99	12.65	37.84	7.32	38.33	14.84
	PROCESS	OVERALL	46.61	17.45	**40.79**	**16.13**	**53.54****	**11.93**	46.03	18.78	43.49	16.07	44.90	17.97
		LP	49.33	20.32	41.35	6.01	54.07	9.68	45.08	23.79	36.67	15.73	44.27	20.02
		HP	43.89	15.46	40.23	23.14	53.01	14.79	46.99	14.41	50.31	14.45	45.53	17.57

Notes. ***Bold text*** *= significant finding,* ***Red text*** *= unclear pattern of changes over time and/or no ANOVA test significance, * = 10-week sig<0.05, ** = 5-week sig<0.05, OVERALL N = 12, Low Participator (LP) N = 6, High Participator (HP) N = 6*

6.3.3.4 Complexity

Changes in complexity across time for both groups are shown in Table 6.7. **In classroom practices and tests, all changes in complexity measures for both ProdS and ProcS were insignificant.** According to the ANOVA results in Table 6.3, students who undertook process or product-related GSF did not speak with statistically more or fewer clauses per t-unit, more words per t-unit or words per speaking turn across the semester.

However, one result which I felt needed highlighting was that **between the second and final test, ProcS spoke in significantly shorter turns, but ProdS did not**. ProcS said a statistically lower number of words in each turn from Week 8 (M = 11.08, SD = 4.54) to Week 13 (M = 8.19, SD = 2.45), t (11) = 2.64, p = 0.02, CI = 0.48 to 5.30, d = 0.79 (medium effect size). Although the ANOVA results (Table 6.3) did not find significant changes for the words per turn for ProcS, I decided to include this t-test result in my discussion of the results. This was firstly because the mean values of the words per turn did decrease across time within tests for ProcS (especially for LPs). This is something which I also observed while scoring ProcS tests (Appendix C, Part 2), as well as within the transcripts. The shortening of speaking turns between Weeks 3 (Appendix M) and Week 13 (Appendix N) gives a clear example of this. ProcS, especially LPs, were noticeably taking shorter turns as time went on in order to reach their scores during the shared speaking time of tests. This is discussed further as a significant finding for Process GSF within this study (Section 6.3.6.2).

Table 6.7. Complexity paired-sample t-test results

			CLASSROOM TASKS						TESTS					
			Week 2		Week 7		Week 12		Week 3		Week 8		Week 13	
Variable	Group	Sub-Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Clauses/Pruned T-unit		OVERALL	1.33	0.44	1.27	0.20	1.30	0.33	1.33	0.29	1.32	0.30	1.31	0.30
	PRODUCT	LP	1.38	0.58	1.25	0.22	1.21	0.42	1.27	0.32	1.27	0.40	1.16	0.16
		HIP	1.28	0.28	1.29	0.21	1.38	0.22	1.40	0.27	1.36	0.18	1.47	0.35
		OVERALL	1.36	0.21	1.28	0.27	1.28	0.16	1.50	0.25	1.40	0.31	1.42	0.20
	PROCESS	LP	1.35	0.23	1.26	0.22	1.35	0.13	1.44	0.18	1.40	0.20	1.34	0.17
		HP	1.37	0.21	1.31	0.33	1.21	0.16	1.57	0.31	1.40	0.41	1.50	0.21
Words/Pruned T-unit		OVERALL	8.72	2.85	9.03	1.50	10.31	2.34	9.00	1.61	8.80	1.32	9.97	3.12
	PRODUCT	LP	8.47	3.46	8.61	1.07	11.36	2.70	8.53	1.68	8.46	1.57	8.94	1.16
		HP	8.97	2.41	9.46	1.84	9.26	1.46	9.47	1.53	9.14	1.03	11.00	4.19
		OVERALL	8.49	1.36	9.01	1.81	8.72	1.45	10.57	2.14	**9.37**	**1.31**	**10.60****	**1.81**
	PROCESS	LP	7.88	1.17	8.39	1.94	8.44	1.11	11.15	2.49	9.20	0.90	10.10	2.20
		HP	9.10	1.34	9.64	1.58	9.00	1.79	9.99	1.75	9.53	1.70	11.09	1.34
Words/Pruned turn		OVERALL	6.06	3.64	9.42	8.26	6.98	3.99	7.94	4.99	8.96	5.14	8.23	4.72
	PRODUCT	LP	3.86	1.33	6.56	4.28	4.74	2.06	5.30	1.70	6.13	2.93	5.07	2.17
		HP	8.26	3.97	12.28	10.58	9.23	4.32	10.58	5.92	11.80	5.50	11.39	4.51
		OVERALL	7.61	2.63	8.60	2.18	7.66	1.26	12.68	9.47	**11.08**	**4.54**	**8.19****	**2.45**
	PROCESS	LP	7.50	3.10	8.60	2.47	7.62	1.65	14.90	13.15	10.94	4.92	7.27	1.99
		HP	7.71	2.38	8.60	2.08	7.69	0.89	10.46	3.52	11.22	4.59	9.11	2.68

Notes. ***Bold text*** *= significant finding,* ***Red text*** *= unclear pattern of changes over time and/or no ANOVA test significance, * = 10-week sig<0.05, ** = 5-week sig<0.05, OVERALL N = 12, Low Participator (LP) N = 6, High Participator (HP) N = 6*

6.3.3.5 Task process-focused performance

Changes in process-focused performance across time for both groups are shown in Table 6.8. These measures were based on the Process GSF variables created using the data in RQ1 (Table 6.2) which were considered to be important signs of student performance. Changes in process-focused measures across time was potentially the clearest difference between ProdS and ProcS within the study. **For the classroom practices and tests, all of the changes in process measures by ProdS were insignificant.** In short, ProdS did not increase their opinions, reasons, questions, agreements or disagreements during discussions across the semester. However, the t-test data (Table 6.8) did show for classroom practices that ProdS increased their combined opinions, agreements and disagreements with statistical significance between Week 2 (M = 4.42, SD = 3.32) and Week 7 (M = 6.50, SD = 4.21), t (11) = -2.70, p = 0.02, CI = -3.79 to -0.38, d = 0.55 (medium effect size). Despite this, changes in this measure over time were not found to be significant with the ANOVA test results (Table 6.3).

Table 6.8. Task process-focused paired-sample t-test results

				CLASSROOM TASKS						TESTS					
				Week 2		Week 7		Week 12		Week 3		Week 8		Week 13	
Category	Variable	Group	Sub-Group	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Speaking	Opinions	PRODUCT	OVERALL	3.08	1.98	4.08	2.78	4.33	2.23	3.00	1.65	4.08	1.62	3.75	2.09
			LP	1.83	1.47	2.33	1.75	3.67	2.16	1.83	1.33	3.67	1.37	2.67	1.63
			HP	4.33	1.63	5.83	2.56	5.00	2.28	4.17	0.98	4.50	1.87	4.83	2.04
		PROCESS	OVERALL	**2.42**	**1.24**	3.08	1.83	**4.42***	**1.62**	**2.50**	**1.31**	3.67	2.23	**4.67***	**1.44**
			LP	1.83	1.17	3.67	1.97	3.83	2.14	2.00	1.10	3.33	2.50	4.83	1.94
			HP	3.00	1.10	2.50	1.64	5.00	0.63	3.00	1.41	4.00	2.10	4.50	0.84
	Reasons	PRODUCT	OVERALL	3.50	2.24	4.17	2.37	4.50	2.97	4.33	2.46	5.42	2.71	5.58	4.17
			LP	2.33	1.97	3.17	2.40	2.67	1.86	3.00	1.10	3.83	2.71	3.50	2.43
			HP	4.67	1.97	5.17	2.04	6.33	2.80	5.67	2.80	7.00	1.67	7.67	4.68
		PROCESS	OVERALL	4.42	3.18	5.67	2.74	5.17	1.53	**4.50**	**2.54**	5.67	2.53	**7.00***	**3.30**
			LP	2.00	1.26	4.67	1.97	4.83	1.17	3.17	1.60	4.67	1.97	6.00	2.68
			HP	6.83	2.56	6.67	3.20	5.50	1.87	5.83	2.71	6.67	2.80	8.00	3.79
Understanding	Questions	PRODUCT	OVERALL	4.50	4.76	4.67	4.48	5.17	4.59	3.75	4.20	3.83	5.64	4.17	4.32
			LP	2.83	1.72	4.00	5.06	4.00	1.55	2.67	2.80	1.17	0.75	2.50	3.02
			HP	6.17	6.34	5.33	4.18	6.33	6.38	4.83	5.31	6.50	7.23	5.83	5.04
		PROCESS	OVERALL	6.00	4.67	6.17	3.13	7.92	3.23	**4.42**	**3.48**	7.33	7.33	**8.00***	**5.05**
			LP	3.33	3.08	4.67	1.86	7.00	3.46	2.50	2.26	4.00	3.03	7.67	4.50
			HP	8.67	4.63	7.67	3.56	8.83	2.99	6.33	3.56	10.67	9.07	8.33	5.96
Reacting to others	Opinion / (dis)agreement turns	PRODUCT	OVERALL	**4.42**	**3.32**	**6.50****	**4.21**	6.25	2.30	6.83	2.48	6.83	2.41	6.92	2.50
			LP	2.17	1.33	3.83	2.32	5.17	1.60	6.00	2.97	5.50	1.76	5.17	1.60
			HP	6.67	3.20	9.17	4.07	7.33	2.50	7.67	1.75	8.17	2.32	8.67	1.97
		PROCESS	OVERALL	5.92	1.83	6.33	2.39	7.00	2.73	**4.92**	**2.35**	6.17	2.37	**7.08***	**1.88**
			LP	4.67	1.21	6.50	2.95	5.67	2.34	4.00	2.10	5.33	1.97	7.00	1.67
			HP	7.17	1.47	6.17	1.94	8.33	2.58	5.83	2.40	7.00	2.61	7.17	2.23

Notes. ***Bold text*** *= significant finding,* ***Red text*** *= unclear pattern of changes over time or and/no ANOVA test significance, * = 10-week sig<0.05, ** = 5-week sig<0.05, OVERALL N = 12, Low Participator (LP) N = 6, High Participator (HP) N = 6*

The data in Table 6.8 shows much more positive improvements for the process measures across time for ProcS compared to ProdS. **Within classroom practices, ProcS significantly increased their opinions** between Week 2 (M = 2.42, SD = 1.42) and Week 12 (M =4.42, SD = 1.62), t (11) = -3.46, p = 0.01, CI = -3.27 to -0.73, d = 1.39 (large effect size). Also, **within tests, ProcS significantly increased (1) opinions** between Week 3 (M = 2.50, SD = 1.31) and Week 13 (M = 4.67, SD = 1.44), t (11) = -4.06, p < 0.01 , CI = -3.34 to -0.99, d = 1.58 (large effect size); **(2) reasons** given between Week 3 (M = 4.50, SD = 2.54) and Week 13 (M = 7.00, SD = 3.30), t (11) = -2.17, p = 0.05, CI = -5.37 to 0.03, d = 0.85 (large effect size); **(3) questions** between Week 3 (M = 4.42, SD = 3.48) and Week 13 (M = 8.00, SD = 5.05), t (11) = -2.44, p = 0.03, CI = -6.82 to -0.35, d = 0.83 (large effect size); and **(4) combined opinions, agreements and disagreements** between Week 3 (M = 4.92, SD = 3.25) and Week 13 (M = 7.08, SD = 1.88), t (11) = -2.63, p = 0.02, CI = -3.98 to -0.35, d = 1.02 (large effect size) (although the ANOVA data in Table 6.3 was not significant for this final variable). Therefore, a benefit for the students who focused on Process GSF with discussions was that they began to demonstrate more of the interactions, related to those goals, which were deemed as important signs of discussion performance in RQ1 (Table 6.2).

Some interesting differences were found between the changes in performances by LPs and HPs among the ProdS and ProcS. **In classroom practices, LPs and HPs among the ProdS showed similar changes in process measures, but LPs among the ProcS increased their reasons and questions more than HPs.** LPs among the ProcS increased their reasons between Week 2 (M = 2.00, SD = 1.26), Week 7 (M = 4.67, SD = 1.97) and Week 12 (M = 4.83, SD = 1.17). However, HPs actually decreased their reasons between Week 2 (M = 6.83, SD = 2.56),

Week 7 (M = 6.67, SD = 3.20) and Week 12 (M = 5.50, SD = 1.87). Also, the same LPs increased their mean values for questions between Week 2 (M = 3.33, SD = 3.08), Week 7 (M = 4.67, SD = 1.86) and Week 12 (M = 7.00, SD = 3.46), while HPs had similar mean values for Week 2 (M = 8.67, SD = 4.63), Week 7 (M = 7.67, SD = 3.56) and Week 12 (M = 8.83, SD = 2.99).

However, **within tests, LPs among the ProdS had consistently lower mean values than HPs for every process measure.** Even though LPs and HPs had similar changes to mean opinions, agreements, disagreements, reasons and questions across time, HPs were always saying around twice as many of each as LPs during tests. Meanwhile, the development of LPs and HPs within tests for ProcS was very different. Even though LPs among the ProcS started with much lower mean values for each measure compared to HPs (similarly to ProdS), **by the final test in Week 13, LPs and HPs among the ProcS were speaking with a similar number of all of the process measures.** Therefore, the students in the groups which focused on Process GSF became much more equal in terms of the performance measures for assessment in tests (Table 6.2) compared to ProdS (where some students had much higher performance measures than others in the same group).

In summary, ProdS showed no significant changes in process measures across the semester. In addition, LPs among them were consistently showing less of these measures than HPs, especially within the tests. However, ProcS significantly increased their process measures over time, with LPs showing larger improvements than HPs, performing at almost the same level as them by the final test. The possible implications for learning for all of these differences are addressed later (Sections 6.3.6.2 and 6.3.6.3).

6.3.4 Additional performance considerations

This section gives details of additional performance changes which I observed within the transcripts for ProdS and ProcS. These additional perceptions of discussion performance are included in the discussion of the overall learning with TBLT tasks over time, as well as the potential effects of the two types of GSF on learning (Section 6.3.6). By looking at a wider variety of changes in speech and interactions between students across time, a clearer picture can be created for overall changes in performance. Measures in this section are related to collaboration and efforts made by students to understand one another to reach decisions together during discussions, which are also considered as important indications of communicative competence (Celce-Murcia, 2007; Celce-Murcia et al., 1995; Ellis, 2003, p. 8). On the whole, the Product GSF appeared to focus students more on collaborating and understanding each other as a group to reach a task outcome (with regards to their turn content, clarifications and word usage), while the Process GSF seemed to encourage students to undertake more frequent turn-taking and voice their individual opinions. Detailed statistical data and analysis are not used for all of these measures. This was because I was not confident to do so with measures which relied heavily on my own interpretations of student meaning and intentions and because the occurrences of such spoken actions were few in number and difficult to assess for changes over time.

6.3.4.1 Outcome-promoting, on-task and off-task turns

Within the discourse of discussions across the semester, I made three important classifications for turns with regards to task outcome. The first type of turn was labelled as *outcome-promoting*, which I deemed to be taken in order to promote an agreed choice among

group members. An example is shown below (see Appendix L, lines 8-12) where a group is discussing which part-time job is best for a university student:

1 (Student 4): *But I have other idea. I am juggler. I show juggling performance other people. If I am park or other event. It is real. Yes.*
2 (Student 3): *Yes. But we cannot do juggling.*
3 (Student 4): *I know.*
4 (Student 3): Yes. *So what do you think our opinion? I think the best idea is university campus. You too?*
5 (Student 1): *Okay.*
6 (Student 2): *Okay.*
7 (Student 4): *Okay.*

Student 3 (turn 4) interrupts the discussion of juggling to push students to align around a previously mentioned job, one on a university campus. This outcome-promoting turn prompts signs of agreement from the other students (‘utterances of ‘okay’).

The second, and most common, type of turn within discussions were labelled as *on-task*, which were connected to the required outcome of the task, but did not directly promote final decisions between group members as outcome-promoting turns were considered to do.

Lastly, I labelled some turns as *off-task*, in which I judged the content to be neither outcome-promoting or on-task, and have no relevance to achieving an outcome (group decision) for the topic. An example of an ‘off-task’ turn is shown in turn 4 below (Appendix N, lines 40-43), where the students are discussing which country would be best for them to visit together:

1 (Student 3): *My best choice is America because America has many sightseeing and I like baseball so I want to watch the baseball game. For example New York Yankees.*

2 (Student 1): *Where do you want to go there?*

3 (Student 3): *I want to go New York or Los Angeles.*

4 (Student 2): *What is your favorite food in America?*

Student 3 (turns 1 and 3) is considering specific cities and reasons to take a trip there. However, Student 2 (turn 4) asks a question related to food preferences which I deemed not to be connected to the decision the group are trying to make. Also, nobody in the group uses the responses to this question to help make their group decision and so I classified the turn by Student 2 (turn 4) as 'off-task'. Differences in the mean number of total off-task and outcome-promoting turns within discussions can be seen in Table 6.9.

Table 6.9. Mean total group off-task and outcome-promoting turns

			CLASSROOM TASKS			TESTS		
Variable	Group		Week 2	Week 7	Week 12	Week 3	Week 8	Week 13
Off-task turns/minute	PRODUCT	Mean	1.25	0.67	0.25	0.50	0.25	0.00
		SD	1.66	1.07	0.45	0.80	0.62	0.00
	PROCESS	Mean	1.67	2.67	4.00	0.83	4.25	4.00
		SD	1.56	2.27	2.22	0.94	3.89	2.13
Outcome-promoting turns/minute	PRODUCT	Mean	0.67	1.67	1.17	1.08	1.67	0.92
		SD	1.23	1.72	1.03	1.24	1.56	0.79
	PROCESS	Mean	0.08	0.25	0.33	0.00	0.17	0.17
		SD	0.29	0.45	0.65	0.00	0.39	0.58

Notes. Means are for the total of all four group members during the six different groups' discussions (PRODUCT N = 3, PROCESS N = 3)

ProdS said fewer and fewer total off-task turns and more total outcome-promoting turns across time within classroom discussions and tests. This suggests that students who focused on the outcome of their discussions as their goals were less likely to take speaking turns which did not help lead to that outcome. On the other hand, **ProcS took more total off-task**

turns and showed no increase in total outcome-promoting turns within classroom discussions or tests. Thus, students who focused on Process GSF with discussions did not use turns to make final decisions and would instead go 'off-task' and no closer to making those decisions together. In fact, ProcS almost never said anything which I perceived in transcripts to promote an outcome in such a way. This may not be surprising, as ProcS were not assessed for output of their discussions, but demonstrates a clear difference in discussion content to the groups of ProdS.

6.3.4.2 Clarifications

The transcripts also showed that **ProdS were generally clarifying each other's turns more than ProcS by the end of the semester**. After listening to another group member speak, ProdS would sometimes take time to clarify the speaker's opinion by using short clarification questions, such as 'why?', or by repeating the part of the turn they wished to clarify as a question (see Appendix K, Lines 7-12 and Appendix L, Lines 31-36 for examples). However, this was almost never seen for ProcS, who by Week 13 would follow other speaking turns with turns connected to getting scores within the GSF (asking a follow-up question, or giving their own opinion). This was of course to be expected from ProcS, as the GSF guided them to develop their responses to each in this way, but the lack of clarifying which occurred between speakers made me wonder whether the students actually understood what each other was saying. The teachers in the study rated clarifying and paraphrasing as important for assessment (Table 6.1) and my decision to remove it from the GSF may have caused ProcS to not develop this skill, while ProdS continued to practise it.

6.3.4.3 Turn-taking strategies

Another interesting development within the discussions was that **ProdS and ProcS developed different turn-taking strategies**. ProcS became noticeably faster at switching between speakers by Week 13 and, as a result, were able to take more turns within discussions across time. This was almost certainly because ProcS knew that the more turns they took, the higher scores they could get with the GSF system they were using. Because of this, ProcS developed strategies focused on increasing the turns they could take, such as following a similar order of speakers and turn content each week. A clear example of this in Week 13 can be seen in Appendix N, where student 1 speaks first, answers follow-up questions from the others, other group members agree and disagree with the opinion and then the routine is repeated by students 2, 3 and 4. In addition, the switching of speakers between ProcS was often coordinated by giving turns to each other with utterances such as 'What do you think about my opinion?' and 'What is your choice?' (see Appendix N, Lines 15, 19, 35, 39, and 57 for examples). This 'giving a turn to someone else' was reported by the teachers as being important for assessment (Table 6.1). Although it was removed from the GSF, it appears that the Process GSF became better than the Product GSF at encouraging it among students.

6.3.4.4 Possessive pronoun usage

Differences in language use were also apparent between the two groups by the end of the semester. One particularly interesting point which I noticed within the transcripts was **the use of different possessive pronouns by ProdS and ProcS when referring to opinions and group choices**. By Week 13, ProdS referred to opinions and choices using the word 'our' more often than ProcS, who used the words 'my' and 'your' noticeably more. This can be seen by

comparing the Week 13 transcripts in Appendix L (ProdS) and Appendix N (ProcS). Within these final tests, ProdS said ‘my’ four times, ‘our’ three times, but no one said ‘your’. However, ProcS said ‘my’ nine times, ‘your’ 18 times, but no one said ‘our’. This may show the more collaborative nature of discussions developed with the use of Product GSF, and a more individualized mind-set about performance with the Process GSF, which I address more in the discussion (Section 6.3.6.2).

6.3.5 RQ2 results summary

Table 6.10 shows a summary of statistically significant changes and effect sizes for discussion performance in classroom tasks and tests for both ProdS and ProcS across the semester. Increases and decreases detailed in black in the table indicate both a significant ANOVA and follow-up t-test changes for the variable. Results in red show uncertain data due to significant t-test (but not ANOVA) results, and/or unclear patterns of changes across time. Table 6.11 also shows a summary of the performance changes for LPs and HPs among the ProdS and ProcS.

Table 6.10. Summary of significant ANOVA repeated measures and follow-up t-test results

		CLASSROOM TASKS		TESTS	
Variable	Sub-variable	PRODUCT	PROCESS	PRODUCT	PROCESS
Participation	Words spoken (8m)				
	Turns/min		W7 to W12 increase* (d = 0.64)		W3 to W8 increase* (d = 0.75)
Fluency	SRA		W2 to W12 increase** (d = 1.17)	W3 to W8 increase** (d = 1.16) W8 to W13 decrease* (d = 0.57)	W3 to W13 increase* (d = 0.80)
	SRB		W2 to W12 increase** (d = 1.32)	W3 to W8 increase** (d = 1.18) W8 to W13 decrease* (d = 0.68)	
	Pauses/min of speech		W7 to W12 decrease** (d = 1.59)	W3 toW13 decrease** (d = 0.95)	W3 to W8 decrease** (d = 0.90)
	Repetitions/min of speech				W3 to W8 increase* (d = 0.77) W3 to W13 increase* (d = 0.65)
	Reformulations/min of speech			W8 to W13 increase** (d = 1.12)	
Accuracy	Errors/100 words	W7 to W12 decrease** (d = 0.81)	W7 to W12 decrease** (d = 0.87)	W3 to W13 decrease* (d = 0.54)	
	PEFC	W2 to W7 decrease* (d = 0.71)	W7 to W12 increase** (d = 0.90)		
Complexity	Clauses/pruned t-unit				
	Words/pruned t-unit				W8 to W13 increase* (d = 0.78)
	Length of pruned t-unit turns				W8 to W13 decrease* (d = 0.79)
Process-focused	Opinions		W2 to W12 increase** (d = 1.39)		W3 to W13 increase** (d = 1.58)
	Reasons				W3 to W13 increase** (d = 0.85)
	Questions				W3 to W13 increase** (d = 0.83)
	Opinion / (dis)agreement turns	W2 to W7 increase* (d = 0.55)			W3 to W13 increase** (d = 1.02)

Notes. ***Bold text*** *= significant ANOVA and follow-up t-test finding,* ***Red text*** *= no ANOVA test significance or unclear pattern of changes over time, * = medium t-test effect size (Cohen's d > 0.5), ** = large t-test effect size (Cohen's d >0.8).*

Table 6.11. Summary of significant differences between LP and HP performances

Variable	Group	Classroom tasks summary	Tests summary
Participation	Product	Larger increase in words spoken by LPs.	Larger increase in words spoken by HPs.
	Process	Larger increase in words spoken and turns taken by LPs.	Larger increase in words spoken and turns taken by LPs.
Fluency	Product		Larger decreases in pauses and repetitions for LPs. Larger increases in reformulations for LPs.
	Process		Larger increases in SRA and SRB for LPs. Larger decreases in pauses and reformulations for LPs.
Accuracy	Product		
	Process		
Complexity	Product		
	Process		Larger decreases in words per turn for LPs.
Process-focused	Product		
	Process	Larger increases in reasons and questions for LPs.	Larger increases in total opinions, agreements and disagreements, as well as reasons and questions by LPs.

6.3.6 RQ2 discussion

I now discuss the changes in performance found across the semester for the students. I start by addressing part (a) of RQ2, by discussing general changes found for both groups, regardless of the GSF undertaken. This evaluates the potential for the use of a TBLT approach for improving group discussions performance. Next, I address part (b) of RQ2, by discussing the possible effects of the Product and Process GSF on learning and performance across the semester by looking at statistically significant performance measure changes (Table 6.10) and other performance changes which I observed across time (Sections 6.3.4.1 - 6.3.4.4). After that, I address part (c) of RQ2 by discussing differences in performance changes for LPs and HPs among the ProdS and ProcS (Table 6.11). Key findings are written in bold and are used to discuss what potential TBLT and the two types of GSF may have for helping students of different levels improve at group discussions. Finally, I give details of the main limitations I encountered within the analysis.

6.3.6.1 Overall discussion performance changes with a TBLT approach

The data showed that **the number of words spoken by ProdS and ProcS did not change with statistical significance across time within classroom practices or tests** (Table 6.4). I found this to be of surprise, as a lot of the research discussed earlier suggests that the presence of focused goals within tasks will increase the degree to which students engage within tasks to improve their performance across time (e.g. Miller et al., 1996; Ryan & Deci, 2000; Wolters, 2004). This finding shows that it cannot be assumed that TBLT group discussions will lead to more speech among groups over time, even when goals are applied. According to the several theories discussed in Section 2.2, students who practise listening to, speaking in, and interacting with others in a second language will improve their communicative competence in that language. However, such improvements do not appear to come in the form of more word spoken. This poses a limitation for the use of TBLT with low-level learners if increased participation is a goal within the learning.

A positive performance change found for classroom discussions was that **classroom accuracy (errors per 100 words spoken) improved for both ProdS and ProcS with statistical significance across the second-half of the semester** (Table 6.6). Even with a new topic each week, the use of a TBLT-style approach to classroom discussions (with a ten-minute pre-discussion pair practice, group discussion and post-discussion language form focus) appears to have helped the students make less errors in speech. This is an important finding, as I am unaware of the existence of such data proving that classroom TBLT group oral discussions can lead to improvements in accuracy over time. Therefore, teachers can expect a TBLT approach to group discussions (with pre-task rehearsal, task, and post-task language focus) to improve L2

accuracy across classroom practices (via fewer spoken errors), even if they do not have the time in class to observe groups to see this happening.

However, **test accuracy did not improve for ProdS or ProcS** (Table 6.6), where teachers would be able to directly observe groups one at a time and perhaps consider the errors being made in speech in their assessment. Exactly why accuracy improved across classes, but not between tests, cannot be confirmed with the data available, but has two potential implications for TBLT with group discussions. Firstly**,** a lack of a pre-discussion rehearsal before tests may make it too difficult for students to transfer learning and decrease spoken errors across time. Oral task repetition has been shown to improve accuracy in follow-up tasks with the same content (Bygate, 1996, 2001; Bygate & Samuda, 2005; Lynch & McClean, 2000, 2001). It was my impression, as the teacher assessing the tests, that the lack of test preparation compared to classroom practices was a major factor for this lack of improvement. Students did not have time to practise conceptualizing, formulating and articulating (De Bot, 1992; Levelt, 1989) their opinions and responses to other opinions on that topic beforehand and could not use their short-term working memory (Baddeley, 1986, 1993; Baddeley & Hitch, 1974; Guará-Tavares, 2011, 2013, 2016) to retain pre-prepared language. Therefore, during the tests, students clearly experienced more cognitive load than in classes (having to undertake the construction of this language on the spot) and this was probably the reason they made more errors in speech. This lack of preparation time for tests was also reported as a source of difficulty for performance by both groups of students within surveys and interviews (Section 6.4.3).

Secondly, the added pressure of discussion test conditions may make students more focused on quickly getting test scores (group decisions for ProdS or delivering opinions, reasons

and questions for ProcS), rather than demonstrating other abilities, such as not making spoken errors, which had no direct connection with their score (Huang & Hung, 2013). This added pressure may also have accounted for why both ProdS and ProcS were able to decrease their pauses in tests with significance across time. When students were under pressure to get their test scores, they most likely tried to achieve those scores in a quicker time than in classroom discussions and, therefore, paused less. Teachers need to consider how students should be allowed to prepare for group discussion tests, because if the conditions for practices and tests are different (such as the lack of a preparation stage in this study potentially affecting performance), the assessment may not show teachers how accurately they are using language in classes, but only a faster version of student discussions, with less pauses, but more spoken errors.

Another significant finding was that **complexity of language use did not improve significantly for ProdS or ProcS within classroom practices or tests** (Tables 6.3 and 6.7). In other studies, TBLT has been found to improve complexity of language use when it was combined with pre-task planning stages (see Ellis, 2009 and Javad Ahmadian et al., 2015 for recent summaries). This lack of improvement for complexity with TBLT *group discussion* tasks, which included a planning stage, may be a weakness in its ability to improve language use. However, complexity of language use may not be a suitable measure of performance for low-level learners due to its contradiction with other measures (Section 3.3.1), and because it was not highlighted by the teachers in the study as an important measure of performance (Table 6.2) and was therefore not the focus of either of the GSF approaches (which were based on interactional process or task outcome only). If teachers wish to use a TBLT and/or GSF approach to improve complexity, they must consider what additional support, or alternative approach to TBLT might

be suitable to do so beyond those used in this study. This is discussed more in the conclusions (Section 7.2).

In summary, if we consider the overall changes in performance measures across time in the study, then the use of TBLT for group discussions shows limitations for language development for low-level students. Although classroom accuracy (spoken errors) and test fluency (pauses) did improve, there were no significant improvements seen for classroom participation, repetitions, reformulations or complexity, or for test accuracy or complexity. As one focus of this study was to assess the effectiveness of GSF to improve the performance measures rated most highly for discussions by the teachers (Table 6.1), the lack of development of some common oral task performance measures discussed above is a concern. If teachers wish to see improvements which were not seen in this study (such as accuracy and/or complexity in tests), then the potential 'trade-off' between CAF measures (the Skehan and Robinson debate discussed in Section 2.2.2), as well as between rubric scores and other aspects of performances (such as measures related to interactions), needs consideration within learning, as well as the need for additional support for low-level learners, such as planning stages. These important points are discussed further in the conclusions section.

6.3.6.2 Discussion performance changes with Product and Process GSF

Within classroom practices, only ProcS improved fluency with statistical significance (increased speech rates and decreased pauses shown in Table 6.5). Although ProdS practised the same style of discussions each week, they did not improve this aspect of their performance within classes, which questions whether this style of TBLT and GSF is appropriate

for students if a focus of the course is to improve fluency. Students who can learn to deliver speech more quickly and with less pauses are considered better language users within research (Skehan, 2009; Tavakoli & Skehan, 2005). Although other research (such as Martin et al., 2016) has suggested improved discussion task outcome is possible with the inclusion of group goals (such a making lists of outcomes), this study shows that urgency to speak, and resultant fluency during group discussions, would be more likely to be improved with the motivation provided by individual performance goals (Elliot, 1999) rather than group goals.

Another important finding was that **within tests, only ProcS significantly increased their number of turns, opinions, reasons and questions** (Tables 6.4 and 6.8, and compare Appendices M and N for an example). These improvements matched what the teachers had reported as important indicators of discussion performance in RQ1 (Table 6.1) and this study has shown, for the first time, that by directly addressing these measures with individual goals that students can improve at them over time. This increased effort and improvement in performance may be expected with the use of such goals (Bong, 2009; Miller et al., 1996), but this study has now provided specific observational data to support this claim for group discussions across time.

In addition, student survey and interview responses in RQ3 also showed that ProcS reported their GSF sheet (Appendix F) to be motivating to 'aim for new targets' and for making them 'think of a lot of questions for points' (Table 6.16), but the Product sheet (Appendix E) was not reported to be motivating in this way. This further suggests the inclusion of Process GSF, unsurprisingly perhaps, encouraged students to focus on process measures (such as interacting as much as possible with opinions, reasons and questions), and that they probably increased their number of turns, opinions, reasons and questions because of that focus. It is important that

teachers consider this positive finding for their own classrooms, if they have groups of students who do not speak very often or only with poor fluency during classroom discussion practices. The inclusion of immediate feedback on individual student performance with the use of a sheet such as the Process sheet has the potential to increase the number of these interactions between students in the L2 within discussions.

On the other hand, **the number of turns and process measures for ProdS did not change with any significance within classroom practices or tests** (see Tables 6.4 and 6.8, and compare Appendices K and L for an example). This was perhaps to be expected, as these measures of performance were not the focus of the Product GSF. However, the vast amount of TBLT literature adheres to the belief that requiring a shared 'outcome' for group tasks, as only ProdS had in this study, will encourage interaction between students and resultant SLA (Bygate et al., 2013; Ellis, 2003, p. 8; Skehan, 1998, p. 101). Although ProdS did interact across the semester, with respect to the measures in Table 6.2, only with the presence of Process GSF (ProcS) were students able to actually increase these interactions across time. According to many years of SLA research, the degree to which students acquire language is directly connected to the interactions which students have with each other in that language (Gass & Mackey, 2007; Hatch, 1978; Larsen-Freeman & Anderson, 2013; Long, 1996; Pica, 1994; Savignon, 2002). According to this *interactionist* viewpoint (Section 2.2.1), more interaction can be assumed to result in more SLA and the lack of increases in interactions for ProdS seems to be a weakness to the Product GSF approach compared to the Process GSF. However, deciding how to categorize and count the number of occurrences of different types of interactions within discussions, rather than assessing the specific content of them, is discussed later as a limitation of the analysis (Section 6.3.8).

Another important finding was the increased urgency in turn-taking by ProcS, but not ProdS. As discussed in Section 6.3.4.3, **ProcS showed more turn-taking 'cues' across time** (such as 'What do you think?') to give turns to others and began to follow similar speaking orders across time. These can be considered important for performance, as research considers turn-taking abilities to be an indication of communicative competence (Sacks et al., 1974), the teachers in the study considered the ability to give turns to others important for assessment (Table 6.1 and Section 6.2.2.2), and becoming quicker at switching speaker helped ProcS take more turns and get better discussion scores.

However, **within tests, ProcS decreased mean turn length and increased repetitions**. Although the ANOVA test was not significant for this, I decided to discuss this change, as it matches with my teacher journal observations of ProcS taking noticeably shorter turns across time (Appendix C, Part 2). This possible trade-off between performances expands upon the long-standing argument between Skehan and Robinson (Section 2.2.2) about how the focus on one measure of linguistic performance may prohibit a focus on another at the same time. For the discussion tests, it appears that trade-off may have been shown in this study to apply to other performance change, such as between improved elements of participation (turns taken and mean words spoken) and worsened elements of fluency (repetitions) and complexity (turn length) for ProcS. Therefore, the empirical data in this study has shown that students who are pushed by Process GSF to speak as much as they can in group discussion tests may sacrifice other performance variables in this way, especially if they are not the focus of the GSF. Choosing the focus of GSF within tasks needs careful consideration to support SLA, as it may lead to negative effects for language use which was not connected to the goals used (Seijts & Latham, 2001). In

the case of ProcS, their GSF led them to focus on gaining scores though taking as many turns as possible to demonstrate the process-related acts shown in Table 6.2 of RQ1. There was no direct teacher assessment or GSF related to linguistic accuracy or complexity, so any lack of improvement or worsening of such measures might be the resultant, but expected, 'trade-off' between different aspects of performance of using this focus of process goals. The students' survey and interview responses in RQ3 suggested that ProcS felt more 'pressure' to speak quickly within the discussions than ProdS (Section 6.4.6.3). Problems with delivering as much speech in as short a time as possible was reported much more often by ProcS, rather than how to improve vocabulary and/or grammar usage by Week 13 (Table 6.13).

A final important difference found between ProdS and ProcS was the content of the turns taken. **Within classroom practices and tests, the Product GSF encouraged more clarifications, reformulations and 'outcome-promoting' turns, as well as less 'off-task' turns than the Process GSF** (Section 6.3.4). I interpreted this as a more '*collaborative style'* of discussions where students were spending more time trying to understand each other's ideas and reach an agreed outcome together. This was also demonstrated by the common use of possessive pronouns such as 'we' and 'our', when referring to opinions and choices (Section 6.3.4.4). As discussed in the literature review (Section 2.3.1), collaboration in learning is believed to be beneficial, as students will support each other to improve their social, cognitive and language skills (Ahmadian & Tajabadi, 2017; Johnson et al., 2013; Oxford, 1997). However, it is important for teachers to determine how important they consider collaboration to be for the learning they want students to undertake during classroom practices, as well as during discussion tests for assessment. Product GSF, as defined in this study, can help focus students on reaching

an agreed group outcome more than Process GSF. However, this may also lead to what I observed for ProdS, namely, discussions in which students created an output 'list' of ideas with little explanation (Appendix C, Part 2). This may be considered a task 'outcome', which is often stated as essential for SLA within TBLT (Ellis, 2003, p. 8; Prabhua, 1987; Skehan, 1998, p. 101). However, the observational data suggested that the freedom of language use for ProdS encouraged them to focus more on reaching the task output for points, at the expense of focusing on interacting about the details of those ideas, such as asking each other questions.

Meanwhile, **the Process GSF encouraged more 'off-task' turns, and less 'outcome-promoting' turns and clarifications than the Product GSF.** Although the overall process-related performance measures were found to improve across time for ProcS, especially within tests, I felt that their discussions adopted a very *'individual speech style*' over time. This was also shown by the increasing use of individually-focused possessive pronouns, such as 'my' and 'your', when discussing opinions and choices (Section 6.3.4.4). Although ProcS took more speaking turns and showed a higher number of the performance measures than ProdS by the end of the semester, ProcS never made final decisions about their choice as a group for the task. Reaching such decisions can be argued to be an important outcome of a task for SLA to to occur with a TBLT approach to learning (Section 2.4.3). If adopting a Process GSF approach with discussions, teachers need to decide whether they consider a concrete outcome to discussions to be important for their students' learning. The different effects of the Product and Process GSF discussed here are revisited in the conclusion section to aid with recommendations for teachers of discussion courses.

In summary, the data showed some important differences in effects of Product versus Process GSF on group discussion performance. The Process GSF motivated students to speed up their discussions, resulting in faster speech rates, development of faster turn-taking tactics (such as more cues) and increases in the process measures used as goals (opinions, reasons and questions in this case). However, this came at the sacrifice of other performances, with shorter turns, more repetitions, and more '*individual speech style'* discussions (more 'off-task' turns, and less 'outcome-promoting' turns and clarifications) emerging. On the other hand, the Product GSF motivated students to reach agreed outcomes, resulting in more '*collaborative style'* discussions (more clarifications, reformulations and 'outcome-promoting' turns, as well as less 'off-task' turns), but with no significant improvement in speech rates, turn-taking tactics or process measures. These effects of GSF, as well as potential trade-off of performances shown within the data, are important new findings in research, as they show how the focus of goals within discussions can alter the efforts and behaviors of student during discussions which will affect their longer-term learning. The implications of this for language learning in general are discussed further in the conclusions section.

6.3.6.3 LP/HP performance changes

Across both ProdS and ProcS groups, LPs increased classroom words spoken more significantly over time than HPs. Although group discussions have been described as having many participation-related merits, such as motivating learners to speak, providing more individual speaking time and feedback, and increasing opportunities for language use (Foster, 1998; Long & Porter, 1985), improvements in participation were mainly among LPs (via

increases in words spoken and turns taken). This may have been because of LPs benefiting more from the GSF rubric used, as lower performers have often been found to benefit more from using rubrics than higher performers (Balan, 2012; Black & Wiliam, 1998b). It may have also been due to HPs feeling obligated to help LPs in their groups to speak. The observational data analysis I undertook could not be used to support this claim, but from my own teaching observations, it did appear that some HPs were speaking less than they could because they were supporting LPs with their turns (Appendix C, Part 3). Consideration of how different students within the same groups may improve their participation and resultant amount of L2 practice is needed, regardless of any GSF which may take place. Module One of this PhD summarized key areas related to group set-up which can affect how much students will speak during group discussions, such as the differing proficiency levels among speakers (Appendix A). Further research into the grouping of LPs/HPs in classes, such as having LP-only and HP-only groups, may help to increase participation across time, but cannot be assumed without more classroom-based research.

Among the ProdS, LPs improved their classroom discussion participation (words spoken and turns taken), but not within tests (Table 6.4). During classroom discussions, LPs took on more active participant roles within the discussions in order for their group to reach their decided outcome together. However, during tests, HPs were often seen to 'take control' of the discussions more than LPs, which lead to them having much higher performance measures as individuals compared to LPs within all of the tests (Tables 6.4 - 6.8). As a teacher, I was not surprised to see this, as within any group task which depends on a shared outcome, students are likely to rely on their most capable members in their group to achieve this. This is possibly the reason why HPs took more of the speaking role than LPs when it came to tests, as the group

members probably wanted to achieve the best score possible using the highest speaking abilities within the group. Emergent 'leaders' within group work have been found to take control and help speed up the process of groups making decisions together in tasks (Forsyth, 2000, 2016. p. 255; Hackman & Johnson, 2013, p. 203). This establishment of leaders is a natural part of discussions, and one which students may see as essential for being able to complete tasks and enhance the learning experience (Ehrman & Dörnyei, 1998, p. 154). During tests, I often saw one or two of the four group members take a more 'dominant' leader-like role within their group (who always turned out to be HPs in the data) and the other students in the group spoke a lot less than them. Although this was perhaps an expected establishment of roles within group work, it made it much more difficult for me to assess the 'performance' of the quieter students (LPs). Although the test score was shared between ProdS, based on the group's combined outcome, if a teacher were to assess group members individually in this style of task, it would be very challenging if some members do not speak much. Therefore, an important observation made for the Product GSF used within this study was that it highlighted the difficulty for a teacher to understand individual performance within outcome-focused TBLT group tasks (Long, 2014, pp. 332-334). If teachers are to adopt a Product GSF approach to learning, it is harder to establish performance rubrics which are fair for speakers within the same group than with an individual approach, such as with the Process GSF. This may lead to demotivation from speakers who do most of the talking within discussions, but get the same score as those in their group who may not have spoken at all. The longer-term impact of such 'unfair' scoring for learners could be very negative, especially for students who do most of the speaking, and needs further research.

Among the ProcS, LPs made larger improvements than HPs in participation, fluency and process measures within both classroom practices and tests (Tables 6.4, 6.5 and 6.8). By using feedback on individual performances within tasks, the students who spoke less in discussions at the start of the semester (LPs) improved more than their more 'talkative' counterparts, and the interactions within groups seemed to be more evenly split between speakers by the end of the semester. This was a very different from the tests by ProdS, in which HPs continued to take much more of the speaking role than LPs. ProcS LPs seemed to have benefited more in terms of performance related to the GSF than HPs, which suggests that such individual performance feedback can benefit the more silent students in group discussions, who may also be the students with lower proficiency (Balan, 2012; Black & Wiliam, 1998b).

However, the smaller improvement in performance observed for HPs compared to LPs poses a serious problem with using Process GSF within group discussions. By allowing LPs within groups to speak more and more across time, HPs among the ProcS may be limited in terms of the time they can use to demonstrate increases in their own process-related performances. This was possibly the reason why they improved their performance measures less across time than LPs, and ended up with similar performance measurements to LPs by the end of the semester. As tests were limited to eight-minute discussions within the study, students had to share that time to take as many turns as possible to get their four individual scores. This may have explained why turns during tests became more frequent and shorter across time for ProcS (Section 6.3.6.2), which was especially true for LPs.

In addition to the key findings above, an important observation which I made during the study was that **the process goals may have put students (especially LPs) under more stressful**

conditions within tests than the product goals. During tests, something I noted in my teacher's journal (Appendix C, Part 3) was that when a student was pausing a lot, or speaking at a speech rate which was clearly slower than the others (often LPs), other group members (often HPs) became quite restless and added pressure to the speaker to finish their turn as quickly as possible. They did this using impatient hand gestures, and even seemed to 'cut off' students who may not have finished their turns. Also, some students looked very stressed before and during tests, as well as disappointed in their performance afterwards (often LPs). This was my opinion from watching their facial expressions and overhearing them say things such as 'this is too hard', 'I'm very stressed', and 'That was bad'. This was also shown with slightly less reduction in anxiety reported by ProcS compared to ProdS across time in RQ3 (Table 6.12). This new finding shows that personal performance goals for group discussions are likely to create more stressful learning conditions than goals shared by groups, as suggested in another classroom-based study by Wang (2014). Although there are many apparent overall performance improvements across time for ProcS, but not for ProdS (see Table 6.10 for a clear comparison), teachers need to consider the enjoyment and anxiety of students during their learning. If students are made more accountable for their individual performance within discussions, and are clearly feeling more pressure to perform, the longer-term implications for intrinsic motivation may be counter-productive. RQ3 addresses this more using self-reported data (Section 6.4.6.3).

To sum up, some important differences were found in performance between LPs and HPs. Firstly, across both groups, LPs were found to increase their words spoken more than HPs, suggesting that the use of GSF with TBLT discussions may be better at improving the spoken participation of students who say less than others. Secondly, ProdS LPs only increased their

words spoken in class, but not in tests. This suggests the potential for domination of speech by ProdS HPs if the discussion goals are shared by the group, which may prohibit the spoken participation and learning of LPs. Thirdly, ProcS LPs improved their class and test participation, fluency and process measures more than HPs did across time. The performances by LPs and HPs became quite even by the final class and test, suggesting significant improvements for LPs, but potential limitations to improvements by HPs. Finally, although ProcS LPs showed the most improvement in performance across the study, the stress which their individual goals may have added to their learning is a serious concern. These key points are revisited in the conclusions section to explain their deeper implications for the use of goals within communication courses.

6.3.7 RQ2 key findings summary

This study has provided new empirical data on the actual learning undertaken by low-level learners across time with a TBLT approach (RQ2, Part a). Although improvements were observed for classroom accuracy and fluency during tests across both ProdS and ProcS groups, no significant improvements were seen for classroom participation, repetitions, reformulations or complexity, or for test accuracy or complexity. This casts doubt over the use of TBLT with group discussions as an approach to learning with low-level learners and consideration of additional support (such as pre-task planning stages) may be required.

Another original finding was the difference in overall effect on performance of Product and Process GSF (RQ2, Part b). Observational data showed how product and process-focused goals can influence low-level learner behaviors and performances within group discussions, as well as how ‘trade-off’ between different aspects of performance may occur, depending on the

focus of the goals. This goes beyond the possible trade-off between CAF measures debated by Skehan and Robinson (Section 2.2.2) and suggests further trade-off between some other types of performance (Seijts & Latham, 2001) which teachers should be aware of when selecting goals for discussions. Specifically, process goals resulted in faster speech in classes, fewer pauses in classes and tests, more turns taken with more opinions, reasons and questions within them in tests, as well as faster turn-taking between speakers (using cues such as 'What do you think?'). However, students also developed more *'individual speech style'* discussions, (cutting each other off more, clarifying less, using more individualized words such as 'your' and 'my', taking more off-task turns and less outcome-promoting turns) using shorter turns with more repetitions in them. On the other hand, product goals resulted in more agreed decisions by groups with more *'collaborative style'* discussions (more outcome-promoting and less off-task turns, more time clarifying checks, more reformulating, and using plural first-person pronouns), but no improvements in speech rates, turn-taking speed, or process measures, as with the process goals.

A final important finding was the difference in performances by LPs and HPs across time (RQ2, Part c). LPs improved their participation more than HPs, demonstrating that TBLT, combined with GSF, may be more beneficial for lower performers (as suggested with rubrics by Balan, 2012; Black & Wiliam, 1998b). This was the first classroom-based study to use longitudinal data to show this for group discussions. Furthermore, LPs using Process GSF showed more improvement in the performance measures than those using Product GSF. This original finding has shown that personal, rather than shared performance-based goals can improve the efforts and performance of lower performers (perhaps the more silent students)

within groups. The implications and teaching recommendations for all of the above findings are further discussed in the conclusions section.

6.3.8 Limitations

Three main limitations need consideration. Firstly, there was **non-equivalency in performances between ProdS and ProcS at the start of the study.** ProcS said more words than ProdS in Weeks 2 and 3 (Table 6.4), and spoke at higher speech rates than ProdS in the Week 3 test (Table 6.5). I selected ProdS and ProcS groups with a mixture of LPs and HPs for the analysis (Section 5.4.2), but the fact that the mean levels of performance for ProdS and ProcS were noticeably different in Weeks 2 and 3 questions the reliability of the comparison of the effects of the two types of GSF over time. In future research, the addition of a pre-study test for students could enable the researcher to establish group equivalency.

Secondly, there were some **unclear patterns of changes in performance** (indicated in red in the tables). Some variables increased/decreased with statistical significance between the start and middle of the semester, but by the end of semester returned to a level similar to that at the start. An example of this was the PEFC values within classroom practices, where both ProdS and ProcS exhibited this unclear rising and falling across time (Table 6.7). Also, some of the follow-up paired-sample t-test results (Tables 6.4 - 6.8) showed significance in changes which the ANOVA tests (Table 3) did not. Examples were words per turn for ProcS (Table 6.7) in tests and the errors in tests for ProdS (Table 6.6). Strictly speaking, if an ANOVA test does not reveal significance for a variable (establishing if a significant change in the variable occurred across time), then follow-up t-test results for the same variable (isolating which time period(s) the

significant changes took place in) should not be considered significant. This made it hard for me to consider t-test results significant or not, if the initial ANOVA test did not find any significance (these results are shown in red within Tables 6.4 - 6.8).

Thirdly, I experienced several **problems with coding speech within the discussion transcripts**. It was (a) difficult to understand what the students meant sometimes. For instance, one student described companies as 'black' (Appendix L, Line 5) which seemed to be understood by the other students. Another student described an American landmark as 'Freedom' (Appendix N, Line 50) which I assumed to be the Statue of Liberty, but could not be sure. I made the decision to count unclear instances such as this one as reasons within the discussion, even though another person watching the discussion may have understood them differently. It was also (b) difficult to judge how much a student should say to be awarded a score for one of the process measures. For example, students sometimes used single words to represent an opinion, reason, or question. One group of students exchanged opinions about what subject they thought they should teach during a part-time job using utterances such as 'student experiments', 'chemistry experiment' and 'mathematics' (Appendix L, Lines 17-22). Compared to other turns which included opinions these were very short, but I decided to count them as opinions within the discussion. Moreover, (c) some turns were quite repetitive of others within the same discussion. An example of this can be seen in Appendix N (Lines 20-28) where a student asks another where they want to go for a trip, even though they have already stated that they want to go to the 'Gold Coast'. As an official rule for myself, I decided not to count turns which were repeating the exact same question, reason or opinion of a person before them within a discussion, or use of the question 'why' (as students could just use it in any context without listening to each

other). I made the students aware of these rules near the start of the semester after considering them following the first test. All three of these coding problems could possibly be reduced in the future with the inclusion of other teachers or raters within the coding process. I was unable to find anyone available to help with such an extensive analysis for the study, but hope to do so in the future. By having different raters independently code the same transcripts I hope to use inter-rater correlation data and exchanging of opinions about the GSF scoring system to improve the reliability of the data and better clarify the GSF scoring for students and teachers.

6.4 RQ3: Student self-reported feelings towards the GSF and discussions

6.4.1 Introduction

This section addresses the third RQ, which was ***'(a) How do ProdS and ProcS report feeling about performing in discussions across the semester? and (b) How do they report feeling about the support the two types of GSF provided for their learning (or not)?'*** Student feelings towards undertaking discussions were explored using student surveys at the start, middle and end of the course. In addition, a second and third part to the surveys and interviews were carried out at the beginning and end of the semester regarding student feelings about difficulties with test performance and the usefulness of the GSF provided. I first explain the results from the surveys and interviews. After that, I discuss the possible implications of these findings, as well as the limitations to the analysis.

6.4.2 Discussion feelings survey results

Table 6.12 shows the results for the first part of the survey (Appendix I) regarding student self-reported feelings about undertaking English discussions during the semester. As can be seen, both ProdS and ProcS showed increases between Weeks 3 and 8, and Weeks 8 and 13, for mean reported levels of enjoyment, motivation, confidence, self-perceived discussion and overall English abilities and out-of-class discussion study time. **Regardless of which GSF was undertaken, student self-reported feelings towards discussions became generally more positive across the semester.** Despite this, some initial non-equivalency did exist between the two groups, shown by the different mean scores for measures reported by the two groups in Week 3. ProcS reported a lower mean anxiety level and higher means for the other six variables compared to ProdS. As students were assigned to one of the two types of GSF based on the classes they were in, this was perhaps unavoidable, where one class of students might be more motivated to undertake discussions than another. This is an important limitation to the data and is discussed more later (Section 6.4.8).

One significant difference shown in Table 6.12 between ProdS and ProcS was for self-reported anxiety. **ProdS reported larger decreases in anxiety levels across time in discussions than ProcS.** ProdS reported decreases in anxiety between Week 3 (M = 4.60, SD = 1.16), Week 8 (M = 3.53, SD = 1.21) and Week 13 (M = 3.40, SD = 1.20). However, ProcS actually reported almost exactly the same level of anxiety in Week 3 (M = 3.92, SD = 1.26) and Week 8 (M = 3.93, SD = 1.56) and eventually a decrease by Week 13 (M = 3.55, SD = 1.53). This difference in self-reported responses was perhaps the only clear contrast between ProdS and ProcS revealed by the surveys.

Table 6.12. Student self-reported feelings towards discussions

	Group	Week 3	SD	Week 8	SD	Week 13	SD
Enjoyment	Product	3.55	1.20	4.83	1.06	5.21	1.12
	Process	4.12	1.09	4.74	1.31	5.58	1.17
Motivation	Product	4.19	1.11	5.05	1.00	5.35	0.88
	Process	4.57	0.99	5.16	1.03	5.54	1.16
Confidence	Product	2.08	0.95	4.66	1.22	4.85	0.84
	Process	2.92	1.16	4.49	1.53	5.24	1.12
Anxiety	Product	4.60	1.16	3.53	1.21	3.40	1.20
	Process	3.92	1.26	3.93	1.56	3.55	1.53
Discussion ability	Product	2.74	0.97	4.57	1.04	5.00	0.88
	Process	3.17	0.99	4.63	1.17	5.19	1.06
Overall English ability	Product	2.21	1.14	4.40	0.83	4.85	0.70
	Process	2.70	1.03	4.56	0.97	5.03	0.88
Out-of-class discussion study	Product	1.87	0.94	4.24	0.60	4.45	0.84
	Process	2.17	1.07	4.39	0.79	4.56	0.79

Notes. Week 3 scale: 1=very low, 4=normal, 7=very high; *Weeks 8 and 13 scale: 1=decreased a lot, 4=no change, 7=increased a lot; Product N = 50, Process N = 82*

6.4.3 Test difficulties survey and peer-interview results

Table 6.13 shows a summary of responses for the open-ended surveys/interviews (Part 2 of Appendix I) in Weeks 3 and 13 regarding student perceptions of the most difficult points for performing in group discussion tests. Once the written survey responses were collected and interviews transcribed and checked with the interviewers, I categorized all of the responses from Weeks 3 and 13 regarding difficulties under seven categories and 17 sub-categories. This was done after extensively reviewing all of the written and spoken responses from the data for both weeks. These were then coded and further categorized under seven main topics shown in Table 6.13. By doing so, all of the responses in the surveys and interviews which referred to a difficulty regarding performance in discussion tests were accounted for in the data.

Table 6.13. Initial (W3) and final (W13) self-reported difficulties for discussion tests

PRODUCT W3 DIFFICULTIES	% of students			PROCESS W3 DIFFICULTIES	% of students		
Category	Surveys	Interviews	Total	Category	Surveys	Interviews	Total
L2 Total	**42.00**	**100.00**	**47.17**	**L2 Total**	**51.22**	**100.00**	**58.06**
Vocabulary	26.00	100.00	30.19	Vocabulary	30.49	54.55	33.33
Grammar	10.00	33.33	11.32	Grammar	14.63	36.36	17.20
Listening comprehension	6.00	0.00	5.66	Listening comprehension	6.10	18.18	7.53
Delivery Total	**44.00**	**66.67**	**45.28**	**Delivery Total**	**43.90**	**54.55**	**45.16**
Expressing ideas in English	44.00	66.67	45.28	Expressing ideas in English	43.90	54.55	45.16
Structuring speech	0.00	0.00	0.00	Structuring speech	0.00	0.00	0.00
Time Total	**22.00**	**66.67**	**24.53**	**Time Total**	**17.07**	**54.55**	**21.51**
Lack of preparation time	12.00	33.33	13.21	Lack of preparation time	7.32	18.18	8.60
Short discussion time	10.00	33.33	11.32	Short discussion time	9.76	36.36	12.90
Delivering speech quickly	0.00	0.00	0.00	Delivering speech quickly	0.00	0.00	0.00
Using time effectively for points	0.00	0.00	0.00	Using time effectively for points	0.00	0.00	0.00
Opinions Total	**16.00**	**0.00**	**15.09**	**Opinions Total**	**4.88**	**27.27**	**7.53**
Coming up with things to say	8.00	0.00	7.55	Coming up with things to say	2.44	27.27	5.38
Giving details (reasons/examples)	8.00	0.00	7.55	Giving details (reasons/examples)	2.44	0.00	2.15
Topic Total	**0.00**	**66.67**	**3.77**	**Topic Total**	**3.66**	**9.09**	**4.30**
Difficult topic content	0.00	66.67	3.77	Difficult topic content	3.66	9.09	4.30
Interaction Total	**0.00**	**0.00**	**0.00**	**Interaction Total**	**1.22**	**0.00**	**1.08**
Asking questions	0.00	0.00	0.00	Asking questions	1.22	0.00	1.08
Answering questions	0.00	0.00	0.00	Answering questions	0.00	0.00	0.00
Agreeing/Disagreeing	0.00	0.00	0.00	Agreeing/Disagreeing	0.00	0.00	0.00
Turn-taking	0.00	0.00	0.00	Turn-taking	0.00	0.00	0.00
Outcome Total	**0.00**	**0.00**	**0.00**	**Outcome Total**	**0.00**	**0.00**	**0.00**
Making a group decision	0.00	0.00	0.00	Making a group decision	0.00	0.00	0.00

PRODUCT W13 DIFFICULTIES	% of students			PROCESS W13 DIFFICULTIES	% of students		
Category	Surveys	Interviews	Total	Category	Surveys	Interviews	Total
Delivery Total	**44.00**	**14.29**	**40.35**	**Interaction total**	**31.71**	**18.18**	**30.11**
Expressing ideas in English	40.00	0.00	35.09	Asking questions	23.17	9.09	21.51
Structuring speech	4.00	14.29	5.26	Answering questions	0.00	0.00	0.00
Opinions Total	**40.00**	**14.29**	**36.84**	Agreeing/Disagreeing	3.66	0.00	3.23
Coming up with things to say	8.00	0.00	7.02	Turn-taking	4.88	9.09	5.38
Giving details (reasons/examples)	32.00	14.29	29.82	**Opinions Total**	**30.49**	**0.00**	**26.88**
L2 Total	**22.00**	**28.57**	**22.81**	Coming up with things to say	7.32	0.00	6.45
Vocab	20.00	28.57	21.05	Giving details (reasons/examples)	23.17	0.00	20.43
Grammar	2.00	0.00	1.75	**Time Total**	**25.61**	**36.36**	**26.88**
Listening comprehension	0.00	0.00	0.00	Short preparation time	0.00	0.00	0.00
Time Total	**14.00**	**14.29**	**14.04**	Short discussion time	2.44	9.09	3.23
Short preparation time	0.00	0.00	0.00	Delivering speech quickly	18.29	18.18	18.28
Short discussion time	6.00	14.29	7.02	Using time effectively for points	4.88	9.09	5.38
Delivering speech quickly	6.00	0.00	5.26	**Delivery Total**	**19.51**	**9.09**	**18.28**
Using time effectively for points	2.00	0.00	1.75	Expressing ideas in English	19.51	9.09	18.28
Topic Total	**4.00**	**57.14**	**10.53**	Structuring speech	0.00	0.00	0.00
Difficult topic content	4.00	57.14	10.53	**L2 Total**	**14.63**	**27.27**	**16.13**
Outcome Total	**4.00**	**14.29**	**5.26**	Vocab	9.76	9.09	9.68
Making a group decision	4.00	14.29	5.26	Grammar	4.88	18.18	6.45
Interaction Total	**2.00**	**0.00**	**1.75**	Listening comprehension	0.00	0.00	0.00
Asking questions	2.00	0.00	1.75	**Topic Total**	**7.32**	**36.36**	**10.75**
Answering questions	0.00	0.00	0.00	Difficult topic content	7.32	36.36	10.75
Agreeing/Disagreeing	0.00	0.00	0.00	**Outcome Total**	**0.00**	**0.00**	**0.00**
Turn-taking	0.00	0.00	0.00	Making a group decision	0.00	0.00	0.00

Notes. Product Survey N = 50; *Process Survey N = 82; Product Interview N = 7; Process Interview N = 11*

The list of categories was sorted from most consistently mentioned to least mentioned by students, with most consistently mentioned difficulties at the top of each list of categories in Table 6.13. If the list in the Table was ordered using the sub-category totals, rather than the overall category totals in bold, the order of some sub-categories would change. For instance, the 'structuring speech' sub-category would be at the bottom of the list in Week 3 for both ProdS and ProcS if this was done, but appears near the top in the table. However, I decided to order the list using overall category totals, so that a more generalized discussion could be made using larger percentages, which I saw as more representative of the overall feelings of the students. This may have caused some problems regarding reliability of the data which are discussed later (Section 6.4.8).

6.4.3.1 Initial reported discussion test difficulties

In Week 3, both ProdS and ProcS reported English usage and delivery, as well as time limitations, as main difficulties for performance in discussion tests. Firstly, about half of the students in both groups reported second language related difficulties (mainly vocabulary and grammar problems) as a main difficulty. Comments such as *"My vocabulary is limited and so I find it very hard to speak in discussions"* were common in the open-ended survey questions. Secondly, 'delivering speech' (specifically, expressing their ideas in English) was also reported very often as a problem. Typical survey comments related to this included *"I can almost never say what I want to in discussion using English"* and *"It's too hard for me to explain my thoughts to others in English"*, although many such statements were fairly vague in explaining why. Thirdly, time limitations were also mentioned as a problem (by around 20% of students) by both

ProdS and ProcS, specifically regarding the lack of preparation time and actual discussion time that students said they had to perform in the tests. Students made statements in surveys such as *"We need more time to prepare, so that we can get better scores in the discussions"* and *"The discussions are too short because group members take a long time to speak during their turns"*, highlighting the general feeling of pressure to perform quickly among most students during discussion test time.

Other similarly ranked difficulties, although less significant, were giving opinions (coming up with things to say and giving details), the discussion topic at hand, interactional problems (asking/answering questions, agreeing/disagreeing, and turn-taking), and making group decisions (a form of task outcome). Due to these similarities among ProdS and ProcS regarding perceptions of test difficulties, any changes to these reported feelings across time could be attributed to the type of GSF that students undertook.

6.4.3.2 Similarities in ProdS and ProcS final reported test difficulties

By the end of the semester, some similarities were seen in the changes of reported difficulties within tests by both groups. Firstly, **fewer ProdS and ProcS reported sub-categories within the 'L2' category as difficulties in Week 13 than Week 3.** Difficulties related mainly to English grammar and vocabulary were main focuses of student responses in Week 3 (Section 6.4.3.1), with around half of the students discussing this as a problem for test performance. However, only 22.81% of ProdS and 16.13% of ProcS reported these as difficulties in Week 13. This decrease does not necessarily mean that fewer students considered vocabulary and grammar less of a difficulty by Week 13 (Appendix C, Part 4 explains how I saw many of

them still focusing on vocabulary within their diaries until the end of the semester), but that within the final surveys and interviews students' focus in explaining the difficulties they were encountering had changed. This may have been due to the GSF used, although the lack of evidence in the data to show this is addressed as a potential limitation later on (Section 6.4.8). Secondly, **more ProdS and ProcS mentioned difficulties under the 'opinions' category in Week 13 than in Week 3.** This was especially so for the sub-category of 'giving reasons and examples' which had become more of a focus point in student responses. Common statements by students included *"I still need time to practise giving reasons and examples. I'm not very good at explaining my opinions to others"* and *"It takes me a long time to think of reasons to explain my opinion".* Thirdly, **more ProdS and ProcS reported the 'discussion topic' as a difficulty in Week 13 than in Week 3, especially within interviews.** Students from both groups made comments such as *"I need to know the topic a long time before the discussion. Some topics need some research, as I don't know what to say on them"* and *"Some topics are too hard. How much I can speak depends on the topic we are given".*

In summary, regardless of what kind of GSF students had undertaken across the semester, reporting of difficulties in tests had shifted away from a heavy focus on language-related problems in Week 3 (mainly grammar and vocabulary) and more towards being able to give reasons and examples which depended more on the topic by Week 13.

6.4.3.3 Differences between ProdS and ProcS final reported test difficulties

There were also some differences between the two groups regarding changes in perceived difficulties within tests across the semester. **ProdS reported sub-categories under the**

'delivery' category (especially 'expressing ideas in English') as their main difficulties within tests in Week 13, but ProcS did not. ProdS made vague comments such as *"Saying what I want is hard if I have to say it in English"* and *"The conversation becomes very slow because no one can say what they want to in English sometimes"*. However, although 45.16% of ProcS reported this difficulty in Week 3, only 18.28% of them mentioned it in Week 13.

Meanwhile, **more ProcS mentioned difficulties under the 'interaction' category in Week 13 (30.11%) than in Week 3 (1.08%).** This had become the most commonly reported difficulty in tests by ProcS, with a heavy focus on problems with asking questions (21.51%), agreeing/disagreeing (3.23%) and turn-taking (5.38%). This was emphasized by ProcS with comments in surveys such as *"It's hard to keep asking questions with new forms such as 'how much', because other group members have already asked them" and "I need to practise this more because it keeps the discussion flowing. It sometimes becomes silent because I can't ask a question quickly enough"*. In contrast, only 1.75% of ProdS reported these interaction-related sub-categories as difficulties in Week 13.

Also, **difficulties related to the discussion 'outcome' (specifically, 'making a group decision') were reported by some ProdS (5%) in Week 13, but not by any ProcS**. An example of a statement within the interviews was *"Sometimes it is too hard to make a choice together because we all have different opinions and do not agree".* Although only 5% of ProdS mentioned the discussion outcome as a problem, their greater focus on group progress, compared to individual progress by ProcS, was more evident within the language used in responses. **ProdS used the first person plural pronoun 'we' or possessive adjective 'our' to describe problems, while ProcS often used first person references such as the personal pronoun 'I'**

and the possessive adjective 'my'. This lent support to the observational data findings in RQ2 (Section 6.3.4.4.) that there was more of an orientation to group performance by ProdS and to individual performance by ProcS, which was the intended effects of the two different types of GSF used. I wanted to further analyze and discuss the use of language in interviews, but was unable to do so in this thesis due to limited space.

Finally, **by Week 13, difficulties related to 'time' (especially with regards to 'delivering speech quickly') were reported more by ProcS than ProdS.** 18.28% of ProcS reported this as a difficulty in Week 13, but none of them mentioned it in Week 3. Comments which demonstrated a belief in speed of articulation being a problem for the students included *"I cannot reply quickly enough to the group members. I need more time to think about what to say"* and *"I get frustrated because I cannot say what I want to quickly enough within the time we have".* In contrast, fewer ProdS mentioned 'time' as a problem for performance in Week 13 (14.04%) than in Week 3 (24.53%).

6.4.4 GSF survey and peer-interview results

The following section summarizes the results from survey and peer-interviews regarding the use of Product and Process GSF across the semester. A discussion and possible implications of these results are given later (Sections 6.4.6.2 and 6.4.6.3).

6.4.4.1 Overall reported feelings about the GSF and performance

A six-point Likert scale survey given to the students at the end of the semester addressed the general feelings students had towards the sheets and diaries used (Part 3 of Appendix I). Tables 6.14 and 6.15 show a summary of the results.

Table 6.14. Week 13 student self-reported usefulness of sheet/diary

	PRODUCT MEAN	SD	PROCESS MEAN	SD
Across the whole course, how helpful was the GROUP DISCUSSION SHEET for improving your performance in discussions?	4.50	0.97	4.74	0.88
Across the whole course, how helpful was your DIARY for improving your performance in discussions?	4.39	0.90	4.59	0.93

Notes. Scale: 1 = not at all helpful, 6 = very helpful; Product N = 62; *Process N = 93*

Table 6.15. Week 13 student self-reported future usage preferences for sheet/diary

	PRODUCT		PROCESS	
	Yes	No	Yes	No
Would you want to use the SHEET again in the future for classroom discussions?	54	8	80	13
	Yes	No	Yes	No
Would you want to use the DIARY again in the future for classroom discussions?	51	11	85	8

Notes. Product N = 62; *Process N = 93*

Both ProdS and ProcS generally reported that their sheets (Product M = 4.50, Process M = 4.74) and diaries (Product M = 4.39, Process M = 4.59) were helpful for improving performance and that most of them wished to use both the sheets (87% of ProdS and 86% of ProcS) and diaries (82% of ProdS and 91% of ProcS) again in the future. In addition to this positive finding for GSF, there was no clear difference in opinions between ProdS and ProcS regarding usefulness of the sheets or diaries. This may have been because both types were viewed as positive additions to the learning, and/or that the students did not want to offend the person who had created and encouraged the use of the diaries and sheets for learning

(myself) and so negative opinions were less commonly reported. The next section adds to the survey data here by using more precise reasons given by students.

6.4.4.2 Reported feelings about the GSF sheets

Following the closed-ended questions about how useful students found the sheets and diaries for improving performance (Section 6.4.4.1), open-ended survey questions and interviews were used to gather more specific data about why the sheets and diaries were perceived as useful or not by the students. I categorized the responses regarding the discussion sheets under eight categories and 22 sub-categories (Table 6.16). All of the survey and interview responses which I judged to refer to a positive or negative (labelled as a 'problem' in the table) point regarding the discussion sheets were coded. These codes were then used to create the categories, which were then sorted from most consistently mentioned to least mentioned, with most consistently mentioned categories (based on percentage of students) at the top of the table.

Table 6.16. Final (W13) self-reported helpfulness of discussion sheet

PRODUCT GROUP	% of students			PROCESS GROUP	% of students		
Category	Surveys	Interviews	Total	Category	Surveys	Interviews	Total
Structure Total	**28.00**	**28.57**	**28.07**	**Clarity of Performance Total**	**63.41**	**81.82**	**65.59**
Helped us organize our discussion	8.00	28.57	10.53	Could see our strong/weak points	40.24	45.45	40.86
Good for summarizing content of discussion	14.00	0.00	12.28	Easy to see my performance level each time	15.85	27.27	17.20
Helped us get many ideas out	6.00	0.00	5.26	Can see our score for each category	7.32	9.09	7.53
Reviewing Total	**26.00**	**42.86**	**28.07**	**Clarity of Focus Total**	**24.39**	**27.27**	**24.73**
Can review the content of the discussion again	26.00	42.86	28.07	Helped me know what to say	23.17	27.27	23.66
Clarity of Performance Total	**24.00**	**28.57**	**24.56**	Made me consider my sentence structure	1.22	0.00	1.08
Could see our strong/weak points	18.00	14.29	17.54	Showed how to do well in tests	0.00	0.00	0.00
Easy to see my performance level each time	4.00	14.29	5.26	**Sheet Usage Problems Total**	**10.98**	**36.36**	**13.98**
Can see our score for each category	2.00	0.00	1.75	Do not think paper was necessary	2.44	9.09	3.23
Clarity of Focus Total	**18.00**	**14.29**	**17.54**	Forgot to add circles to sheet during discussion	2.44	0.00	2.15
Helped me know what to say	14.00	14.29	14.04	Writing scores quickly on sheet during discussion difficult	4.88	27.27	7.53
Made me consider my sentence structure	0.00	0.00	0.00	Boring to fill out each time	0.00	0.00	0.00
Showed how to do well in tests	4.00	0.00	3.51	Didn't know when to score things (confusing	1.22	0.00	1.08
Transfer Problems Total	**6.00**	**0.00**	**5.26**	**Motivation Total**	**10.98**	**18.18**	**11.83**
Never reviewed what was written again	2.00	0.00	1.75	Motivated me to aim for new targets	8.54	18.18	9.68
Not enough time to review it	2.00	0.00	1.75	Made me think of a lot of questions for points	2.44	0.00	2.15
Scores depended too much on the topic	2.00	0.00	1.75	**Transfer Problems Total**	**0.00**	**0.00**	**0.00**
Hard to remember what you said for one sheet for	0.00	0.00	0.00	Never reviewed what was written again	0.00	0.00	0.00
Transfer Total	**0.00**	**14.29**	**1.75**	Not enough time to review it	0.00	0.00	0.00
Good for next time	0.00	14.29	1.75	Scores depended too much on the topic	0.00	0.00	0.00
Sheet Usage Problems Total	**2.00**	**0.00**	**1.75**	Hard to remember what you said for one sheet for the next time	0.00	0.00	0.00
Do not think paper was necessary	0.00	0.00	0.00	**Structure Total**	**0.00**	**0.00**	**0.00**
Forgot to add circles to sheet during discussion	0.00	0.00	0.00	Helped us organize our discussion	0.00	0.00	0.00
Writing scores quickly on sheet during discussion difficult	0.00	0.00	0.00	Good for summarizing content of discussion	0.00	0.00	0.00
Boring to fill out each time	2.00	0.00	1.75	Helped us get many ideas out	0.00	0.00	0.00
Didn't know when to score things (confusing	0.00	0.00	0.00	**Reviewing Total**	**0.00**	**0.00**	**0.00**
Motivation Total	**0.00**	**0.00**	**0.00**	Can review the content of the discussion again	0.00	0.00	0.00
Motivated me to aim for new targets	0.00	0.00	0.00	**Transfer Total**	**0.00**	**0.00**	**0.00**
Made me think of a lot of questions for points	0.00	0.00	0.00	Good for next time	0.00	0.00	0.00

Notes. Product Survey N = 50; Process Survey N = 82; Product Interview N = 7; Process Interview N = 11

Two similar points were reported about the usefulness of the Product and Process sheets. Firstly, **the GSF sheets were often reported to be helpful for clarifying performance by ProdS (24.56%) and especially ProcS (65.69%), mainly with regards to seeing your own strong and weak points.** Statements by ProdS showing this belief included *"The sheet helped because we could see what we were good and bad at each time"* and *"We could understand what we did not talk about enough in the discussions"*, with similar comments from ProcS, such as *"The sheet shows me which things I'm not doing well at, such as disagreeing with others enough"* and *"If I forget to ask enough questions in the discussion the sheet helps show me that"*. Secondly, **the GSF sheets were reported to be useful for clarifying what to focus on during discussions by ProdS (17.54%) and ProcS (24.73%), especially by helping them 'know what to say'.** Common responses on this included *"Before we used the sheet, I could not think of things to say. The sheet helped me a lot with this"*. Appendix P shows more example comments made by ProcS on both of these points above (coded as A and D in the responses). The fact that the students reported both of the GSF sheets to help with clarifying focus and performance for discussions suggests an advantage of using such an approach in learning. This is addressed more in the discussion (Section 6.4.6.2).

There were also some differences in the self-reported data with regards to the Product and Process sheets. On the whole, **the Product sheet was reported to help with structuring and reviewing discussions, but not with transferring learning from one discussion to the next.** Firstly, the Product sheet was reported to help students 'structure' what they wanted to say (28.07%), especially in terms of helping them organize the discussion and summary for its content. This was explained with statements such as *"The sheet helped structure our ideas*

together, so it was useful for helping us speak a lot" and *"It helped us think about how to do the discussion and organize what we said each week"*. Secondly, ProdS reported that the Product sheet helped them 'review' their discussion content well after discussions had been completed (28.07%). An example comment by a student was *"The sheet was good for reviewing what we had said in the group and remembering it again"*. Thirdly, a problem reported (but by only 5.26%) was that it was difficult to transfer learning from one discussion to the next with the Product sheet. Comments by ProdS which demonstrated this were *"We never reviewed what was written again"* and *"There wasn't enough time to review things"*. Although these points were made, they were individual comments by just two students, and so I did not consider them an important focus of discussion in this thesis compared to benefits discussed by more students. Therefore, I judged the discussion sheet used by ProdS to not have any significantly reported problems for use with classroom discussions. More discussion about the potential implications for learning for the three points above about the Product sheet comes later (Section 6.4.6.3).

Another reported difference between the GSF sheets was that **the Process sheet was reported as being more motivating more often than the Product sheet, but also more often as being difficult to use.** Firstly, 11.83% of ProcS reported that they were more motivated to perform in discussions because they had used the Process sheet. Statements showing this belief included *"It was much better when we used the sheet because we were all aiming for goals which were motivating us"* and *"I felt motivated when I was trying to get a higher score each time"*. In contrast, none of the ProdS mentioned benefits directly related to motivation, such as the desire to get higher scores each week, during either the surveys or interviews. Secondly, unlike ProdS, 13.98% of ProcS reported problems they experienced with regards to using the

discussion sheet. Comments included *"The sheet was useful, but too hard to use during the discussion. The conversation was too quick to give points on"* and *"I stopped using it after a few weeks. It was stopping me from focusing on what was said, so I just focused on using it after the discussion".* More specific details of what the students meant by these comments were not available which was a limitation of the data (Section 6.4.8). However, it appears that one concern with the use of an individual discussion performance feedback sheet is that the workload put on students during task time (monitoring and recording each individual spoken act related to the assessment used as with ProcS) is much higher compared to when students assess the joint outcome of the group as a whole (as with ProdS). Even though both sheets were used during and after the discussions, it appears that more consideration is needed with regards to the challenge of gathering process performance feedback for students within discussions with the use of sheets. This is addressed more in the discussion (Section 6.4.6.3).

6.4.4.3 Reported feelings about the GSF diaries

The open-ended survey questions and interview responses regarding the perceived usefulness of the diaries used by both groups were also analyzed, coded and categorized. Table 6.17 shows the overall summary, with the student responses put under 10 categories and 23 sub-categories which emerged from my analysis.

Table 6.17. Final (W13) self-reported helpfulness of discussion diary

PRODUCT GROUP	% of students			PROCESS GROUP	% of students		
Category	Surveys	Interviews	Total	Category	Surveys	Interviews	Total
Clarity of Performance Total	**26.00**	**14.29**	**24.56**	**Clarity of Performance Total**	**34.15**	**18.18**	**32.26**
Could see our weak points	10.00	0.00	8.77	Could see our weak points	14.63	9.09	13.98
Easy to see my performance level each time	14.00	14.29	14.04	Easy to see my performance level each time	9.76	0.00	8.60
I could understand scores easily with graphs	2.00	0.00	1.75	I could understand scores easily with graphs	9.76	9.09	9.68
Reviewing Total	**24.00**	**14.29**	**22.81**	**Clarity of Progress Total**	**26.83**	**27.27**	**26.88**
Ability to review discussions helped improve areas	6.00	0.00	5.26	Useful to see changes in my scores across weeks	26.83	27.27	26.88
Helped improve my grammar with written review	18.00	14.29	17.54	**Reviewing Total**	**17.07**	**45.45**	**20.43**
Motivation Total	**12.00**	**42.86**	**15.79**	Ability to review discussions helped improve areas	15.85	27.27	17.20
Goals were motivating to do better each time	8.00	14.29	8.77	Helped improve my grammar with written review	1.22	18.18	3.23
Seeing improvements made me happy	2.00	28.57	5.26	**Motivation Total**	**18.29**	**9.09**	**17.20**
Motivating when you reached a goal	2.00	0.00	1.75	Goals were motivating to do better each time	8.54	9.09	8.60
Clarity of Focus Total	**16.00**	**14.29**	**15.79**	Seeing improvements made me happy	9.76	0.00	8.60
Able to practise what to say next time	0.00	0.00	0.00	Motivating when you reached a goal	0.00	0.00	0.00
Could see goals clearly	6.00	14.29	7.02	**Clarity of Focus Total**	**8.54**	**18.18**	**9.68**
Helped me focus on needed points	6.00	0.00	5.26	Able to practise what to say next time	6.10	18.18	7.53
We could set goals	4.00	0.00	3.51	Could see goals clearly	2.44	0.00	2.15
L2 Improvements Total	**6.00**	**28.57**	**8.77**	Helped me focus on needed points	0.00	0.00	0.00
Could correct English mistakes each time	4.00	0.00	3.51	We could set goals	0.00	0.00	0.00
Improved our English writing skills	2.00	28.57	5.26	**Diary Use Problems Total**	**4.88**	**9.09**	**5.38**
Reviewing Problems Total	**0.00**	**42.86**	**5.26**	I was not focusing on the diary's detailed points	1.22	0.00	1.08
Not helpful writing things down	0.00	14.29	1.75	Tough writing in the diary every week	1.22	9.09	2.15
Did not use English I reviewed in future discussions	0.00	28.57	3.51	I wanted to return to the discussion again when reviewing mistakes	1.22	0.00	1.08
Clarity of Progress Problems Total	**0.00**	**14.29**	**1.75**	Scoring system hard to understand	1.22	0.00	1.08
Cannot measure level across time with changing	0.00	14.29	1.75	**Progress Problems Total**	**2.44**	**0.00**	**2.15**
Clarity of Progress Total	**0.00**	**0.00**	**0.00**	It did not increase my knowledge of English	2.44	0.00	2.15
Useful to see changes in my scores across weeks	0.00	0.00	0.00	**Reviewing Problems Total**	**0.00**	**0.00**	**0.00**
Progress Problems Total	**0.00**	**0.00**	**0.00**	Not helpful writing things down	0.00	0.00	0.00
It did not increase my knowledge of English	0.00	0.00	0.00	Did not use English I reviewed in future	0.00	0.00	0.00
Diary Use Problems Total	**0.00**	**0.00**	**0.00**	**Clarity of Progress Problems Total**	**0.00**	**0.00**	**0.00**
I was not focusing on the diary's detailed points	0.00	0.00	0.00	Cannot measure level across time with changing	0.00	0.00	0.00
Tough writing in the diary every week	0.00	0.00	0.00	**L2 Improvements Total**	**0.00**	**0.00**	**0.00**
I wanted to return to the discussion again when reviewing mistakes	0.00	0.00	0.00	Could correct English mistakes each time	0.00	0.00	0.00
Scoring system hard to understand	0.00	0.00	0.00	Improved our English writing skills	0.00	0.00	0.00

Notes. Product Survey N = 50; Process Survey N = 82; Product Interview N = 7; Process Interview N = 11

Four important similarities emerged from the self-reported data about the diaries. **Both the Product and Process diaries were reported to be motivating, as well as useful for clarifying performance, clarifying focus and reviewing English used in discussions.** The most commonly reported benefit was that the diaries provided 'clarity of performance' for both ProdS (24.56%) and ProcS (32.36%). Comments connected to this by ProdS included "*We can see our score each time, so I think we can speak more by knowing this"* and *"It helped when we could see with our own eyes how we were doing each week"*. ProcS made similar comments and also suggested that the use of performance graphs within the diary were helpful also (9.68%). These included *"I could see my scores clearly with the graphs in the diary and knew how well I was doing"* and *"I could see what I was not good at each time with the graphs and it helped me focus more on them next time"*. The second most reported benefit was for 'reviewing'. ProdS (22.81%) and ProcS (20.43%) made statements which suggested a belief that being able to review discussions again with their diary helped improve their performance. One ProdS said in their survey that "*Reviewing my mistakes each week after the discussion was very useful and helped me improve each week".* An example ProcS interview comment was *"I could review my weak areas with no time pressure with the diary and this helped me improve for the next time"*. The third most reported benefit was that the diaries were 'motivating' for ProdS (15.79%) and ProcS (17.20%). For ProdS, this was said to be especially so because of having goals to motivate them to do better and seeing improvements in performance across time. Example statements in the interviews related to this included *"It was motivating to see improvements in different areas of the diary sections"*, *"It was nice to see how you were getting better. It motivated me to try harder"*, and *"The scores I could see in my diary motivated me to try and speak more each*

week". ProcS also made statements referring to the motivation which having and reaching goals provided. Such survey comments included *"The diary helped me aim for targets which I thought was very motivating"* and *"I like seeing the green bars go high when I reach a goal. It makes me try harder each time"*. A final commonality between the groups was that the diaries were reported to improve 'clarity of focus' for both ProdS (15.79%) and ProcS (9.68%), especially because of the use of goals. One ProdS stated in their interview that *"The diary helps us see our goals and makes it easy to focus on them"* and an example survey statement by a ProcS was *"The diary was good for setting goals for next time which I could aim for"*. These four overall similarities in opinions about the GSF diaries by both ProdS and ProcS are significant findings, as they suggest that the use of an electronic diary to record and track performance goals, scores and language focuses across time can create a sense of improved clarity of focus and performance, motivation and improvement in performance among students. Possible implications this may present for learning are discussed later (Section 6.4.6.2).

Two significant differences between the reported usefulness of the two diaries also arose in the data. Firstly, **only the Product diary was reported to help with improving English, but also to be difficult to use to review English.** ProdS stated that the diary helped with 'L2 improvements' (8.77%) with example interview responses such as *"We could notice our mistakes afterwards, so we could focus on improving for next time. This was very helpful and interesting"*. However, three of the seven interviewed ProdS reported problems with 'reviewing' with the diary. Two of the ProdS said *"I did not know what to write down or review and was not sure about how to improve with the diary" and "I only wrote a few simple sentences each time in the diary. I wanted to improve my English more and this didn't help me".* Secondly, **only the**

Process diary was reported to help clarify progress of performance, but also to be difficult to understand how to use. The Process diary was believed to provide 'clarity of progress' by ProcS (26.88%), but this was never reported in the data for the Product diary. Statements made in the interviews by ProcS included *"The diary helped me compare my scores across time and see what I got better or worse at"* and *"The diary was good for seeing how I got better or worse each week"*. However, problems with 'diary usage' were mentioned in both surveys and interviews (5.38%) for the Process diary. One student stated in their survey that *"We spent too much time focusing on English corrections afterwards. We needed more time to practise speaking with each other"*. In summary, an important difference was that many of the problems discussed for the ProdS diary related to not being able to improve at language use, while many for the ProcS diary were more related to focusing on language improvements as a distraction from the learning. This significant difference in focus within learning due to the two types of GSF used is discussed again later (Section 6.4.6.3).

6.4.5 RQ3 results summary

Across the semester, most of the students reported increases in enjoyment, motivation, confidence, discussion ability and overall English ability and decreases in anxiety during discussions (Table 6.12). This may be very expected with GSF (for reasons explained in Sections 4.2 - 4.4), but tells us little about the differences between the two GSF approaches used in the study. The only significant difference between the groups in Table 6.12 was the fact that ProdS reported a greater reduction in anxiety over time than ProcS. Students focusing on individual process-focused performance goals may, therefore, feel more anxious whilst undergoing

discussions than those who focus on overall shared group goals. This is revisited as an important point in the discussion (Section 6.4.6.3).

Table 6.18 below gives a summary of the difficulties reported by the two groups within discussion tests by the end of the semester. Three of the main challenges reported by both ProdS and ProcS for discussions was the ability to express their own ideas in English, give reasons and examples, and the discussion test topic. However, ProdS were also concerned with their English vocabulary, while ProcS reported interacting with others in discourse (via asking questions) and delivering speech quickly enough as problems. Possible implications of these reported difficulties are discussed in the next section.

Table 6.18. Summary of main final reported difficulties for discussion tests

Group	Main reported difficulties (percentage of students reporting it)
ProdS	Expressing ideas in English (35.09%), giving reasons/examples (29.82%), English vocabulary (21.05%) and the discussion topic (10.53%)
ProcS	Asking questions (21.51%), giving reasons/examples (20.43%), delivering speech quickly (18.28%), expressing ideas in English (18.28%) and the discussion topic (10.75%).

In addition, the sheets and diaries used by both ProdS and ProcS were generally viewed as useful for helping with performance and most of the students wished to use them again for future classroom discussions (Tables 6.14 and 6.15). Table 6.19 gives a summary of the student perceptions of the sheets and diaries used to implement the Product and Process GSF (Tables 6.16 and 6.17). Differences between ProdS and ProcS in the table are discussed in the next section.

Table 6.19. Summary of student perceptions of benefits and problems with discussion sheets and diaries

Group	GSF instrument	Reported benefits summary	Reported problems summary
ProdS	Sheet	Provided clarity of focus and performance, help with structuring discussions, and useful for reviewing English used in discussions	Difficult to transfer learning from the sheet to next discussion
	Diary	Provided clarity of focus and performance, motivated students to perform better, useful for reviewing English used in discussions, and helped improve spoken English grammar	Uncertainty in what English to review and focus on improving in the post-discussion practice section
ProcS	Sheet	Provided clarity of focus and performance, and motivated students to perform better	Knowing how to use it and filling it out quickly enough
	Diary	Provided clarity of focus and performance (especially through the graphs), motivated students to perform better, useful for reviewing English used in discussions, and made progress in performance clearer	Knowing how to use it to help improve at discussions

6.4.6 RQ3 discussion

I now discuss the results for the student surveys and peer-interviews in more detail, with all the key findings for RQ3 shown in bold. I start by addressing the student feelings about difficulties undertaking discussions across the semester and make possible connections between those difficulties and performance problems observed (RQ2). Next, I discuss the self-reported data about the usefulness of the GSF and how it may be connected to observed performance changes across the semester (RQ2). After that, I discuss specific differences in the reporting of the usefulness of the Product and Process GSF used and highlight potential connections between the two types and changes in observed performance across time (RQ2). Finally, I give an overview of the limitations to the data analysis.

6.4.6.1 Reported feelings about discussion performance

Across the study, both ProdS and ProcS reported increases in enjoyment, motivation and confidence within discussions, as well as increases in self-perceived discussion and overall English abilities and out-of-class discussion study time (Table 6.12). Thus, this study has shown that **low-level students' self-perceived enjoyment, efforts and performance within group discussions will improve across time with a TBLT approach supported by GSF.** It was my belief that the students were basing these increased ratings of their efforts and performance on the scores displayed within their GSF sheets and diaries, as that is how their performance was being determined. These positive effects on learning are expected when using clear performance rubrics (Andrade et al., 2009; Panadero et al., 2012) with feedback (Chappuis, 2005; Leung, 2005a; Stiggins et al., 2004) and this study has shown these effects to also be possible with TBLT group discussions. However, this is based on self-reported data only and the two separate effects of a TBLT approach and the GSF on these self-reported feelings (Table 6.12) cannot be determined. This lack of clarity of separate effects of TBLT and GSF within the observational and self-reported data is discussed as an overall limitation to the thesis in the conclusions.

Another important finding was that by Week 13, both ProdS and ProcS mentioned tests difficulties related to giving opinions and the discussion topic more, and problems with English vocabulary and grammar less than in Week 3 (Section 6.4.3.2). There are several possible interpretations of this data. Firstly, **the use of TBLT group discussions and the GSF used in the study may make students value meaning-focused performance more than form-focused performance across time**. This is the intended influence of TBLT and may mean that students concern themselves more with negotiation of meaning during discussions, rather than specific

language forms, 'notice' their language gaps as a result (Batstone, 1996, p. 273) and therefore improve their performance over time (Ellis, 2003; Littlewood; 2004; Long, 2014; Robinson 2007, 2011; Willis & Willis, 2008). This may have aided the improvements seen in accuracy during classroom discussions (which followed a TBLT pre-task, task and post-task approach) by both ProdS and ProcS (Section 6.3.6.1) although this relationship is not shown with any data in this study.

Secondly, **the use of TBLT group discussions and GSF may require additional support, such as a pre-test planning stage, to help low-level students prepare their opinions.** This may have been an ongoing problem for the discussion tests, as considering and preparing to exchange opinions for discussion topics can take time, but the tests did not allow the ten-minute planning stage which the classroom practices did. I discussed in the literature review how pre-task planning can generally improve oral performance (Ellis, 2009; Guará-Tavares, 2011, 2013, 2016; Javad Ahmadian et al., 2015). Also, the results in Module Two of this PhD showed that having a ten-minute rehearsal planning stage, for the same group discussion set up used in this thesis (in terms of group size, topics and discussion length), can improve how much students speak, as well as how many reasons they will give (Appendix B). The lack of such rehearsal for the students for tests may have accounted for their inability to improve their accuracy for example, as they did improve accuracy (errors in speech) in classroom practices across time when they were allowed a planning-stage (Section 6.3.6.1).

Thirdly, **group discussion topics used by low-level students may need to be more familiar and easier to discuss than those in the study in order to support transfer of learning.** In the survey and interviews in Week 13, ProdS and ProcS reported that some of the

topics were too hard to discuss and that their performance depended heavily on what the topic was. Although I chose topics (Appendix J) which I believed the students had adequate knowledge on, did not threaten face, and were simple to discuss, some groups perceived some topics to be too difficult for them to discuss. I sometimes heard comments from students such as *"that topic was too hard"* and perhaps the fact that the topic was new each time made it harder for the students to transfer the improvements they were seeing in classroom performance (such as accuracy for example) to the tests. Although transfer of learning from task-to-task is expected for students when the goals, conditions and set-up are the same each time (Barnett & Ceci, 2002; Benson, 2016; Blume et al., 2010), some topics may have created a high cognitive load for students and prevented such transfer between tasks (Skehan,1998, p. 99). However, it is also possible that the students were unaware or unwilling to admit in the self-reported data that problems with their test performance were due to factors other than the topic and that they were blaming it for whatever test scores they got. More detail about why the students felt the topics were too difficult was not revealed in the survey or interview data and would have been very useful for better understanding this potential problem.

6.4.6.2 Similarities in the reported effects of Product and Process GSF

Similar things were reported by both groups with regards to the usefulness of the GSF to help with discussion performance. On the whole, all of the Product and Process GSF sheets and diaries were reported to be helpful with improving performance (Table 6.19). More specifically, using the GSF sheets and diaries was reported to provide clarity of focus and performance, as well as with reviewing English. Therefore, this study has shown that **the use of GSF with group**

discussions can help students feel that they understand their performance better and how to focus on what they need to improve at. Although the use of goals and feedback were discussed earlier to be believed to support learning in this way (Sections 4.2 and 4.3), this study has shown this to also be applicable to *group discussion tasks*. The list of Process GSF which I created in RQ1, as well as the lists ProdS used to assess outcome, were short and concise and so helped create this focus. This positive finding means that the addition of feedback like this for discussions can be implemented as a helpful addition to learning, as students will find it easier to know how to perform well and this will increase their chances of improving their performance across time (Chappuis, 2005; Leung, 2005a; Stiggins et al., 2004). However, exactly how this reported improvement in focus helps, and what performance the students were referring to in the data is unclear. I assumed that they were referring to the GSF scores they used, but could not prove this. There was also no clear connection between these reports and changes in performance in the RQ2 data either, suggesting that reported improvements due to the GSF were more motivational in nature, rather than performance-based.

This study has also shown how **electronic performance diaries which support group discussion GSF can be used to motivate students to perform**. This was reported within both groups, and with specific reference among ProcS to graphs showing performance feedback being helpful. The application of classroom-based technology to assist formative-assessment and feedback on tasks has been shown to motivate students (Irving, 2015; Maier, Wolf & Randler, 2016) which has been shown to also be the case for group discussions in this study. This is an important finding for teachers who do not have time to monitor and analyze performance for every student themselves, as the more frequent and immediate feedback which such technology

provides would be expected to improve performance over time (Black & Wiliam, 2009; Davison, 2004; Huhta, 2008; Sadler, 1998; Tunstall & Gipps, 1996; Wiliam, 2018).

6.4.6.3 Differences in the reported effects of Product and Process GSF

Three important differences were seen in the self-reported data by the Product and Process GSF. Firstly, **the Process GSF seemed to focus students more on interactional performance than the Product GSF** (such as asking questions, agreeing and disagreeing). In addition to a greater improvement in the interactional performance measures (Table 6.8) and more '*individual speech style*' discussions than ProdS (Section 6.3.6.2), the changes in self-reported data by ProcS between Week 3 and Week 13 (Table 6.18) suggested that they began to see a discussion task more as one which involves interacting in this way. This was also shown in the observational data, as ProcS were seen to use language indicating more of a concern for individual performance within interactions (Section 6.3.7). This was perhaps unsurprising, as the GSF list they were given to use (Table 6.2) was intended to focus them on such goals to demonstrate performance. Therefore, process-related goals might be more appropriate for language learning with group discussions than product-related goals, as the higher number of interactions which they result in are believed to be important for SLA (Firth & Wagner, 2007; Gass & Mackey, 2007; Gass & Varonis, 1994; Hatch, 1978; Long, 1996; Pica, 1994; Sun, 2011). However, connections between interactive language use and SLA have not been firmly proven in research (Keck et al., 2006) and this is discussed more in the conclusions section.

Secondly, **the Process GSF and assessment may have made students feel under more pressure to perform than the Product GSF**, which was discussed in Chapter 4 as a potential

problem with the use of individual performance goals (Section 4.2.2). Within tests, ProcS said 'delivering speech quickly' was a difficulty (Table 6.18), suggesting they may have felt extra pressure to get scores as quickly as possible. This extra pressure to perform for ProcS, but perhaps not ProdS, was also discussed in RQ2, where there was a noticeable effort by the students to pause less, speak more quickly and take more turns within the group across time with more '*individual speech style*' discussions (Section 6.3.7). Despite this, ProcS did say that the sheet (used during classroom discussions) motivated them to perform better and rises in the number of turns taken during classes and tests seen within RQ2 (Table 6.4) may be connected to this motivation. Therefore, with Process GSF, it is very important to consider the balance between pressure on students to perform and improvements in process-measures. It may be desirable to 'push' students to participate and interact more within discussions using such personalized accountability for performance with Process GSF, to ensure that they practise using the language as much as possible. This study has shown how such accountability with personal goals for group discussions can create a sense of pressure and resultant anxiety among students (ProcS reported higher anxiety levels than ProdS by Week 13 in Table 6.12), which may have longer-term implications for student motivation. Making performance scores 'public' in classes, or linked to interpersonal or competitive goals between members, have been shown to create problems with motivation and engagement (Covington, 2000; Latham et al., 2016; Ryan & Patrick, 2001; Ryan & Pintrich, 1997) and appear to also be a danger for group discussions.

Thirdly, **the cognitive load placed on students during discussions by the Process GSF was higher than by the Product GSF**. The in-task self-regulated assessment and recording of personal performance (ProcS) was clearly more demanding than the post-task overall group

performance style (ProdS). The survey/interview responses suggest that ProcS had difficulties filling out their sheet quickly enough compared to ProdS (Table 6.19). I also observed ProcS forgetting to use the sheet, or even intentionally putting it aside at points, when they seemed very focused on their discussions (Appendix C, Part 3). Even with the short list of process goals created in this study (Table 6.2) the workload may be too high and distracting from the task itself. Other research has suggested that students have limited attention to what they can focus on during task time (Skehan, 1996, 1998, p. 97) and that some goals may distract them away from what will help them learn or perform well (Seijts & Latham, 2001). This study confirms these beliefs and builds on them by using longitudinal data (RQ2 and RQ3) to show that the addition of self-regulated performance tracking during group discussions, while perhaps being motivating, can also be distracting from the language practice and resultant learning. Listening to and responding to others in a discussion, whilst also assessing and recording personal performance, may have been a reason for the more '*individual speech style*' discussions by ProcS, as opposed to '*collaborative style*' discussions by ProdS (who did not need to record individual performance during discussions) and the possible 'trade-off' of other performance such as complexity (shorter turn length) and fluency (more repetitions) (Section 6.3.7). Careful thought is needed regarding the design of any GSF applied to group discussion tasks, as well as additional support which may be required to reduce cognitive load during discussions (such as the planning stage discussed in Section 6.3.7), to ensure that students can focus on improving their performance as much as possible. This possible 'trade-off' of student focus between in-task self-regulated performance tracking and performance development is extremely relevant to the application of GSF to group discussions and is discussed further in the conclusions section.

6.4.7 RQ3 key findings summary

The self-reported data collected in RQ3 has revealed some important findings which contribute to language teaching research. Firstly, the students in the study reported increasingly positive feelings towards engaging in learning through TBLT group discussions across time, as well as towards using the GSF provided. This has shown the potential for a combined TBLT and GSF approach to positively effect low-level students' attitudes towards second language group discussions. Furthermore, the data has shown that feedback provided by GSF can help students understand their performance in discussions and how to direct future efforts to improve (Atkin et al., 2001; Chappuis, 2005; Leung, 2005a; Stiggins et al., 2004), and that the use of classroom technology (such as the automated electronic diaries in the study) can be motivating to improve performance (Irving, 2015; Maier et al., 2016).

Secondly, I believe that the data has shown how the type of performance goals applied to group discussions will influence the focus and efforts made by students. Both product and process goals appeared to make students value meaning-focused more than form-focused performance across time, which would be expected to result in negotiation of meaning and improvements in communicative competence (Ellis, 2003; Littlewood, 2004; Long, 2014; Robinson, 2011; Willis & Willis, 2008). Also, process goals (focused on individual speech and interactions during discussions) seemed to focus students more on interactional performance (mainly asking questions and agreeing/disagreeing) than product goals. The use of personal performance goals for tasks has been shown to 'push' students to engage more in discussions in an interactive way, which supports the literature (Elliot & McGregor, 2001) and discussion of the observed effects on performance in RQ2 (Section 6.3.7). However, the stress which individual

goals were reported as putting ProcS under to perform, as described in other research (Latham et al., 2016, p. 3), may make students focus more on getting task scores than improving their actual language use or completing the task as a group (Seijts & Latham, 2001).

Finally, the data showed that additional support within learning may be required to support low-level students using a TBLT approach and GSF with group discussions. Students continued to report finding it difficult to give opinions within discussions and that topics were sometimes too difficult to discuss. ProcS also reported finding it difficult to use the Process sheet to assess and record performance at the same time as having a discussion. The cognitive load placed on low-level students undertaking TBLT group discussion, whilst using GSF, should be as low as possible to help them focus on using the language and improving over time. Suggestions to support this are the use of a pre-discussion planning stage (Ellis, 2009; Javad Ahmadian et al., 2015), topics which are as familiar to the students and connected as possible (Skehan, 1998, p. 99) and GSF being done outside of the discussion time (by reviewing recordings of discussions post-task for example). Implications and recommendations based on all of the above findings are discussed further in the conclusions section.

6.4.8 Limitations

Four main limitations need considering. Firstly, there was some **non-equivalency between ProdS and ProcS.** These were similar to the non-equivalency problem discussed for the observational data in Weeks 2 and 3 in RQ2 (Section 6.3.8) and again highlight the problem with comparing changes in data between two uneven groups. In RQ3, there were (a) differences in self-reported feelings towards discussions in Week 3 (Table 6.12), with ProcS generally

giving more positive responses than ProdS. Also, (b) the number of ProdS and ProcS students interviewed was both quite small and uneven. Only a total of 18 students (7 ProdS and 11 ProcS) were interviewed at both the start and end of the semester. The reliability of the data collected would have been better if more ProdS and ProcS had been available for interviews and if more students at the university could have been found to do the interviews. In addition, (c) the number of survey responses were uneven for the two groups (ProdS N = 50, ProcS N = 82) due to the difference in sizes of the classes undertaking the different approaches. This is why I used percentages of responses, rather than the number of students, but it means problems such as a larger group of students being statistically more likely to mention a wider variety of points than a smaller one.

A second limitation was **a lack of detail in responses, especially within peer-interviews.** Understanding exactly what students were referring to was sometimes difficult and required my own interpretation of what they meant. Answers given were generally short and sometimes not elaborated on (see the coding comments columns in Appendices O, P and Q for examples). If using peer-interviews again in the future, I would consider making two main changes to tackle this problem. Firstly, student interviews could be trained as interviewers for a longer period of time before starting a study. Their lack of ability to elicit more data on occasions was apparent to me by the end of the study and was a problem for the quality of data gathered. Secondly, student interviewers could socialize more with the students they will interview before the first interviews. By doing so they could build better rapport and potentially get more open and detailed responses from the interviewees.

Thirdly, there were **difficulties in coding the responses**. The interpretation problems discussed above made it hard to form categories and sub-categories for the overall data. With regards to reported test difficulties (Appendix O), one problem was with coding the responses related to 'expressing ideas in English'. This was a very common response in surveys, but I was unsure if it referred to problems with English vocabulary and grammar or delivering speech in English. Also, many students stated that their GSF sheet provided 'clarity of performance', but not specifically if this meant helping see their weak points or just an overall performance score. With regards to the reported usefulness of the GSF diaries, students often reported 'seeing scores' or 'having goals' as helpful. However, it was difficult to know sometimes if students meant that they were motivated by these scores and goals, or if they simply provided clarity for focus and performance. I decided to use broadly worded categories within the data tables to summarize such opinions in order to reduce the chance of misinterpreting responses. However, this still required my interpretation of short and unclear responses in some cases and also made my findings less specific about the student reports of potential problems encountered in tests, as well as the potential effects of GSF on learning.

Finally, there were some **challenges with interpreting the data tables**. It was (a) difficult to interpret the importance of factors based on the percentage of students who mentioned them (Tables 6.13, 6.16 and 6.17). I determined all of the points mentioned by any of the students as relevant, even though some points were mentioned much more often than others. I felt this would capture a full view of potential opinions rather than only the most common. However, this lack of accuracy in the potential 'relevance' of factors in the self-reported data was a potential weakness of the approach used. By then ordering factors in the tables by percentages,

some of them near the bottom of the lists may have appeared less significant than they actually were. Using pre-determined rating list surveys would have avoided this problem (rating factors from 1 to 6 for relevance for example) by making sure each student responded in some way to the same factors at the start and finish of the semester. However, as I previously stated, I did not wish to use such a list, as it may not cover potential factors which students would want to report as important for their learning. Additionally, (b) it was challenging to interpret discussion test difficulties mentioned in Week 3, but not in Week 13. If students described certain factors as problems at the start of the semester, but not at the end, this did not necessarily mean that they no longer considered those factors as problems. Although the fact that students mentioned certain difficulties first may suggest that those were more relevant, the students may have just chosen to focus their discussion on other things in their limited time to do the survey or interview on that particular day. Rating a set list of potential difficulties in both Weeks 3 and 13 could have avoided this problem by ensuring students rate the same set of potential difficulties at the start and end of the semester. However, I decided to use the open-ended survey/interview approach to gather as much data as possible from students which may be missed by using pre-formulated surveys for examples. Another possible explanation is that the students may not have wanted to repeat themselves in a second interview. Because they were interviewed twice by the same student (with the intent of helping them build rapport and give more open and detailed responses), they may have assumed the interviewer already knew the problems they had discussed in the first interview and decided not to repeat them. If this was the case, it could be avoided by alternating the interviewers in the future, so that a different student is interviewing the students each time. Also, students interviewing other students when meeting them for the

first time would mean they have less rapport than if they met multiple times to discuss the questions. One solution could be to have the interviewers socialize with the interviewees as many times as possible before the interviews in order to build rapport, as discussed above.

6.5 Summary of research question findings

This chapter addressed the overall research question for this thesis, which was '*What are the effects on learning of using Goal-Setting and Feedback (GSF) with TBLT group discussions across a semester?*' Table 6.20 summarizes the three RQs addressed within this thesis, as well as the overall findings regarding each RQ with the study data.

Two sets of performance goals were created for use in the study. Process goals were related to interactions (opinions, reasons, examples, questions and (dis)agreements), which I also adapted to make product goals (group choice, reasons, examples and other choices). The combined TBLT and GSF approach used was reported as enjoyable and motivating, as well as effective for creating more meaning-focused (rather than form-focused) discussions. Language use improvements were also seen for both groups across time (classroom accuracy and test fluency), especially for LPs, but lacking in other areas (classroom participation, repetitions, reformulations or complexity, or for test accuracy or complexity), suggesting 'trade-off' in performance due to limitations in focus, and a need for additional support in the learning (such as planning time and careful topic choice).

Table 6.20. Summary of RQ1-3 findings

Research question	Summary of findings
RQ1: What are appropriate discussion performance goals for the Japanese university students in this study?	- *Process* goals: interactions between students involving giving opinions (opinions, reasons, and examples), understanding one another (questions and answers), and (dis)agreeing (with at least one reason). - *Product* goals: decisions between students involving an agreed choice, reasons and examples for that choice, and other possible choices (with reasons).
RQ2, Part (a): How does observable discussion task performance change for the students across a semester using a TBLT approach (regardless of the type of GSF used)?	- Classroom accuracy (spoken errors) and test fluency (pauses) improves. - No significant improvements for classroom participation, repetitions, reformulations or complexity, or for test accuracy or complexity (due to a potential 'trade-off' between CAF measures, as well as between rubric scores and other aspects of performances, suggesting the need for additional support, such as planning stages).
RQ2, Part (b): What different effects do Product and Process GSF have on observable performance across a semester?	- Process GSF motivates students to speed up their discussions, resulting in faster speech rates, development of faster turn-taking tactics (such as more cues) and increases in the process measures used as goals (opinions, reasons and questions in this case). However, they also may 'trade-off' other performance aspects via shorter turns, more repetitions, and more '*individual speech style*' discussions (more 'off-task' turns, and less 'outcome-promoting' turns and clarifications). - Product GSF motivates students to reach agreed outcomes, resulting in more '*collaborative style*' discussions (more clarifications, reformulations and 'outcome-promoting' turns, as well as less 'off-task' turns), but with no significant improvement in speech rates, turn-taking tactics or process measures.
RQ2, Part (c): Are these effects the same for Low (LPs) and High Participators (HPs)?	- Spoken participation improves more overall for LPs. - ProdS HPs may dominate the speaking role in tests. - ProcS LPs are expected to show more improvements in participation, fluency and process measures than ProcS HPs. - More anxiety is expected in the learning for ProcS (especially LPs) than ProdS, due to the pressure of personal (rather than group) goals.
RQ3, Part (a): How do ProdS and ProcS report feeling about performing in discussions across the semester?	- Enjoyment, efforts and performance within group discussions are reported to improve across time. - Students report valuing meaning-focused performance more than form-focused performance across time. - Low-level students using TBLT and GSF with group discussions may desire additional support, such as a pre-test planning stage and careful topic choice (which are easy to discuss and are connected), to help them perform better across time.
RQ3, Part (b): How do they report feeling about the support the two types of GSF provided for their learning (or not)?	- GSF with group discussions is reported to help students understand their performance better and what to focus on to improve. - Electronic performance diaries which support group discussion GSF are reported to motivate students to improve performance. - Process GSF is reported to focus students more on interactional performance than the Product GSF. - Individual Process GSF is reported to put students under more pressure during performance than group Product GSF. - the cognitive load placed on students during discussions by the Process GSF is reported to be higher than by the Product GSF.

The GSF used was reported as being helpful to understand discussion performance and how to focus efforts to improve, as well as motivating to improve performance when tracked using the electronic diaries. The different effects of the two types of GSF used, as well as possible trade-off between different aspects of performance created by both, were also shown. The Process GSF motivated students to speed up discussions, resulting in improved participation, fluency and process measures (especially for the LPs), but came with higher cognitive load and stress within performance than Product GSF, encouraging shorter turns and more '*individual speech style*' discussions (more 'off-task' turns, and less 'outcome-promoting' turns and clarifications). On the other hand, the Product GSF promoted more '*collaborative style'* discussions (more clarifications, reformulations and 'outcome-promoting' turns, as well as less 'off-task' turns), but with little improvement in the performance measures. These findings are used to draw conclusions for this thesis within the next chapter in terms of their overall contributions to and implications for research and language teaching.

CHAPTER 7. CONCLUSIONS

This thesis examined how the use of Goal-Setting and Feedback (GSF) in Task-Based Language Teaching (TBLT) group discussions affects learning and task performance for students across time. This chapter explains the contributions made to research by the findings summarized in Table 6.20, followed by recommendations for language teaching which arise from those findings. Finally, the overall limitations of this thesis and suggested future research directions are discussed.

7.1 Contributions to research

Four important contributions to research have been made. Firstly, this is the first study I am aware of to use mixed-method longitudinal empirical data to show how a TBLT approach affects the learning and performance of low-level students undertaking group discussions across time. Although TBLT group discussions are perhaps expected by researchers and teachers to result in improved language use across time (as discussed in Chapter 2), there is a clear gap in the research to show what changes actually occur. Although the students in the study showed some improvements in CAF (accuracy and fluency-related) when a pre-discussion rehearsal stage was used in classes, most of the CAF measures did not improve, especially when the planning was not present (within testing). This suggests that **the improvements in linguistic performance which TBLT can nurture within group discussions with low-level learners appear to be limited**. Although discussions were reported as being enjoyable and motivational, the cognitive load placed on the students to improve their performance over time was also clear in the data. Thus, low-level learners may struggle to transfer learning across time and improve

performance, indicating possible 'trade-off' between CAF measures within group discussions, as suggested by Skehan (1998, p. 97) for other task types. This challenges the assumption which often appears to be made in the literature (e.g. Ellis, 2003; Long, 2014; Prabhu, 1987; Skehan, 1998; Willis & Willis, 2008) regarding the suitability of TBLT as a stand-alone approach for improving performance with low-level learners, when the focus is on *group discussions*. The use of an alternative approach, or additional supportive learning for TBLT, may be required to improve multiple aspects of linguistic performance over time (such as tasks focused on accuracy and/or complexity to address the lack of improvements which was evident in the study). Recommendations to address this within language teaching are discussed in the next section.

Secondly, this study contributes important data to goal-setting and feedback research. This study has added to the substantial literature on the effects of goal-setting, feedback and performance rubrics (see Chapter Four) by showing the impact of Goal-Setting and Feedback (GSF) on language learning through *group discussions*, a neglected aspect of the field. The self-reported data suggested that the addition of GSF helped clarify performance and motivated students to improve, and that the use of goals related to the process and product of discussions, as opposed to CAF performances, may have encouraged more meaning-focused (as opposed to form-focused) discussions. More significantly, the study showed how **different types of goals can be used to influence students' focus within, and resultant stress and performance, during group discussions**. Group product goals motivate students to collaborate more and reach goals together, while individual process goals motivate them to interact more often and at a greater speed across time, but with higher cognitive load and stress levels than the product goals. These findings have shown a deeper level of 'trade-off' occurring within performance, going

beyond that between CAF measures discussed above. This is a highly significant finding which helps elaborate on what 'trade-off' involves when considering the wider range of possible performances associated within *group discussions*. As the application of goals to discussions will direct the limited focus which students have towards certain areas of performance, there will inevitably be less focus (if any) on other areas. Although this study and other research has shown how the use of performance rubrics can help focus students on understanding their ability and how to improve it, the lowering of attention on performance which is unrelated to goals may result in less improvement in such areas than when GSF is not applied. Recommendations for applying GSF to language teaching are discussed in the next section.

Thirdly, this study has shown important differences in the effects of a combined TBLT and GSF approach for different students, based on their level of participation within discussions (LPs and HPs). TBLT has been reported as a more beneficial approach for students of a higher proficiency (Burrows, 2008; Tseng, 2006) and that having a lower level than others in a group task may prevent students from speaking at all (Foster, 1998). This was not the case in the study when GSF was applied to TBLT group discussions. LPs showed more improvements in participation than HPs and more improvements in fluency and process measures than HPs when using process goals. This new data demonstrates that **applying individualized GSF to TBLT groups discussions can help improve performance for different levels of performers, especially those who usually speak less than others within groups.** This finding reflects findings in the literature related to performance goal rubrics, which suggest that lower performers improve performance more than higher performers when using individual performance rubrics (Balan, 2012; Black & Wiliam, 1998b). The study has therefore

demonstrated, for the first time, how such individual rubrics can be applied to TBLT group discussions to support the learning of all students within groups and not just those who usually speak the most.

Finally, **this study has shown an efficient and fruitful method of data collection for classroom-based oral task research which can be replicated by other studies**. Observational data of multiple groups within the same classes was collected across time by combining audio and video files (Section 5.6.1), and combined with electronic surveys to capture large amounts of useful MMR data about student performance. In addition, the community research style peer-interviewing system used (Section 5.7.2) gathered large amounts of interview data whilst I was observing and grading student tests. This study has shown that careful organization of time and resources, as well as the inclusion of student interviews in data collection, can create a clearer picture of how students are learning. Recommendations for the application of such data collection methods within other language learning contexts is discussed in the next section.

7.2 Recommendations for language teaching

This section discusses four main recommendations for teachers, institutions and governments regarding language teaching, based on the above findings. Firstly, **the actual effects on learning shown in this study with a TBLT approach to group discussions with low-level learners need considering before applying such an approach to communication courses**. The data has shown that a TBLT approach should not be assumed to result in language use improvements (especially CAF measures) among low-level students which the literature and past study findings with other types of tasks may suggest (Chapter Two). Specifically, the

weekly use of a ten-minute pre-discussion rehearsal stage, an eight-minute discussion, and 30-minute post-discussion language review stage helped students improve accuracy and fluency within those discussions (but not within tests). In order to improve complexity, however, teachers may need to combine group discussion tasks with form-focused tasks which enable students to focus on using new vocabulary or grammatical forms. For instance, students could practice specific grammatical structures with 'production-practice activities' (Ellis, 2003, p. 261) or set lexical patterns within post-task review (Willis & Willis, 2008, pp. 194-196), with the objective of them transferring such learning to performance within future discussion tasks. However, the longer-term effects of this additional learning on performance cannot be determined from my study, as it did not go beyond the effects of TBLT as an overall approach.

Secondly, **the findings within the study should be used to help guide the design and implementation of GSF into communication courses**. Both sets of goal types used (shared group product and individual process) were shown to successfully direct students' focuses and behaviors towards improving at those specific performances across time. The product goals focused students more on collaborating to agree outcomes together, while the process goals focused student more on giving opinions and interacting during tasks, which can both be replicated for other courses using the sheets and diaries created (Appendices E and F). These positive findings for GSF suggest that performance measures for tasks should be determined for communication courses (such as by surveying/interviewing teachers, as in this study), and rubrics developed for students to use across time to help clarify performance and motivate them to improve in the ways intended. Teachers should make a list of what they consider the most important performance measures for their students (as done in Table 6.2 for the study) and

ensure that rubrics are aligned with those goals, and ideally with course assessment to further motivate students to succeed (Assessment Reform Group, 1999, p. 7). Those rubrics should be as quick and easy for students to use as possible by keeping the list short and measures easy to score (preferably by counting them as in the study). Adequate time should be allocated to using the rubrics during classes by teachers carefully organizing and perhaps cutting some of the work usually undertaken by students in that time. In addition, the use of individual performance goals, as opposed to shared group goals, was shown to improve performance more among students who usually spoke less within discussions (LPs), but added more stress to their learning. Thus, teachers should consider applying such accountability to each student for their performance across time, while remaining aware that it does add pressure and anxiety to the learning. Collecting and analyzing feedback from students across courses (such as surveys or observations by the teacher) is recommended to ensure such anxiety is not counter-productive to the learning (such as students having negative feelings regarding performance or avoiding participating). If stress levels are high, then the use of group goals may reduce the pressure felt by students and perhaps be more suitable.

Thirdly, **the findings should be used to help train both students and teachers how to apply GSF to communicative language courses**. This would be especially useful in Japan, where the government intends to improve the conversation skills of students at the school and university level (MEXT, 2009, 2013), but where TBLT faces challenges such as a lack of teacher skills and acceptance of it as an approach by both students and teachers because of its unclear connection to progress and testing (Deng & Carless, 2009; Lai, 2015). By showing how GSF may be applied to clarify and measure performance in the way found in the study, learning

would be expected to improve. At the start of courses, students should be allowed to see and discuss the potential benefits and problems associated with the approach (perhaps using the summary of findings in Table 6.20) before being asked to undertake it themselves. This greater understanding of the expected outcomes from using TBLT/GSF is likely to help students utilize it more than within the study, where it was undertaken for the first time. In addition, the same findings would be effective at helping train teachers to apply GSF to TBLT group discussions within their communication courses. The teaching and data collection methods, sheets and diaries used, as well as actual effects on performance found within the study, would help teachers understand how to use goals to support learning more and what effects to expect (or not). They would then be able to apply such knowledge to their own courses by adapting the rubrics to match their own approach to determining performance for different types of tasks with different students.

Finally, an implication for institutions and policy makers is that **classrooms should be equipped with technology to support GSF within communication courses**. Mainly, simple technology which allows students to record and assess their own discussions after they have been completed (such as the classroom management system and microphones on each desk in the study) would make post-task reviewing of performance possible. The study findings suggest that it is important to allow students to perform GSF outside of the actual task time with no need to track scores at that time, allowing them to focus more on improving their language use. Also, such recordings can provide teachers with more feedback on student performance across time which they would otherwise not get. In addition, computers, or similar technology, which can allow students the ability to track and monitor their GSF across time (such as the diaries in the

study) would automate and speed up the ability of students to self-regulate their performance tracking across courses, as well as provide another source of feedback for teachers on student performance.

7.3 Thesis limitations and future research directions

Three main limitations and areas of important future research were found for the study. Firstly, only ten teachers and 132 low-level students participated in the study and data was only collected for group discussions within a single university in Japan. Future research projects which use data from a wider variety of institutions and countries, larger groups of participants, different levels of learners (perhaps of a higher level), and with task types other than group discussions, would shed more light on the overall effects of TBLT and GSF on learning. This could then be used to check the above findings on a more universal basis and help guide the application of TBLT and GSF to improve language courses.

Secondly, the view of performance within discussions was limited to the amount of discussion which could be made within this thesis, and future studies which analyze different aspects of task performance may reveal alternative effects of TBLT and GSF. The analysis in the study could not expand further within the wordcount allowed to cover alternative aspects of performance, such as lexical complexity, which may reveal changes not discussed in this thesis. Such additional analyses of the effects of TBLT and GSF may further broaden the understanding of the learning they can achieve.

Thirdly, the separate effects which TBLT, SRL, goal-setting, Formative Assessment (FA) and supportive technology (electronic diaries in the study) had on the performance and self-

reported feelings by students could not be confidently isolated. For instance, the data did not show how using a TBLT approach alone (without the GSF) might affect learning. The study used all of these learning factors as a combined approach to providing students with the greatest support possible within their learning, based on past literature, but only the self-reported data in RQ3 was helpful for distinguishing which of them may have accounted for changes in performance seen in RQ2. Although the data provided in this thesis contributed significantly to the understanding of the effects of TBLT and GSF, future research projects which isolate and examine the separate effects of TBLT and GSF may help to consolidate understandings of how to operationalise them more effectively.

REFERENCES

Ahmadian, M., & Tajabadi, A. (2017). Patterns of Interaction in Young EFL Learners' Pair Work: The Relationship between Pair Dynamics and Vocabulary Acquisition. *3L: The Southeast Asian Journal of English Language Studies, 22*(3), 98-114.

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy.* New York: Longman.

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation, 10*(3), 1-11.

Andrade, H., Wang, X. L., Du, Y., & Akawi, R. L. (2009). Rubric-referenced self-assessment and self-efficacy for writing. *Journal of Educational Research, 102*(4), 287-301.

Angouri, J. (2010). Quantitative, qualitative or both? Combining methods in linguistic research. In: Litosseliti L (ed.), *Research Methods in Linguistics* (pp. 29-45). London: Continuum.

Aspinall, R. (2005). University entrance in Japan. In J. S. Eades, R. Goodman and Y. Hada (eds.), *The 'Big Bang' in Japanese Higher Education: The 2004 reforms and he dynamics of change* (pp. 199-218). Rosanna, VIC; Trans Pacific Press.

Assessment Reform Group. (1999). *Assessment for Learning: beyond the black box*. Cambridge: University of Cambridge School of Education.

Atkin, J. M., Black, P., & Coffey, J. (2001). *Classroom assessment and the national science standards.* Washington, DC: National Academies Press.

Bachman, L. F. (2000). Modern language testing at the turn of the century: Assuring that what we count counts. *Language testing, 17*(1), 1-42.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice: Designing and developing useful language tests* (Vol. 1). Oxford: Oxford University Press.

Baddeley, A. D. (1986). *Working memory*. London/New York: Oxford University Press.

Baddeley, A. D. (1993). Working memory or working attention? In A. Baddeley & L. Weiskrantz (Eds.), *Attention: Selection, awareness, and control* (pp. 152-170). Oxford: Oxford University Press.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. A. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47-89). New York: Academic Press.

Balan, A. (2012). Assessment for learning: A case study in mathematics education. Doctoral dissertation. Malmö University, Malmö, Sweden.

Bargh, J. A., Gollwitzer, P. M., Lee-Chai, A., Barndollar, K., & Trotschel, R. (2001). The automated will: Nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology, 81*(6), 1014-1027.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*, 612-637.

Batstone, R. (1996). Noticing. *ELT Journal 50*(3), 273.

Benson, S. D. (2016). Task-based language teaching: An empirical study of task transfer. *Language Teaching Research, 20*(3), 341-365.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2002). *Working inside the black box: Assessment for learning in the classroom*. London: King's College School of Education.

Black, P., & Wiliam, D. (1998a). Inside the Black Box: Raising standards through classroom assessment. *Phi Delta Kappan, 80*(2), 139-148.

Black, P., & Wiliam, D. (1998b). Assessment and classroom learning. *Assessment in Education: principles, policy & practice, 5*(1), 7-74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education), 21*(1), 5-31.

Blume, B. D., Ford, K. J., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*, 1065-1105.

Bong, M. (2009). Age-related differences in achievement goal orientation. *Journal of Educational Psychology, 101*, 879-896.

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing*, *20*(1), 89-110.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77-101.

Briggs, C. L. (1986). *Learning how to ask: A sociolinguistic appraisal of the role of the interview in social science research* (No. 1). Cambridge: Cambridge University Press.

British Educational Research Association (BERA), (2011). Accessed at https://www.bera.ac.uk/wp-content/uploads/2014/02/BERA-Ethical-Guidelines-2011.pdf on April 1st, 2015.

Brown, G. & Yule, G. (1983). *Teaching the spoken language.* Cambridge: Cambridge University Press.

Brown, H. D. (2007). *Principles of language learning and teaching.* White Plains, NY: Pearson Longman.

Burrows, C. (2008). An evaluation of task-based learning (TBL) in the Japanese classroom. *English Today, 24*(4), 11-16.

Butler, Y. (2011). The implementation of communicative and task-based language teaching in the Asia-Pacific region. *Annual Review of Applied Linguistics, 31*, 36-57.

Bygate, M. (1996). Effects of task repetition: appraising the developing language of learners. In J. Willis and D. Willis (Eds.) *Challenge and Change in Language Teaching* (pp. 136-146). Oxford: Heinemann.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan & M. Swain, *Task-based learning: language teaching, learning and assessment* (pp. 23-48). London: Longman.

Bygate, M. & Samuda, V. (2005). Integrative planning through the use of task repetition. In R. Ellis (Ed), *Planning and task performance in a second language* (Vol. 11, pp. 37-74). Amsterdam/New York: Benjamins.

Bygate, M., Skehan, P., & Swain, M. (2013). *Researching pedagogic tasks: Second Language Learning, Teaching and Testing.* London: Routledge.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In Richards, J. C., & Schmidt, R. W. (Eds.), *Language and Communication*, (pp. 2-27), London: Longman.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics,1*, 1-47.

Candlin, C. (1987). Towards task-based language learning. In Candlin, C. and Murphy, D., (eds), *Language learning task*. Englewood Cliffs, NJ: Prentice-Hall.

Carless, D. (2004). Issues in teachers' reinterpretation of a task-based innovation in primary schools. *TESOL Quarterly, 38(4),* 639-662.

Carless, D. (2007a). Student use of the mother tongue in the task-based classroom. *ELT Journal, 62*(4), 331-338.

Carless, D. (2007b). The suitability of task-based approaches for secondary schools: Perspectives from Hong Kong. *System, 35*(4), 595-608.

Carless, D. (2009a). Learning-oriented assessment: Principles, practice and a project. In Luanna H. Meyer, Susan Davidson, Helen Anderson, Richard B. Fletcher, Patricia M. Johnston & Malcolm Rees (eds.), *Tertiary Assessment & Higher Education Student Outcomes: Policy, Practice & Research* (pp. 79-90). Wellington, New Zealand: Ako Aotearoa.

Carless, D. (2009b). Revisiting the TBLT versus P-P-P debate: Voices from Hong Kong. *Asian Journal of English Language Teaching, 19,* 49-66.

Celce-Murcia, M. (2007). Rethinking the role of communicative competence in language teaching. In Soler, E. & Jorda, M. (eds.) *Intercultural language use and language learning* (pp. 41-57). Netherlands: Springer.

Celce-Murcia, M., Dörnyei, Z., & Thurrell, S. (1995). Communicative competence: A pedagogically motivated model with content specifications. *Issues in Applied linguistics, 6*(2), 5-35.

Chang, F. (2011). The Causes of Learners' Reticence and Passivity in English Classrooms in Taiwan. *The Journal of Asia TEFL, 8*(1), 1-22.

Chappuis, S. (2005). Is formative assessment losing its meaning? *Education Week*, *24*(44), 38.

Chen, G., & Kanfer, R. (2006). Toward a systems theory of motivated behavior in work teams. In B. Staw & L. Cummings (Eds.). *Research in organizational behavior* (Vol. 27, pp. 223-267). Greenwich, CT: JAI Press.

Clapham, C. (2000). Assessment and testing. *Annual Review of Applied Linguistics, 20*, 147-161.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*(1), 171-200.

Creswell, J. W. (2009). Editorial: Mapping the field of mixed methods research. *Journal of Mixed Methods Research, 3*(2), 95-108.

Creswell, J. W., & Plano Clark, V.L. (2011). *Designing and conducting mixed methods research* (2nd edition), Thousand Oaks, CA: Sage.

Cuesta, M. R. C. (1995). A task-based approach to language teaching: the case for task-based grammar activities. *Revista Alicantina de Estudios Ingleses, 8*, 91-100.

Dann, R. (2002). *Promoting assessment as learning: Improving the learning process*. New York, NY: Routledge.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305-334.

Davison, C., & Leung, C. (2009). Current issues in English language teacher-based assessment. *TESOL Quarterly, 43*(3), 393-415.

De Bot, K. (1992). A bilingual production model: Levelt's 'Speaking' model adapted, *Applied Linguistics, 13*, 1-24.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behavior.* New York: Plenum.

Deng, C. & Carless, D. (2009). The communicativeness of activities in a task-based innovation in Guangdong, China. *Asian Journal of English Language Teaching, 19*, 113-134.

Denscombe, M. (2014). *The good research guide: for small-scale social research projects.* Berkshire, UK: McGraw-Hill Education.

Denzin, N. (1970). Strategies of multiple triangulation. In Denzin, N. (ed.), *The Research Act in Sociology: A Theoretical Introduction to Sociological Method* (pp. 297-313). New York: McGraw-Hill.

Donato, R. (1994). Collective scaffolding in second language learning. In J. Lantolf & G. Appel (Eds.), *Vygotskian approaches to second language research* (pp. 33-56). Norwood, NJ: Ablex.

Dörnyei, Z. (2001). *Motivational strategies in the language classroom.* Cambridge: Cambridge University Press.

Dörnyei, Z. (2003). *Questionnaires in second language research: Construction, Administration, and processing.* Mahwah, NJ: Lawrence Erlbaum Associates.

Dörnyei, Z. (2007). *Research methods in Applied Linguistics*. Oxford: Oxford University Press.

Ehrman, M. E., & Dörnyei, Z. (1998). *Interpersonal dynamics in second language education: the visible and invisible classroom*. Thousand Oaks, CA: Sage.

Elder, C., & Iwashita, N. (2005). Planning for test performance: Does it make a difference? In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 219-238). Amsterdam: John Benjamins.

Elliot, A. J. (1999). Approach and avoidance motivation and achievement goals. *Educational Psychologist, 34,* 169-189.

Elliot, A. J., & McGregor, H. A. (2001). A 2 x 2 achievement goal framework. *Journal of Personality and Social Psychology, 80,* 501-519.

Ellis, R. & Barkhuizen, G. (2005). *Analysing learner language.* Oxford: Oxford University Press.

Ellis, R. (2002). Grammar teaching – practice or consciousness-raising? In J. C. Richards & W. A. Renandya (eds.), *Methodology in language teaching: An anthology of current practice* (pp. 167-174). Cambridge: Cambridge University Press.

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

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity and accuracy in l2 oral production. *Applied Linguistics 30*(4), 474-509.

Fengying, M. (2003). Motivating students by modifying evaluation methods. *English Teaching Forum, 41*, 38-41.

Firth, A., & Wagner, J. (2007). Second/foreign language learning as a social accomplishment: Elaborations on a reconceptualized SLA. *The Modern Language Journal*, *91*(1), 800-819.

Forsyth, D. (2000). One hundred years of group research: Introduction to the special issue. *Group Dynamics: Theory, Research, and Practice, 4*(1), 3-6.

Forsyth, D. (2016). *Group dynamics.* Belmont, CA: Cengage Learning.

Foster, P. (1998). A classroom perspective on the negotiation of meaning. *Applied Linguistics*, *19*(1), 1-23.

Foster, P. (1999). Key concepts in ELT. *ELT Journal, 53*(1), 69-70.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition*, *18*, 299-323.

Fredericks, J., Blumenfeld, P., & Paris, A. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59-109.

Fulcher, G. (2003). *Testing second language speaking*. Harlow: Pearson.

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment*. London, England & New York, NY: Routledge.

Gardner, A. K., Diesen, D. L., Hogg, D., & Huerta, S. (2016). The impact of goal setting and goal orientation on performance during a clerkship surgical skills training program. *The American Journal of Surgery*, *211*(2), 321-325.

Gass, S., & Mackey, A. (2007). Input, interaction and output: An overview. In K. Bardovi-Harlig & Z. Dornyei (Eds.), *AILA Review* (pp. 3-17). Amsterdam: Benjamins.

Gass, S., & Varonis, E. M. (1994). Input, interaction, and second language production. *Studies in second language acquisition*, *16*(3), 283-302.

Ghaith, G. (2002). Using Cooperative Learning to Facilitate Alternative Assessment. *Forum*, *40* (3), 26-31.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning and +/− Hereand-Now: effects on L2 oral production. In Maria del Pilar García-Mayo (ed.),

Investigating Tasks in Formal Language Learning (pp. 44-68). Clevedon: Multilingual Matters.

Goodson, L. & Phillimore, J. (2010). A community research methodology: working with new migrants to develop a policy related evidence base. *Social Policy and Society, 9*(4), 489-501.

Goodson, L. & Phillimore, J. (2012). *Community research for community participation: from theory to method.* Bristol, UK: Policy Press.

Gorsuch, G. (1998). Yakudoku EFL instruction in two Japanese high school classrooms: An exploratory study. *JALT Journal, 20*(1), 6-32.

Gower, R., & Walters, S. (1983). *Teaching Practice Handbook.* Oxford: Heinemann.

Greene, J. C. (2008). Is mixed methods social inquiry a distinctive methodology? *Journal of Mixed Methods Research, 2*(1), 7-22.

Greene, J. C., Caracelli, V.J., & Graham, W.F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), 255-274.

Guará-Tavares, M. G. (2011). Pre-task planning, working memory capacity and L2 speech performance. *Organon (UFRGS), 26*, 245-266.

Guará-Tavares, M. G. (2013). Working memory capacity and L2 speech performance in planned and spontaneous conditions: a correlational analysis. *Trabalhos em Linguística Aplicada (UNICAMP), 52*, 9-29.

Guará-Tavares, M. G. (2016). Learners' processes during pre-task planning and Working Memory Capacity. *Ilha do Desterro, 69*(1), 79-94.

Guillot, M. N. (1999). *Fluency and its teaching* (Vol. 11). Clevedon: Multilingual Matters.

Hackman, M. Z., & Johnson, C. E. (2013). *Leadership: A communication perspective.* Illinois, USA: Waveland Press.

Hamp-Lyons, L. (2007). The impact of testing practices on teaching: Ideologies and alternatives. In J. Cummins & C. Davison (Eds.), *International handbook of English language teaching* (pp. 487-504). New York: Springer.

Harlen, W., & James, M. (1997). Assessment and learning: differences and relationships between formative and summative assessment. *Assessment in Education: Principles, Policy & Practice*, *4*(3), 365-379.

Hart, W., & Albarracín, D. (2009). The effects of chronic achievement motivation and achievement primes on the activation of achievement and fun goals. *Journal of Personality and Social Psychology, 97*(6), 1129-1141.

Haskell, R. E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. San Diego, CA: Academic Press.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 401-435). Rowley, MA: Newbury House.

Hedge, T. (1993). Key concepts in ELT. *ELT journal, 47*(3), 275-277.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics*, *30*(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, Accuracy and Fluency: Definitions, Measurement and Research. In Housen, Kuiken & Vedder (eds), *Dimensions of L2 Performance and Proficiency* (pp. 1-20). Amsterdam: John Benjamins.

Howatt, A. P. R. (1984). *A History of English Language Teaching.* Oxford: Oxford University Press.

Huang, H. T. D., & Hung, S. T. A. (2013). Comparing the effects of test anxiety on independent and integrated speaking test performance. *TESOL Quarterly, 47*(2), 244-269.

Huang, S. C. (2011). Convergent vs. divergent assessment: Impact on college EFL students' motivation and self-regulated learning strategies. *Language Testing, 28*(2), 251-271.

Huhta, A. (2008). Diagnostic and Formative Assessment. In B. Spolsky & F. M. Hult (Eds.), *The Handbook of Educational Linguistics* (pp. 467-482). Hong Kong: Blackwell Publishing.

Hunt, K. (1965). *Grammatical structures written at three grade levels* (Rep. No. 3). Champaign,IL: National Council of Teachers of English.

Hymes, D. (1972). On communicative competence, in Pride, J. B. and Holmes, J. (eds), *Sociolinguistics.* Harmondsworth: Penguin.

Hymes, D. (1974). *Foundations in sociolinguistics: an ethnographic approach*. Philadelphia: Centre for Curriculum Development.

Irving, K. E. (2015). Technology-Assisted Formative Assessment. In Urban, M. and Falvo, D. (eds), *Improving K-12 STEM Education Outcomes through Technological Integration* (pp. 380-398). IGI Global.

Ivankova, N. V., & Greer, J. L. (2015). Mixed Methods Research and Analysis. In Paltridge B. and Phakiti, A. (Eds.), *Research methods in applied linguistics: A practical resource* (pp. 63-81). New York: Bloomsbury.

Iwashita, N., & Li, H. F. (2012). Patterns of corrective feedback in a task-based adult EFL classroom setting in China. In A. Shehadeh & C. A. Coombe (Eds), *Task-based teaching*

in foreign language contexts: Research and implementation (pp. 137-161). Amsterdam: John Benjamins.

Javad Ahmadian, M., Tavakoli, M., & Vahid Dastjerdi, H. (2015). The combined effects of online planning and task structure on complexity, accuracy and fluency of L2 speech. *The Language Learning Journal, 43*(1), 41-56.

Johnson, D. W., Johnson, R., & Holubec, E. (2013). *Cooperation in the classroom* (9th ed.). Edina, MN: Interaction Book Company.

Johnson, M. (2001). *The art of non-conversation: A reexamination of the validity of the oral proficiency interview*. New Haven: Yale University Press.

Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 143-164). Amsterdam: John Benjamins.

Keck, C., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A quanititative meta-analysis. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91-132). Amsterdam: Benjamins.

King, J. (2012). Silence in the second language classrooms of Japanese universities. *Applied Linguistics, 34*(3), 325-343.

King, J. (2013). *Silence in the second language classroom.* New York: Palgrave Macmillan.

Klem, A. M., & Connell, J. P. (2004). Relationships matter: Linking teacher support to student engagement and achievement. *Journal of school health, 74*(7), 262-273.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5-24). Ann Arbor: University of Michigan Press.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. Addison-Wesley Longman Ltd.

Kuhl, J. (2000). A functional-design approach to motivation and self-regulation. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), *Handbook of self-regulation* (pp. 111-169). San Diego, CA: Academic Press.

Lai, C. & Lin, X. (2015) Strategy training in a task-based language classroom, *The Language Learning Journal, 43*(1), 20-40.

Lai, C. (2015). Task-based language teaching in the Asian context: Where are we now and where are we going? In Thomas, M., Reinders, H. (Eds.), *Contemporary task-based language teaching in Asia* (pp. 12-29). London: Bloomsbury.

Lai, C., Zhao, Y., & Wang, J. (2011). Task-Based Language Teaching in Online Ab Initio Foreign Language Classrooms. *The Modern Language Journal*, *95*(1), 81-103.

Lantolf, J. (1993). Sociocultural theory and the second language classroom: The lesson of Strategic Interaction. In J. E. Alatis (Ed.), *Strategic interaction and language acquisition: Theory, practice, and research* (pp. 220-233). Washington, DC: Georgetown University Press.

Larsen-Freeman, D., & Anderson, M. (2013). *Techniques and Principles in Language Teaching* (3rd edition). Oxford: Oxford University Press.

Latham, G., & Locke, E. (2007). New developments in and directions for goal-setting research. *European Psychologist, 12*(4), 290-300.

Latham, G., Seijts, G., & Slocum, J. (2016). The goal setting and goal orientation labyrinth. *Organizational Dynamics, 4*(45), 271-277.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387-417.

Leung, C. (1999). Teachers' response to linguistic diversity. In A. Tosi & C. Leung (Eds.), *Rethinking language education: From a monolingual to a multilingual perspective* (pp. 225-240). London: Centre for Information on Language Teaching and Research.

Leung, C. (2005a). Convivial communication: Re-contextualizing communicative competence. *International Journal of Applied Linguistics, 15*(2), 119-144.

Leung, C. (2005b). English as an additional language policy: Issues of inclusive access and language learning in the mainstream. *Prospect, 20*(1), 95-113.

Leung, C., & Lewkowicz, J. (2006). Expanding horizons and unresolved conundrums: Language testing and assessment. *TESOL Quarterly, 40*(1), 211-234.

Levelt, W. J. M. (1989). *Speaking: from intention to articulation*. Cambridge, MA: MIT Press.

Littlewood, W. (2004). The task-based approach: Some questions and suggestions. *ELT journal, 58*(4), 319-326.

Littlewood, W. (2007). Communicative and task-based language teaching in East Asian classrooms. *Language Teaching, 40*(3), 243-249.

Locke, E. A., & Latham, G. P. (2002). Building a practically useful theory of goal setting and task motivation: A 35-year odyssey. *American Psychologist, 57*(9), 705-717.

Long, M. (1977). Group work in the teaching and learning of English as a foreign language - problems and potential. *English Language Teaching Journal, 31*(4), 285-292.

Long, M. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and Assessing Second Language Acquisition* (pp. 77-99). Clevedon, England: Multilingual Matters.

Long, M. (1989). Task, group, and task-group interactions. *University of Hawai'i Working Papers in ESL, 8*, 1-25.

Long, M. (1990). Task, group and task-group interactions. In Anivan, S. (ed.), *Language teaching methodology for the nineties* (pp. 31-50). Singapore: Regional English Language Centre/Singapore University Press.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie and T. Bhatia (eds.), *Handbook of second language acquisition* (pp. 413-468). San Diego: Academic Press.

Long, M. (2014). *Second language acquisition and task-based language teaching*. Oxford: John Wiley & Sons.

Long, M. & Porter, P. (1985). Group work, interlanguage talk, and second language acquisition. *TESOL Quarterly, 19*, 305-325.

Lynch, T., & Anderson, K. (1992). *Study speaking. A course in spoken English for academic purposes*. Cambridge: Cambridge University Press.

Lynch, T., & Maclean, J. (2000). Exploring the benefits of task repetition and recycling for classroom language learning. *Language Teaching Research, 4*, 221-250.

Lynch, T., & Maclean, J. (2001). A case of exercising: Effects of immediate task repetition on learners' performance. In M. Bygate, P. Skehan and M. Swain (Eds.), *Researching pedagogic tasks: Second Language learning, teaching and testing* (pp. 141-162). Harlow: Longman (Pearson Education).

Maehr, M. L. (1984). Meaning and motivation: Toward a theory of personal investment. In R. Ames & C. Ames (Eds.), *Research on motivation in education: Student motivation* (Vol. 1, pp. 115-144). New York: Academic Press.

Maier, U., Wolf, N., & Randler, C. (2016). Effects of a computer-assisted formative assessment intervention based on multiple-tier diagnostic items and different feedback types. *Computers & Education, 95,* 85-98.

Mann, S. (2011). A critical review of qualitative interviews in applied linguistics. *Applied Linguistics, 32*(1), 6-24.

Martin, B., McNally, J., & Taggar, S. (2016). Determining the importance of self-evaluation on the goal-performance effect in goal setting: Primary findings. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement*, *48*(2), 91-100.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20,* 83-108.

MEXT [Japanese Ministry of Education, Culture, Sports, Science and Technology]. (2009). Koutougakkou gakushu shidou yoryo gaikokugo eigoban kariyaku [Study of course guideline for foreign languages in senior high schools; provisional version]. Retrieved from www.mext.go.jp/a_menu/shotou/new-cs/youryou/eiyaku/1298353.htm on Nov 31st, 2017.

MEXT [Japanese Ministry of Education, Culture, Sports, Science and Technology]. (2013). Report on the Future Improvement and Enhancement of English Education. Retrieved from http://www.mext.go.jp/en/news/topics/detail/1372625.htm on Nov 31st, 2017.

Miller, R. B., Greene, B. A., Montalvo, G. P., Ravindran, B., & Nichols, J. D. (1996). Engagement in academic work. The role of learning goals, future consequences, pleasing others, and perceived ability. *Contemporary Education Psychology*, *21*, 388-422.

Moskowitz, G. B., & Grant, H. (2009). *The psychology of goals.* New York: Guilford Press.

Nahrgang, J. D., DeRue, D. S., Hollenbeck, J. R., Spitzmuller, M., Jundt, D. K., & Ilgen, D. R. (2013). Goal setting in teams: The impact of learning and performance goals on process and performance. *Organizational Behavior and Human Decision Processes*, *122*(1), 12-21.

Nishino, T. (2008). Japanese secondary school teachers' beliefs and practices regarding communicative language teaching: An exploratory survey. *JALT Journal*, *30*(1), 27-50.

Nishino, T., & Watanabe, M. (2008). Communication-oriented policies versus classroom realities in Japan. *TESOLl Quarterly*, *42*(1), 133-138.

Norris, J. M. (2008). *Validity evaluation in language assessment*. Frankfurt am Main, Germany: Peter Lang.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578.

Nunan, D. (1989). *Designing tasks for the communicative classroom.* Cambridge: Cambridge University Press.

Ogilvie, G., & Dunn, W. (2010). Taking teacher education to task: Exploring the role of teacher education in promoting the utilization of task-based language teaching. *Language Teaching Research, 14*(2), 161-181.

Orsmond, P., Merry, S., & Reiling, K. (1997). A study in self-assessment: Tutor and students' perceptions of performance criteria. *Assessment & Evaluation in Higher Education, 22*(4), 357-368.

Oxford, R. L. (1997). Cooperative learning, collaborative learning, and interaction: Three communicative strands in the language classroom. *The Modern Language Journal, 81*(4), 443-456.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601.

Panadero, E. (2011). Instructional help for self-assessment and self-regulation: Evaluation of the efficacy of self-assessment scripts vs. rubrics. Doctoral dissertation. Universidad Autónoma de Madrid, Madrid, Spain.

Panadero, E., Alonso-Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation, learning and self-efficacy in secondary education. *Learning and Individual Differences, 22*(6), 806-813.

Panadero, E., & Jönsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited. *Educational Research Review, 9*, 129-144.

Parks-Stamm, E. J., Oettingen, G., & Gollwitzer, P. M. (2010). Making sense of one's actions in an explanatory vacuum: The interpretation of nonconscious goal striving. *Journal of Experimental Social Psychology, 46*(3), 531-542.

Philp, J., & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics, 36,* 50-72.

Philp, J., Oliver, R., & Mackey, A. (2006). The impact of planning time on children's task-based interactions. *System, 34*(4), 547-565.

Pica, T. (1994). Research on Negotiation: What Does It Reveal About Second-Language Learning Conditions, Processes, and Outcomes? *Language learning*, *44*(3), 493-527.

Poupore, G. (2016). Measuring group work dynamics and its relationship with L2 learners' task motivation and language production. *Language Teaching Research, 20*(6), 719-740.

Prabhu, N. S. (1987). *Second language pedagogy* (Vol. 20). Oxford: Oxford University Press.

Price, M., Handley, K., Millar, J., & O'Donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*, 277-289.

Reeve, J. (2012). A self-determination theory perspective on student engagement. In S.L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 149-172). New York, NY: Springer.

Reeve, J., Deci, E. L., & Ryan, R. M. (2004). Self-determination theory: A dialectal framework for understanding the sociocultural influences on learner motivation. In D. McInerney, & S. Van Etten (Eds.), *Research on sociocultural influences on motivation and learning: Big theories revisited* (vol. 4, pp. 31-59). Greenwich, CT: Information Age.

Reynolds-Keefer, L. (2010). Rubric-referenced assessment in teacher preparation: An opportunity to learn by using. *Practical Assessment, Research & Evaluation, 15*(8), 1-9.

Rignall, M., & Furneaux, C. (1997). *Speaking: students' book (English for academic study series)*. Hertfordshire: Prentice Hall Europe.

Ritchie, J., Spencer, L., & O'Connor, W. (2003). Carrying out qualitative analysis. In J. Ritchie & J. Lewis (Eds.), *Qualitative research practice: A guide for social science students and researchers* (pp. 219-262). London: Sage.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*, 27-57.

Robinson, P. (2003). Attention and memory during SLA. In C. M. Doughty and M. H. Long (eds.), *Handbook of Second Language Acquisition* (pp. 631-678). Oxford: Blackwell.

Robinson, P. (2007). Triadic framework for TBLT: Task complexity, task difficulty, and task condition. *The Journal of Asia TEFL*, 195-225.

Robinson, P. (2011). Task-based language learning: A review of issues. *Language Learning, 61*(1), 1-36.

Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American Psychologist, 55*(1), 68-78.

Ryan, R. M., & Patrick, H. (2001). The classroom social environment and changes in adolescents' motivation and engagement during middle school. *American Educational Research Journal, 38*, 437-460.

Ryan, R. M., & Pintrich, P. R. (1997). "Should I ask for help?" The role of motivation and attitudes in adolecents' help seeking in math class. *Journal of Educational Psychology, 89*, 329-341.

Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language, 50*, 696-735.

Sadler, D. R. (1998). Formative assessment: Revisiting the territory. *Assessment in education: principles, policy & practice, 5*(1), 77-84.

Salvisberg, J. A. (2011). *Diagnostic Oral Skills Assessment: Developing Flexible Guidelines for Formative Speaking Tests in EFL Classrooms Worldwide.* Oxford: Peter Lang AG.

Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 111-141). Amsterdam: John Benjamins.

Savignon, S. J. (2002). Communicative Curriculum Design for the 21st Century. In *Forum* (Vol. 40, No. 1, pp. 2-7).

Schamber, J. F., & Mahoney, S. L. (2006). Assessing and improving the quality of group critical thinking exhibited in the final projects of collaborative learning groups. *The Journal of General Education, 55*(2), 103-137.

Seijts, G. H., & Latham, G. P. (2001). The effect of distal learning, outcome, and proximal goals on a moderately complex task. *Journal of Organizational Behavior*, *22*(3), 291-307.

Sitzmann, T. & Ely, K. (2011). A meta-analysis of self-regulated learning in work-related training and educational attainment: what we know and where we need to go. *Psychol. Bull. 137*(3), 421-442.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied linguistics*, *17*(1), 38-62.

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: The influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 193-216). Amsterdam: John Benjamins.

Smith, J.A., & Osborn, M. (2008). Interpretative phenomenological analysis. In Smith, J. (ed.), *Qualitative psychology: A practical guide to research methods* (pp. 53-80). London: Sage.

Sook, K. (2003). The types of speaking assessment tasks used by Korean Junior Secondary school English teachers. *Asian EFL Journal*, *5*(4), 1-11.

Sparks, C. (2010). Teacher reaction to and understanding of a task-based embedded syllabus in Queensland. *Flinders University Languages Group Online Review, 4*(2), 73-92.

Stiggins, R., Arter, J., Chappuis, J., & Chappuis, S. (2004). *Classroom assessment for student learning—Doing it right, using it well*. Portland, OR: Educational Testing Service.

Storch, N. (2002). Patterns of interaction in ESL pair work. *Language Learning, 52*(1), 119-158.

Stroud, R. (2013). Power-sharing in the Asian TBL classroom: Switching from teacher to facilitator. *On Task*, *3*(1), 4-11.

Stroud, R. (2017). The impact of task performance scoring and tracking on second language engagement. *System*, *69,* 121-132.

Sun, Y. (2011). The influence of the social interactional context on test performance: A sociocultural view. *Canadian Journal of Applied Linguistics, 14*(1), 194-221.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. *Input in second language acquisition*, *15*, 165-179.

Swain, M. (1995). Three functions of output in second language learning. *Principle and practice in applied linguistics: Studies in honour of HG Widdowson, 2*(3), 125-144.

Swan, M. (2005). Legislation by hypothesis: The case of task-based instruction. *Applied Linguistics, 26*(3), 376-401.

Tarone, E. (1980). Communication strategies, foreigner talk, and repair in interlanguage. *Language Learning, 30*(2), 417-428.

Tashakkori, A., & Teddlie, C. (2003). Issues and dilemmas in teaching research methods courses in social and behavioural sciences: US perspective. *International Journal of Social Research Methodology, 6*(1), 61-77.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-276). Amsterdam: John Benjamins.

Thomas, M., & Reinders, H. (2015). *Contemporary task-based language teaching in Asia*. London: Bloomsbury Publishing.

Thornbury, S. (1999). *How to teach grammar.* Harlow: Longman.

Tseng, C. Y. (2006). A study of the effect of task-based instruction on primary school EFL students. Unpublished master's thesis. National Chung Cheng University, Taiwan.

Tsui, A. (2001). Classroom Interaction. In C Carter and D. Nunan (Eds.), *The Cambridge Guide to Teaching English to Speakers of Other Languages* (pp. 120-126). Cambridge: Cambridge University Press.

Tunstall, P., & Gipps, C. (1996). Teacher feedback to young children in formative assessment: A typology. *British educational research journal, 22*(4), 389-404.

Van den Branden, K. (Ed.). (2006). *Task-based language education.* Cambridge: Cambridge University Press.

Vohs, K. D., & Baumeister, R. F. (2016). *Handbook of self-regulation: Research, theory, and applications.* New York: Guilford Publications.

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Vygotsky, L. S. (1986). *Thought and language*. Cambridge: MIT Press.

Wang, Z. (2014). Developing accuracy and fluency in spoken English of Chinese EFL learners. *English Language Teaching, 7*(2), 110-118.

Waters, A. (2009). Managing innovation in English language education. *Language Teaching*, *42*(4), 421-458.

Watson, C. (2009). The impossible vanity: uses and abuses of empathy in qualitative inquiry. *Qualitative Research, 9*(1), 105-117.

Weir, C. J. (2005). *Language testing and validation.* Hampshire: Palgrave McMillan.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106.

Wiliam, D. (2018). *Embedded Formative Assessment*. Bloomington, IN: Solution Tree Press.

Williams, K., & Andrade, M. (2008). Foreign Language Learning Anxiety in Japanese EFL University Classes: Causes, Coping, and Locus of Control. *Electronic Journal of Foreign Language Teaching, 5*(2), 181-191.

Willis, D., & Willis, J. (2008). *Doing task-based teaching*. Oxford: Oxford University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. Honolulu, Hawaii: University of Hawaii Press.

Wolters, C. A. (2003). Regulation of motivation: Evaluating an underemphasized aspect of self-regulated learning. *Educational psychologist*, *38*(4), 189-205.

Wolters, C. A. (2004). Advancing achievement goal theory: Using goal structures and goal orientations to predict students' motivation, cognition, and achievement. *Journal of Educational Psychology, 96*, 236-250.

Woodrow, L. (2006). Anxiety and Speaking English as a Second Language. *RELC Journal, 37*(3), 308-328.

Yu, G. (2010). Lexical Diversity in Writing and Speaking Tasks Performances. *Applied Linguistics. 31*(2), 236-59.

Yuan, F., & Ellis, R. (2003). The effects of pre-task and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics 24*(1), 1-27.

Zimmerman, B. J., & Campillo, M. (2003). Motivating self-regulated problem solvers. In J. E. Davidson & R. J. Sternberg (Eds.), *The Psychology of Problem Solving* (pp. 233-263). New York, NY: Cambridge University Press.

Zimmerman, B. J., & Moylan, A. R. (2009). Self-regulation: where metacognition and motivation intersect. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp. 299-315). New York, NY: Routledge.

Zimmerman, B. J., & Schunk, D. H. (2011). *Handbook of Self-Regulation of Learning and Performance.* New York, NY: Routledge.

APPENDICES

Appendix A. Summary of Module One findings regarding factors affecting group discussion participation for Japanese university students

Category	Sub-category	Mean score (out of five)		Recommendations for instructors
Topic (Barrier)	Discussion topic knowledge	3.48	*Discussion task design*	Instructors should choose topics which learners have adequate knowledge about and interest in to promote participation (Bachman & Palmer, 1996; Cao & Philp, 2006; Hann, 2007). Allow learners to choose the topic to increase the chance of their participation (Deci & Ryan, 2000).
	Discussion topic interest	3.43		
	Relevance of topic to learners' majors	3.14		
Relevance (Boost)	Topics connected to real world issues	3.08		
Choice (Boost)	Learners choosing the topic	3.69		
Support (Boost)	Pre-discussion brainstorming	3.63		Instructors should consider how to allocate time prior to discussions for learners to prepare with brainstorming and practise (Chang, 2011; Hann, 2007; Nakatsuhara, 2011), and with a clear idea of how to take part well in the discussion with a clear scoring system for example (Dörnyei, 2001, p. 28; Tsuo, 2005).
	Pre-discussion peer practice	3.54		
	Seeing a scoring explanation for getting discussion points	3.34		
Support (Boost)	Peer feedback after discussions	3.13		Instructors should design discussion tasks to allow for learners to receive feedback on their discussions after they are completed. This could be done through peer feedback, self-recording, or from the instructor for example (Chang, 2011; Harumi, 2010).
Outcome (Boost)	Learners seeing measurable progress of their discussion skills	3.53		
Support (Boost)	Repeating task types across time	3.47		Discussion tasks should be kept uniform in design across time (in terms of parameters such as time limits, group sizes, outcome expectations and member roles), so as to allow learners to focus mostly on participating orally, rather than understanding the task (Ellis, 2003, p. 226; Fushino, 2010).
Enjoyment (Boost)	Discussions more like games and puzzles	3.59		Designing a discussion task to be more humorous, or more like a game (with game parameters such as points being involved) may result in learners taking part more (Author, 2013a).
	Laughing during the discussion	3.85		
L2 Use (Barrier)	A lack of vocabulary	3.26		Instructors should ensure that adequate time is allocated to learning and retaining relevant vocabulary during discussion skills courses, so learners can feel they have the language necessary to take part in discussions (Ferris, 1998; Harumi, 2010).

Group Dynamics (Barrier)	Members being close friends or not	3.59	*Group set-up*	Instructors should ensure that group members are comfortable with each other in terms of familiarity and comfort with each other to promote open participation (Cao & Phillip, 2006; Forsyth, 2006, p. 29; Fushino, 2010; Nakatsuhara, 2011; Sugiman, 1997). Instructors should change group members (or allow learners to choose their own groups) to make sure group dynamics do not act as a significant barrier for participation in discussions (Tsui, 2001; Williams & Andrade, 2008; Woodrow, 2006).
	Differing English abilities amongst members	3.40		
Culture (Barrier)	A fear of making a mistake when speaking	3.06		
Support (Boost)	Positive feedback or praise from the instructor	3.51	*Instructor actions*	Instructors should monitor group interaction closely during discussions and be ready to interact directly with groups (in order to improve learner participation) if necessary (Cheng & Dörnyei, 2007; Dörnyei, 2001; Guilloteaux & Dörnyei, 2008; Williams & Andrade, 2008). Asking group members questions, giving them help, and praising contributions when participation levels are low can help boost participation (Harumi, 2010; Patrick et al., 2000).
	Direct instructor support during discussions	3.30		

Appendix B. Summary of Module Two findings regarding the effects of group discussion planning for Japanese university students

	LOW participators (N=12)		**HIGH participators (N=12)**		**OVERALL (N=24)**		**(N=24)**			**95% confidence interval**			
Measure (NONE=no planning STR=Strategic REH=Rehearsal)	**MEAN**	**SD**	**MEAN**	**SD**	**MEAN**	**SD**	**Mean change**	**t**	**Sig. (two-tailed)**	**lower**	**upper**	**η^2**	**Effect size**
NONE Words / min	6.85	3.29	16.06	4.97	11.46	6.25							
STR Words / min	10.51	2.55	16.34	5.03	13.43	4.91	1.97	-2.127	**0.044**	-3.89	-0.05	**0.16**	**large**
REH Words / min	10.13	3.49	18.55	5.25	14.34	6.13	2.88	-3.077	**0.005**	-4.82	-0.94	**0.29**	**large**
NONE Syllables / min	8.74	4.02	20.56	6.14	14.65	7.89							
STR Syllables / min	13.99	3.49	20.65	7.06	17.32	6.42	2.67	-2.149	**0.042**	-5.23	-0.10	**0.17**	**large**
REH Syllables / min	13.36	4.41	22.79	6.02	18.08	7.06	3.43	-2.75	**0.011**	-6.00	-0.85	**0.25**	**large**
NONE Turns / min	0.40	0.30	0.50	0.25	0.45	0.27							
STR Turns / min	0.42	0.33	0.54	0.28	0.48	0.30	0.03	-0.83	0.415	-0.11	0.05	0.03	small
REH Turns / min	0.36	0.16	0.73	0.45	0.55	0.38	0.10	-1.602	0.123	-0.23	0.03	0.10	medium
NONE Speech Rate A	67.26	31.24	76.38	19.76	71.82	25.99							
STR Speech Rate A	84.96	20.96	82.15	20.33	83.56	20.24	11.74	-2.263	**0.033**	-22.46	-1.01	**0.18**	**large**
REH Speech Rate A	84.52	18.09	84.05	16.82	84.29	17.08	12.46	-2.591	**0.016**	-22.41	-2.51	**0.23**	**large**
NONE Speech Rate B	62.57	30.02	64.93	21.53	63.75	25.58							
STR Speech Rate B	75.92	19.10	73.30	21.01	74.61	19.68	10.86	-2.193	**0.039**	-21.10	-0.61	**0.17**	**large**
REH Speech Rate B	78.12	21.60	74.53	17.69	76.33	19.39	12.58	-2.616	**0.015**	-22.52	-2.63	**0.23**	**large**
NONE L1 fillers / min speaking	5.16	4.91	7.53	4.87	6.35	4.93							

STR L1 fillers / min speaking	5.64	4.78	9.13	6.58	7.39	5.90	1.04	-1.464	0.157	-2.52	0.43	0.09	medium
REH L1 fillers / min speaking	4.04	2.97	8.52	6.73	6.28	5.58	-0.07	0.084	0.934	-1.55	1.68	0.00	small
NONE Pauses / min speaking	9.94	4.02	8.83	2.01	9.39	3.16							
STR Pauses / min speaking	8.30	3.01	6.85	2.27	7.58	2.71	-1.81	2.764	**0.011**	0.45	3.16	**0.25**	**large**
REH Pauses / min speaking	7.50	2.71	6.62	1.91	7.06	2.34	-2.33	3.026	**0.006**	0.74	3.92	**0.28**	**large**
NONE Repetitions / min speaking	1.12	1.66	3.28	2.72	2.20	2.46							
STR Repetitions / min speaking	2.05	1.56	3.25	3.15	2.65	2.51	0.45	-0.732	0.472	-1.74	0.83	0.02	small
REH Repetitions / min speaking	1.44	1.76	3.03	3.26	2.23	2.69	0.04	-0.07	0.945	-1.21	1.13	0.00	small
NONE Reformulations / min speaking	1.41	1.14	2.67	1.34	2.04	1.38							
STR Reformulations / min speaking	1.53	1.21	1.60	0.88	1.57	1.04	-0.48	1.539	0.137	-0.16	1.11	0.09	medium
REH Reformulations / min speaking	1.12	1.36	2.29	0.96	1.70	1.30	-0.34	1.311	0.203	-0.20	0.88	0.07	medium
NONE Ref clauses / 100 pruned words	6.40	4.61	5.60	3.69	6.00	4.10							
STR Ref clauses / 100 pruned words	2.66	2.37	4.66	2.44	3.66	2.57	-2.34	2.496	**0.02**	0.40	4.29	**0.21**	**large**
REH Ref clauses / 100 pruned words	5.38	4.88	5.54	2.39	5.46	3.76	-0.54	0.705	0.488	-1.05	2.14	0.02	small
NONE Number of reasons / turn	0.93	0.57	1.91	1.13	1.42	1.01							
STR Number of reasons / turn	2.32	1.73	1.96	1.30	2.14	1.51	0.73	-2.324	**0.029**	-1.37	-0.08	**0.19**	**large**
REH Number of reasons / turn	1.77	1.15	1.73	0.93	1.75	1.02	0.33	-1.491	0.15	-0.79	0.13	0.09	medium

Appendix O. Student discussion test difficulty open-ended responses coding examples (ProdS – Week 13)

Test difficulty sub-category	**Code**
Giving details (reasons/examples)	A
English Vocabulary	B
Expressing ideas in English	C
Short discussion time	D
Difficult topic content	E

Written survey response	Code	Coding comments	Spoken interview response	Code	Coding comments
It was hard to think of reasons and examples[A] to match the situation and topic.	A	Is this about the topic (E) or coming up with reasons and examples (A)?	Student: *It's difficult to think of reasons[A] for what I say to my group. Sometimes it's very difficult.* Interviewer: *Yes. That's right. The English is too difficult for you?* Student: *Sometimes, but in general I can't think of reasons or details. I just don't have any ideas. Maybe it would be the*	A B	Coming up with reasons (A) and having a lack of vocabulary (B) clarified by the interviewer as difficulties reported by the student.
I'm poor at thinking of reasons and examples[A].	A	A problem with L2 or coming up with reasons and examples?			
I found it easy to say my opinion, but difficult to give reasons[A] to support it.	A	A problem with using English or the actual content for reasons?			
It was hard to give reasons[A] for my choices.	A	Clearly saying that having reasons themselves was hard, rather than referring to an English problem.			

It was hard to come up with concrete reasons for my opinions[A].	A	A clearer example of issues reported with coming up with reasons, rather than English problems.	*same in Japanese. But I also don't have a lot of English words[B], so I need to practise those more.*		
I found it hard to say what I wanted to in English[C]. I had a lack of vocabulary[B] and I could not think of reasons[A].	A B C	Student explains A and B as reasons for C. This is clearer than some other students.	Student: *I can't say what I want to in discussions. I can't think of what to say.* Interviewer: *I see. It's difficult to do it isn't it.* Student: *Yes. My group members can do it, but I can't. It's too difficult sometimes to say the words in English[B].*	B	Interviewer does not really helps clarify what the issue is believed to be (coming up with reasons (A), vocabulary (B) or articulating the language (C) for example).
My lack of vocabulary[B].	B	Clear example of a student saying that vocabulary was the issue for them			
I could not think of the English words[B] to use.	B	Do they mean having a lack of vocabulary or not being able to come up with reasons?			
I found it difficult to express my thoughts in English[C], because I have limited English vocabulary[B]. I could not have a smooth conversation like I can in Japanese.	B C	Clearly referring to a lack of vocabulary (B) and articulation (C) difficulties.			
It was hard to say want I wanted to in English[C] and I had to use gestures, because I had a lack of English words[B].	B C	Referring to a lack of vocabulary (B) as a problem and ability to say things in English (C), but due to the lack of vocabulary or something else?			

I found it difficult to use my English to communicate my points[C] to the other members	C	No reasons stated why it was difficult.	Student: *I think we need more time to speak*[D]. Interviewer: *The time is too short?* Student: *Yes. If we had a little bit more time, we could give more opinions, so we could get more points.*	D	Interviewer helps clarify that the length of the discussion (D) is seen as a difficulty, but does get any more detail afterwards.
Saying what I wanted to in English[C] in that moment was difficult.	C	Referring to the ability to articulate into L2, or the pressure of doing it quickly enough?			
I could not say the things I wanted to in English[C].	C	More of a clear example of articulation issues? Very general comment (used a lot in surveys) which did not reveal too much about student feelings!			
I knew what I wanted to say each week, but it was difficult to say it well in English[C].	C	Clearer example of articulation problems with L2 (knew what to say, but couldn't in L2, so not a reasons problem!)			
The time limit was too short[D].	D	Made it hard to say things in English or come up with ideas?			
Some topics[E] are much harder than others.	E	Clearly indicating that perceived difficulties were			

		related to the topics themselves.	Student: *I can't speak sometimes. It's too difficult.* Interviewer: *Right. I see. Why do you think so?* Student: *It depends on the topic. If it is too hard to have an opinion about then I just stay quiet. I don't think that's my fault. I think the topic is too hard*[E].	E	Interviewer clarifies that the student clearly feels their performance problems are sometimes down to the topic.

Appendix P. Student GSF sheet usefulness open-ended responses coding examples (ProcS – Week 13)

Sheet usefulness sub-category	**Code**
Could see my weak/strong points	A
Easy to see my performance level each time	B
Can see our score for each category	C
Helped me know what to say	D
Motivated me to aim for new targets	E

Written survey response	Code	Coding comments	Spoken interview response	Code	Coding comments
I could see my weak and strong points[A].	A	Clearly about seeing own strengths / weaknesses.	Student: *I like the sheet because you can see what you are not good at*[A]. Interviewer: *I see. For example?* Student: *I could ask a lot of questions, but I forgot to give reasons. I could see a low reasons score, so it was very useful. Some areas needed practice*[C] *and I could see the levels each week*[B]*, so I could compare them.*	A B C	Interviewer uses a good eliciting question to clarify that the student refers to weak/strong points, seeing performance each time and feedback on the difference categories.
I understood my weaknesses[A].	A				
I could see how well I could speak every class[B] by knowing my strong and weak points[A].	A B	Mentions both seeing ability each time and seeing weak and strong points. Clearer than the two students above.			
It was easy to think about what to say next time[D], because you could see which different areas[C] you were doing well at and badly at[A].	A C D	Student reports that knowing own weaknesses and knowing what to say were difficult. Are these the same thing, or are they appropriately separated as two different points referring			

It helped me see where I was lacking points[A] and to know what to say[D].	A D	to clarity of performance and clarity of focus?	Student: *The sheet helped me know what I should say[D] each time.* Interviewer: *What do you mean?* Student: *I could see the things I was bad at[A], so it was clear that I wasn't asking enough questions, or giving enough reasons sometimes for example.*	A D	Interviewer uses a good eliciting question to clarify that the student refers to weak/strong points, as well as being helped in knowing what to say by the sheet.
We were very motivated to get our next goals[E] because we could see what we were bad at[A].	A E	About motivation due to seeing weak points. Perhaps other students are motivated in this way, but just do not mention it in their responses.			
It was motivating[E] because you wanted to say what the sheet showed you that you could not[A] in previous discussions.	A E				
I could see my score each time[B].	B	It is not clear why this was helpful. Needed to be explained more by the student.			
It was useful for reviewing later because you could see your results visually for different discussions[B] which you cannot do while you are having a discussion.	B	Clearly about seeing your ability, but do they mean in general, or specifically about seeing weaknesses? These are both under the same category in the results tables to eliminate doubt.			
It was easy to see my scores each week[B].	B				

I could see how well I was performing each time[B].	B		Student: *There were some good points for the sheet. For example, I could see what my scores were for the different sections[C]. I felt really motivated to try and do better at the goals for my weak areas[E], because I could see them on the sheet[A] and my strong areas did not need much focus.* Interviewer: *I see.* Student: *Yes. The sheet was good. I think it motivates a lot of students because they can see their overall score each time[B].*	A B C E	The student explains multiple points and the interviewer actively listens to elicit more detail well.
I could see my score on each new sheet[B.] This made me want to improve and get closer to my goals[E] each time.	B E	Says that seeing the score created motivation. A bit clearer than other comments above.			
It was useful to see what you should be saying[D] in a discussion.	D	Do the students mean because of the weak points they see for themselves or seeing their performance each time?			
It helped me understand what kind of things I should say[D].	D				
It made me try hard to reach my goal[E] each time.	E	Talk about motivation because of the goals, but is it the presence of the goal or trying to beat it which really motivated them?			
When I used the sheet, I tried hard to beat the score which I got last time and reach a new target[E].	E				

Appendix Q. Student GSF diary usefulness open-ended responses coding examples (ProdS – Week 13)

Diary usefulness sub-category	**Code**
Easy to see my performance level each time	A
Goals were motivating to do better each time	B
Seeing improvements made me happy	C
Could see goals clearly	D
Did not use English I used in future discussions	E

Written survey response	Code	Coding comments	Spoken interview response	Code	Coding comments
We could see our scores each time[A].	A	Clearly stating that the performance feedback is the helpful point, but not exactly why.	Student: *It was helpful to see your overall score each week*[A]. Interviewer: *Why?* Student: *Because it was easy to see how good our group was at the discussion.*	A	Interviewer doesn't elicit details about reasoning very well.
It was easy to see the results each week[A].	A				
We could see how much score our group had for each discussion[A].	A				
You can see how well you do each time[A].	A				
By setting a goal, I tried harder[B] to reach for higher scores.	B	Having or setting a goal was reported as the clear reason for motivation	Student: *I liked having goals in the diary.*	B C D	Interviewer elicits details about the goals being

All group members were motivated by having a goal[B] for next time.	B	(B), but is not explained any further.	Interviewer: *Really? Why?* Student: *They motivated me[B]. They made it really clear to us what we needed to do[D] and we all felt really happy when we could see our scores getting better[C].*		present and clear, as well as seeing improvements over time as sources of motivation.
Setting goals for the next discussion made us try to get them[B] during that discussion.	B				
Setting a goal for next time motivated me[B] to try hard.	B		Interviewer: *Was the diary helpful?* Student: *Yes, but I couldn't see how to use the English we were practising again in the next discussion. It was very difficult.* Interview: *I see. Why do you think so?* Student: *I couldn't remember it when we were talking. It is hard to use it in the next discussion[E]. The diary wasn't useful for this point.*	E	Interviewer elicits a criticism of the diary about not being able to transfer practised English to the next discussion (E). None of the students mentioned this in the surveys, but interviewed students did.
It was nice to see your scores increasing over time[C].	C	Do the other students mean this, but did not specify motivation?			
It made me feel good when I got a better score[C].	C	Improvement in scores overtime reported as the main reason for satisfaction.			
We were happy when we could see points getting higher[C].	C				
You can see your future goals easily[D].	D	Seeing what the goal was for next time reported as the reason.			
It was easy to see our goals[D] for next time.	D				
After seeing our result, we could see what our goal was[D] for next time.	D				