Task Parallelness: Investigating the Difficulty of Two Spoken Narrative Tasks

Chihiro Inoue
(MA in TEFL, BA in Linguistics)

A thesis submitted for the degree of PhD

Department of Linguistics and English Language
Lancaster University

April 2011

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Declaration

I hereby confirm that this thesis represents an original piece of work on my part. It has not previously been submitted for a higher degree elsewhere. A paper is published out of Pilot Study 3 (Chapter 3, Section 3.4) in the proceedings of the Lancaster University Postgraduate Conference in Linguistics and Language Teaching 2009. This is given in the bibliography at the end of this thesis.

Chihiro Inoue, April 2011

Task Parallelness: Investigating the Difficulty of Two Spoken Narrative Tasks

A thesis submitted for the degree of PhD

Department of Linguistics and English Language, Lancaster University

April 2011

Chihiro Inoue (MA in TEFL, BA in Linguistics)

Abstract

This thesis explores how task parallelness might be established; this is of fundamental importance to any discussion in the areas of language testing and task-based research, where the equivalence of tasks is a prerequisite. Five pilot studies were conducted using two spoken narrative tasks from an ongoing speaking test of English in Japan, the Standard Speaking Test, including two feasibility studies using several linguistic variables to analyse candidate performances, a study of expert judgements of the two tasks, a study of the linguistic performance of native speakers of English, and a study to identify an appropriate pair of tasks for the main study. The main study examined the parallelness of two spoken narrative tasks by Hill (1960) in terms of the ratings calculated by MFRM analysis, the linguistic performances of 65 Japanese candidates and 11 native speakers of English, expert judgements by Japanese teachers of English, and perceptions of the Japanese candidates and native speakers of English. The validity of the linguistic variables was also examined. The results of analyses demonstrated that the two tasks were not actually parallel, despite the effort to ensure a priori parallelness via the pilot studies. The findings were extensively discussed in relation to the theories of task complexity from Robinson (2001) and Skehan (1998), and raised several questions regarding the variables for quantifying the accuracy and syntactic complexity of linguistic performance. Taken together, the findings of this thesis add significantly to

the understanding of task parallelness and the results of my work can be applied not only to the design and selection of tasks but also to the investigation of linguistic performance in the fields of language testing and task-based research.

Acknowledgements

For all my life, I have struggled to find confidence in myself. Despite my long-term wish to be a high-achiever, I had never made the best effort to be one, for fear of failure, always leaving some room for excuses to save my pride, saying, "well, my achievement wasn't very high, but that's because I didn't try very hard." I had been running away from something I should have confronted much earlier, however, I had also been looking for the right project to devote myself to fully, something that I would not be afraid to give my best to. Researching and writing a PhD thesis at Lancaster gave me such an opportunity to face my past and to test my ability to meet the rigorous intellectual challenges. Now, having gained so much from my PhD, I finally feel I have more confidence and have come to better terms with myself than I ever did.

This is all thanks to Lancaster. I am most indebted to Dr. Judit Kormos, who has been an excellent supervisor through good and bad, giving me constant support and encouragement with enormous patience. With her professionalism as an academic and devotion to her family as a wife and mother, Judit is my super role model and someone I will always try to learn from. My heartfelt thanks also go to Prof. Charles Alderson, who was my co-supervisor for the first two years of my PhD. His critical reading and thinking as well as his vast knowledge across the subject gave me a solid foundation for my thesis.

The members of the Language Testing Research Group and Second Language Learning and Teaching Research Group at Lancaster deserve my sincerest appreciation. I could not have completed my thesis without their cooperation in my pilot studies and rating workshops as well as their questions and suggestions on the earlier pieces of my work. I would like to specifically thank my raters, Zahra Al-Lawati, Karen Dunn, Tania Horák, Janina Iwaniec, Gareth McCray, Geoff Shaw-Champion, Hiroko Usami, and

Lynn Wilson, for sparing their precious time out of their MA and PhD research.

I am also grateful to the teachers, students, and senior colleagues in Japan who voluntarily participated in the various phases of data collection. I would like to specifically thank Prof. Masashi Negishi, Prof. Asako Yoshitomi, Prof. Yukio Tono, Dr. Naoyuki Naganuma, Mr. Yoji Kudo, and Mr. Naoyuki Kiryu, for generously sparing their class time to advertise and conduct surveys for my study. The fact that I was able to receive great support from my home university has been an enormous encouragement. My sincerest gratitude also goes to Mr. Hirano at ALC Press, who granted me permission to have access to the recordings data of the NICT JLE Corpus. It was very fortunate for me to have it on top of the test materials of the SST, without which this thesis would not have been born at all.

Without my award of a Scholarship for Overseas Degrees from the Tokyo University of Foreign Studies, funded by the Japanese Ministry of Education, Culture, Sports, Science and Technology, this thesis would never have been possible. Being able to concentrate on studying without worrying about money has been a great luxury. I am also grateful for the British Council Japan Association Scholarship, BAAL UK Student Scholarship and William Ritchie Travel Fund for their financial support in data collection and conference presentations.

I cannot thank my Japanese family enough for their love, support, and understanding all through the way. I hope I have made a daughter and a sister that they can be proud of. Last but not least, I would like to thank my fiancé, Rob, for his constant support and encouragement. On top of his excellent meals and generous lifts to campus, his unwavering love and faith in me kept me going. He has always reminded me that life is still beautiful in difficult times, and without him, I would never have coped with my PhD path so well.

Contents

Tables

Figures

Appendices

Chapter 1: Introduction

1.1. Rationale of This Thesis

Achieving high English language proficiency has become of an increasing importance in Japan as it is considered indispensable for Japanese in order to survive in today's globalising world where English is used as a common international language (The Ministry of Education, Culture, Sports, Science and Technology [MEXT], 2003). In response to this situation, the MEXT has launched a large-scale action plan for better English education in 2003 which aims to improve its Course of Study as well as curricula, teaching methods, and teacher training, and to promote international exchange programs in high schools so that Japanese will acquire more communicative English proficiency with stronger productive skills, especially speaking. Accordingly, it is of no doubt that there needs to be test tasks which can reliably measure the English speaking ability of Japanese learners.

What is vital for reliable English proficiency tests is to have equivalent forms, i.e. comparable test versions to give to a number of candidates over years so that meaningful comparison of scores is possible while maintaining test security. Nevertheless, establishing evidence of equivalence among different test forms or at the task level, especially in productive tests, are seldom provided or carried out at all by test administrators (Weir, 2005: 250), which seriously threatens not only reliability but also validity and fairness of the tests. Moreover, the same problem applies to previous studies on task complexity in task-based research, where equivalence of tasks is a prerequisite but seldom demonstrated (Weir & Wu, 2006). This issue clearly deserves further exploration. Focusing on narrative tasks which are frequently used in English tests in Japan, this thesis intends to explore how evidence of 'parallelness' of speaking

tasks might be established and to examine which variables can be used in establishing the evidence. In turn, it is hoped that a better understanding of task design for producing more reliable speaking tests can be achieved.

The rest of this introductory chapter is structured as follows. The definition of spoken narrative is given (1.2.1), and a brief overview of the use of spoken narrative tasks in related research fields of language testing and task-based research follows (1.2.2). Subsequently, the terms used in this thesis are described (1.2.3), and the structure of this thesis is introduced (1.2.4).

1.2. Introduction to Spoken Narrative Tasks

1.2.1. Definition of Spoken Narrative

According to Labov (1972: 360), a narrative can be minimally defined as "a sequence of two clauses which are temporally ordered". Based on the analysis of hundreds of stories told in natural conversation by informants from various backgrounds, Labov identified six core features of a more fully-developed narrative: *abstract* (summarising the story briefly before the narrative begins), *orientation* (setting the time, place, characters and situation), *complicating action* (telling the events in the story), *result or resolution* (telling what happened at the end), *evaluation* (indicating the point of the story), and *coda* (concluding the narrative) (Labov, 1972: 363-70). Labov's framework has been highly influential in the field of sociolinguistics (Holmes, 2003: 118) and is also frequently cited by studies of second language narrative development (e.g. Liskin-Gasparro, 1996; Verhoeven & Strömqvist, 2001; Montanari, 2004).

Whilst naturalistic data (i.e. obtained from everyday language use) are collected by sociolinguists, the narratives in second language development are often elicited artificially by prompts such as silent-movie clips and picture books. Spoken

narrative tasks, which refer to sequences of a small number of pictures (i.e. 4, 6 or 8 pictures in this thesis), can be classified as one such elicitation prompt for a narrative. In particular, in the fields of language testing and second language acquisition, spoken narrative tasks are administered in order to elicit a relatively long monologue so that the language elicited can be of a certain length that can provide an adequate sample of performance. The next section briefly reviews the use of this task type in these two fields of research.

1.2.2. Spoken Narrative Tasks in Language Testing

In language testing, spoken narrative tasks refer to tasks based on picture sequences that candidates are asked to describe orally in one time frame (Luoma, 2004: 144). More specifically, Luoma noted that candidates should demonstrate their control over the following essential features of a narrative: setting the scene, identifying the characters and referring to them consistently, identifying the main events, and telling them in a coherent sequence. In the light of the narrative features by Labov (1972) mentioned above, candidates may include *orientation*, *complication action*, and *resolution* in a coherent manner in their narration.

Often, criticisms are made of spoken narrative tasks in tests for their lack of authenticity; it is almost impossible to imagine a real-life situation where a person has to tell a story based on a picture sequence. Nevertheless, the use of this task type may be defended on the ground that narrative is a part of the information routine of reporting, which is a common type of discourse in everyday life (Cyril James Weir, 2005: 148-149). Besides, in exchange for the lower authenticity, constraining the content of narration by pictures can lead to higher reliability. As the pictures control the content of the story for all candidates, so comparisons of performances can be relatively

unaffected by background or cultural knowledge, provided that the pictures used are culturally unbiased (Cyril James Weir, 2005: 148). In addition, this task type is well suited to lower-level candidates because "telling simple stories is one of the first things that they are able to do in a second language" (Fulcher, 2003: 70).

To benefit from these advantages, there are a number of speaking tests which utilise narrative tasks, for example: Test of Spoken English,[1] English Language Skills Assessment,[2] and Test in Practical English Proficiency.[3] However, little evidence of the comparability of narrative tasks in their different test versions can be found in published research. Moreover, the comparability of different test versions is seldom demonstrated by testing organisations (Cyril James Weir & Wu, 2006: 169), although it is vital for any language test to ensure meaningful comparisons of scores across a number of administrations whilst maintaining test security. This issue is discussed further in Section 2.3.

1.2.3. Spoken Narrative Tasks in Task-based Research

In the field of second language acquisition, especially in task-based research, spoken narrative tasks are of "well-established and frequently researched task type" (Albert & Kormos, 2004: 286). A number of researchers have utilised them in order to examine the effects of manipulating task administration conditions and/or task characteristics on candidates' performance, including Robinson (2001), Skehan and Foster (1999), Bygate (1999), Ortega (1999) and Yuan and Ellis (2003), to name but a few. These studies are part of an effort to justify and assist in the pedagogic use of tasks

---

[1] Administered by the Educational Testing Service.

[2] Administered by London Chamber of Commerce and Industry Examinations Board.

[3] The most well-known English proficiency test in Japan (The Society for Testing English Proficiency, 2010).

by determining how certain task characteristics affect L2 performance, so that teachers can decide which tasks to implement in their classrooms according to their teaching goals (Skehan, 1998: 97). The underlying rationale for this line of research is that tasks with certain characteristics and/or administration conditions will impose varying processing loads, which may then direct the attention of L2 learners to different aspects of language use (see Section 2.5.1 for further review).

In order for the results and implications of these studies to be valid, it is obvious that the tasks used in a study must be comparable, except for the particular task characteristics or conditions in question. Otherwise, any differences observed in performance cannot be credibly attributed to the task characteristics or conditions in question; they may have been caused by unintended and uncontrolled inherent differences between tasks. Nevertheless, very few such studies have provided evidence of the comparability of tasks beforehand (Cyril James Weir & Wu, 2006: 169). In fact, many of them do not reveal the actual tasks or the source of where the tasks were obtained. The lack of this important piece of information, as well as the lack of comparability evidence of tasks, can cast doubts on the reliability and validity of the findings of such research.

1.3. Terminology

When discussing comparability, several different terms are used by different researchers. It appears that 'comparability' is comprehensive (e.g. Bachman, 1990: Luoma, 2004) and that there are two terms to refer to the comparability of test forms (or tasks): 'parallelness' and 'equivalence'. Although these terms are sometimes used interchangeably (e.g. Weir & Wu, 2006), Alderson, Clapham and Wall (1995: 96) drew a distinction between the 'parallelness' and 'equivalence' of language tests. Being

'parallel' refers to being designed to be as similar as possible, including the same instructions, response types, number of test items and, ideally, the same content (J. Charles Alderson, Clapham, & Wall, 1995). Alderson et al. (1995: 96) further added that if the same candidates took parallel versions, then their scores should yield the same mean and standard deviations. Since this is impossible to achieve in real-life testing situations,[4] many testing organisations develop 'equivalent' versions that are based on the same specifications but differ in the number of test items, response types and content (Alderson et al., 1995: 96-97).

The term 'task difficulty' is used in this thesis to refer to the logit values for tasks calculated by MFRM analysis. Where there is a need to use this term differently so that concepts employed by other researchers can be appropriately introduced, it is clearly explained, for example, as 'task difficulty in the framework by Robinson (2001)' and 'perceptions of task difficulty'. The terms for other characteristics of tasks are introduced and defined in Sections 2.5.1.3 and 2.5.2.

1.4. Organisation of the Thesis

This thesis consists of seven chapters. Chapter 2 presents the review of relevant literature in language testing and task-based research. Drawing on the frameworks of validity by Messick (1989; 1996) and of contextual factors in speaking assessment by McNamara (1996), Skehan (1998) and Bachman (2004), the aspects of spoken performance to be examined and controlled for in order to establish parallelness are identified. Then, by reviewing previous studies on equivalence in language tests and theories of speech production, attention and task-related factors, the operationalisation

[4] Alderson et al. note that, for this reason, the use of parallel tests may be limited to experimental reliability studies (1996: 96).

for the variables is sought. It concludes that parallelness should be evidenced in terms of ratings adjusted by MFRM analysis, perceptions by the candidates and native speakers of English, expert judgements, and the elicited narrative performance characterised in the areas of fluency, accuracy, complexity, and idea units. Chapter 3 describes a series of pilot studies based on tasks from a speaking test in Japan. This chapter provides vital methodological implications for the main study after trialling several variables, collecting expert judgement and native speaker performance data, and selecting appropriate tasks for the main study. The need was also identified to conduct a validation study of the variables to examine the elicited performances. Chapter 4 presents the methodology of the main study, and reports on the instruments, rater training, and methods of analysis in detail. Chapter 5 shows the results for the research questions which address the aspects of spoken narrative performance that have been examined. Chapter 6 discusses and synthesises the findings in the light of relevant literature. Chapter 7 considers the implications for the design of spoken narrative tasks and the implications for the theories of task complexity, in addition to outlining the limitations of this thesis.

Chapter 2: Review of Literature

2.1. Introduction

In this chapter, drawing on the literature in the fields of language testing and task-based research, relevant previous research is reviewed for the purpose of identifying what needs to be considered as evidence of 'parallelness' in spoken narrative tasks. The first half of this chapter mainly handles previous studies in language testing research, discussing the relevant aspects of validity and contextual factors of speaking assessment that should be controlled for (Section 2.2), and related research on the equivalence of test forms and tasks (Section 2.3) along with the methodological implications for this thesis. The latter half of the chapter summarises task-related research, and explores how relevant aspects of validity can be operationalised in this thesis. It includes reviewing models of speech production (Section 2.5.1), discussing relevant task characteristics (Section 2.5.2) and linguistic variables to examine different aspects of narrative performance, such as fluency and accuracy (Section 2.5.3), as well as task-specific variables (Section 2.5.4). Reviewing the variables for linguistic performance leads to the selection of appropriate rating scales (Section 2.6) for candidates' performance. Finally, the research questions are presented at the end of the chapter (Section 2.7).

2.2. Theoretical Frameworks

This section reviews the theoretical frameworks that are indispensable for this thesis. Firstly, it explains the Messick's (1989, 1996) model of validity which is a generic model applicable to all types of educational assessment and which has indeed been most influential in conceptualising the evidence needed for good language tests.

Secondly, with a view to demonstrating the evidence for the parallelness of spoken narrative tasks, models of speaking assessment by McNamara (1996) and Skehan (1998) are introduced and discussed.

2.2.1. Validity

Discussing validity in language testing means ascertaining whether or not a particular test measures what it is intended to measure (Lado, 1961: 321). What language testers intend to measure by their tests is a *construct* which is "a theoretical conceptualisation about an aspect of human behaviour that cannot be measured or observed directly" (Ebel & Frisbie, 1991: 108). Examples of constructs include intelligence, motivation, anxiety, attitude and reading comprehension. Messick (1996) offered a widely accepted definition of construct validity in educational assessment as follows:

> Validity is an overall evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of interpretations and actions based on test scores or other modes of assessment. (Messick, 1996: 245)

Accordingly, Messick identified six aspects of validity that should be evidenced to support a validity argument for a test: content, structural, external, consequential and substantive validity, and generalisability.

The content aspect refers to the relevance and representativeness of the test content in relation to the construct, and is examined by expert judgement and the language elicited as to whether this shares the same characteristics with the language used in real-life situations. The structural aspect questions if the elicited dimensions of candidates' performance match the construct assumed by testers. The external aspect is

also known as concurrent validity, and is often investigated by correlating the scores with another test of the same construct (or of a different construct to show distinctiveness). The consequential aspect, also known as washback, often examines, qualitatively and longitudinally, the effect of a test on society, i.e. not only on candidates but also on all stakeholders (e.g. teachers, admission officers, employers, publishers, textbook designers, etc.).

Although the four aspects of construct validity introduced above are essential when evaluating language tests that consist of a number of items or tasks, the two remaining aspects (substantive validity and generalisability) are of key importance and relevance to the discussion at the task level, which is the focus of this thesis. The substantive aspect concerns whether candidates go through processes, when completing a test (or task), in a way that corresponds to the hypothesised construct (Chapelle, 1999: 262); this has been one of the most difficult aspects to investigate (Fulcher, 2003: 195) because of the methodological difficulties in capturing what is going on in a candidate's mind during test-taking. O'Loughlin (2001), however, suggested ways to demonstrate substantive validity evidence in speaking tests by observation, questionnaires and interviews of candidates; his study is reviewed in detail in Section 2.3.3.

The sixth aspect of validity is generalisability, which is often researched by examining candidates' elicited language under different test conditions and test task characteristics in order to explore how generalisable performance on the test (or task) is. Previous studies which have looked at this aspect of different test versions for equivalence are reviewed in Section 2.3. In addition, although under different frameworks, this aspect is extensively researched within the field of task-based research, which are reviewed in Section 2.5.

Thus far, the six aspects of validity defined by Messick (1989, 1996) have been

briefly described. Now, the review of literature proceeds to define construct, based on which language testers attempt to collect evidence to argue for their tests' validity.

2.2.2. Models of Speaking Assessment

When hypothesising construct, one needs to consider what constitutes a person's language proficiency. The model which is most frequently referred to in the current field of language testing is Bachman and Palmer's (1996) model of communicative language ability (Luoma, 2004: 97). This model is based on the work of Bachman (1990), who reorganised the components of one's communicative competence, drawing on earlier frameworks by Hymes (1972), Canal and Swain (1980) as well as an empirical study by Bachman and Palmer (1982).

Bachman and Palmer's (1996) notion of language ability includes constituents of competence, such as knowledge about the language (*language knowledge*) as well as the capability to implement the knowledge for use (*strategic competence*). Language knowledge includes organisational (grammatical and textual) and pragmatic (functional and sociolinguistic) competencies. Strategic competence is a collection of dynamic strategies (goal-setting, assessment and planning) which are utilised when one engages in communication: estimating the task goal and planning what to say and how to say it, while drawing on necessary language knowledge as well as topical knowledge to complete the task.

The concept of language ability is a primary part of *candidate characteristics* which also incorporate personal characteristics (such as gender, age, L1 and L2 proficiency), topical knowledge, and affective schemata (i.e. emotional attitudes to the topic of a task). Bachman and Palmer (1996: 62) argue that performance should be understood as resulting from a complex interaction between candidate characteristics

and task characteristics, as these two sets of characteristics are considered to affect performance greatly. This discussion is revisited in Section 2.5.1.3.

While Bachman and Palmer's model has contributed immensely to conceptualising an underlying structure of language proficiency (Luoma, 2004: 101), it has been criticised for focusing too much on the individual candidate (Chalhoub-Deville, 1997: 5). McNamara (1996) drew our attention to the contextual factors that influence a candidate's score or rating in speaking assessment. In addition to the task (characteristics) that Bachman and Palmer (1996) noted, McNamara listed not only the test tasks, but also interlocutors, rating scales and raters as additional elements of contextual factors, as summarised in Figure 2.1, below.

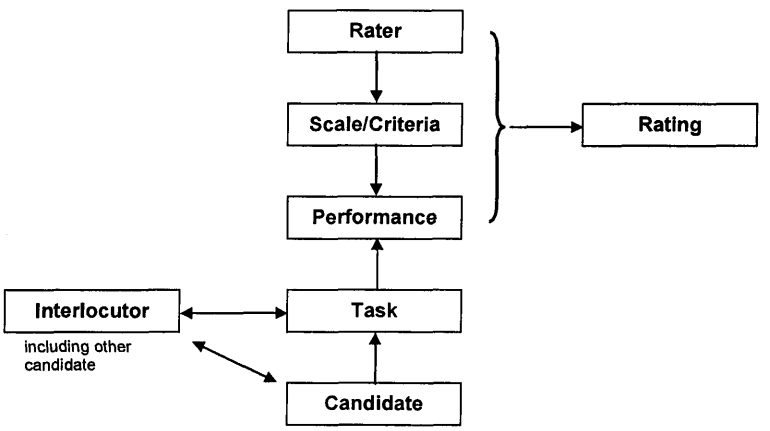


Figure 2.1
Contextual Factors in Speaking Assessment (from McNamara, 1996: 86)

Starting from the bottom in Figure 2.1, a *candidate* speaks to or with an *interlocutor* (or *interlocutors* in the case of a paired or group oral test) on a test task. The *performance* elicited by the task is rated according to the *rating scale(s)* or *criteria* by trained *raters*, who finally produce a final *rating* or score for the candidate. McNamara's model has been a very influential framework when organising research

(Skehan, 1998: 170), which has led to numerous studies on how different contextual factors may influence spoken performance. Such studies have researched the effects of, for example, different candidate characteristics such as gender (O'Sullivan, 2000), personality (Berry, 2004), interlocutors (A Brown, 2003), tasks (Fulcher, 1996b) and raters (Weigle, 1998), to mention just a few.

While recognising the influence that McNamara's model has had in the field of language testing, Skehan (1998) nonetheless argued for its further expansion in order to account for how individual candidates engage with performing a task. Skehan divided task factors into *task qualities* and *task conditions*, and incorporated *competence* and *ability for use* as the two factors that influence a candidate, as presented in Figure 2.2. His notion of ability for use "goes well beyond the role of strategic competence [i.e. by Bachman and Palmer (1996)], and draws into play generalised processing capacities and the need to engage worthwhile language use" (Skehan, 1998: 171).

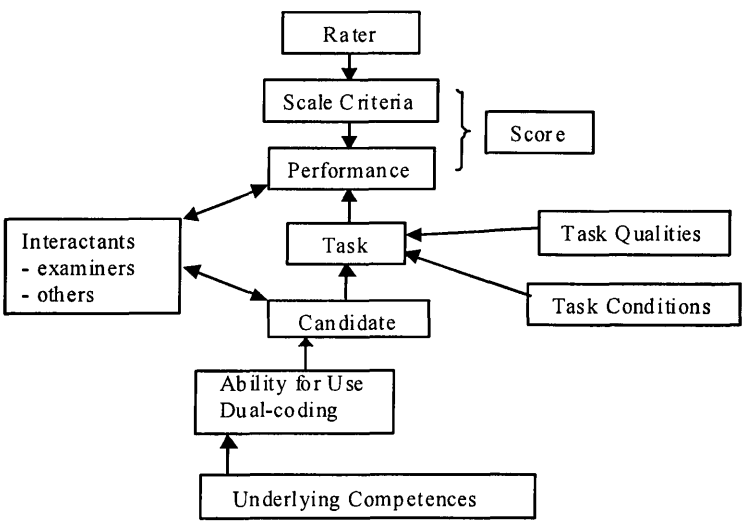


Figure 2.2
An Expanded Model of Speaking Assessment (from Skehan, 1998: 172)

More recently, Bachman (2002) has modified Skehan's model with emphasis on more dynamic interaction among the contextual factors, and reorganised candidate factors and task factors, as shown in Figure 2.3. Bachman argues that:

> Candidates, who will differ in their underlying competencies and ability for use, may find tasks with different qualities and conditions differentially difficult to perform. Different candidates will find different examiners and other interactants differentially easy or difficult to interact with. Different raters may apply the scale criteria differently to different performances, so that they may be differentially lenient or severe. (Bachman, 2002: 466)

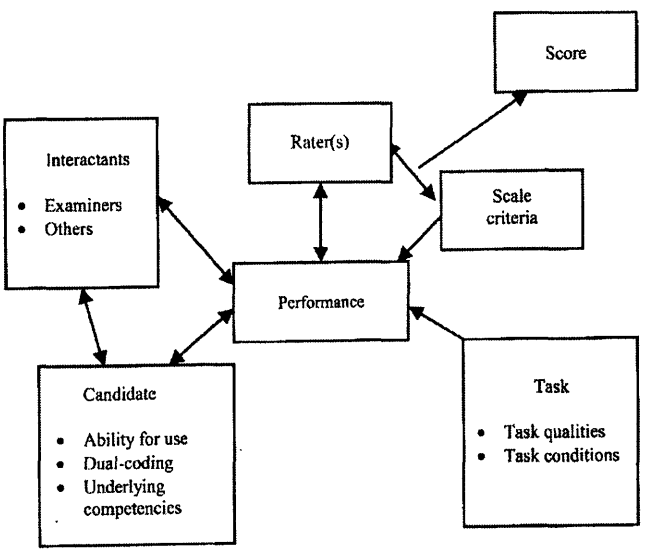


Figure 2.3
Bachman's Model of Interacting Factors in Speaking Assessment (2002: 467)

With such complex interactions between the factors which influence spoken performance and scores or ratings, it is evident that the factors must be strictly controlled for if the parallelness of tasks is to be investigated; the tasks should be

administered by the same interviewer on the same candidates whose performances are then rated by the same raters using the same rating scales. Regarding task factors, Bachman (2002: 469) recommended conceptualising tasks as sets of characteristics and clearly distinguishing between the features inherent in tasks, the attributes of candidates, and the interactions between the characteristics of candidates and tasks. Task and candidate factors are discussed in detail in Section 2.5.1.3.

Thus far, this section has introduced the concept of validity and models of speaking assessment, demonstrating what needs to be investigated to establish task parallelness (i.e. the substantive and generalisability aspects of validity) and that strict control for contextual factors is indispensable. The next section addresses the concept of reliability and then reviews previous studies on the equivalence of tests and tasks, which have methodological implications for this thesis.

2.3. Equivalence of Test Forms and Tasks in Speaking Assessments

2.3.1. Reliability

Investigating the equivalence of two test forms or tasks addresses the issue of a test's reliability. Whilst validity, as explained earlier in Section 2.4.1, refers to whether a test measures what it is intended to measure, reliability focuses on whether a test undertakes measurements consistently. The method traditionally used for examining equivalence is parallel-form reliability[5] (Henning, 1987: 81-82), which is represented by a correlation coefficient between the scores of two test forms, given that the two forms display equivalent means and variance of scores, and correlate equally with a third measure of the same ability. However, it should be noted that parallel-form

[5] Although the name may seem to imply otherwise, using parallel-form reliability is not restricted to parallel forms but also applies to equivalent forms of a test.

reliability is not often applied to test speaking because the design of some speaking tests does not allow for it. An exception is the study by Weir and Wu (2006), which is reviewed and discussed in detail in Section 2.3.4.

Another type of reliability which is crucial when testing speaking is that of raters, whether they rate the same performances similarly and consistently. As discussed by Bachman (2002: 466), whose framework was introduced earlier in Section 2.2.2, raters and their behaviour in using rating scales can influence ratings to a large extent. Traditionally, inter-rater reliability has been widely used to ensure that the ratings given to candidates' performances are consistent, it is based on the correlation coefficient of the ratings given by two or more raters. However, as correlation cannot detect differences in rater severity, or differences in raters' interpretation and use of rating scales, more advanced measures to examine rater behaviour have recently come into use. Examples of such measures include G-study (Kenyon & Tschirner, 2000) and the Rasch model (Bonk & Ockey, 2003). The study by Kenyon and Tschirner (2000), which examined the equivalence of the scores of face-to-face and tape-mediated Oral Proficiency Interviews, is reviewed in Section 2.3.2.

The next three sections review previous studies which have been conducted on equivalence in speaking tests. The objects of comparison in such studies can be broadly classified into three categories: test forms across different administrations of a test, test forms of a test delivered by different modes (e.g. face-to-face vs. tape-mediated), and tasks of a test across different administrations.

2.3.2. Test Forms of a Test

The first line of evidence for test-form equivalence is based on the scores that candidates obtain from each test form. Bachman, Davidson, Ryan and Choi (1995)

investigated the equivalence of the speaking part of Cambridge FCE (Paper 5) using regression analysis as a preliminary to their research project to investigate the comparability of Cambridge FCE and TOEFL. Due to the paired oral test design that included several task types with variations, out of which the examiner was to choose one, each form of Paper 5 was quite different, making it impossible to use parallel-form reliability. Therefore, Bachman et al. focused on measurement errors that were identified by comparing the R-squared values with and without a test form factor. They found few differences in the scores due to the test forms, concluding that the FCE Paper 5 forms were equivalent. They reported that the board decided to follow their recommendations to ensure test-form equivalence for standard procedures (Bachman et al., 1995: 134); however, there appears to be no such evidence consistently made public today.

Another speaking test that utilises other measures to assert the equivalence of test forms is TOEFL iBT. Its Speaking section involves 6 different tasks, leading to "technical and test security constraints" (Educational Testing Service, 2008: 6) which do not allow for equation[6] to be conducted. Therefore, the only measures to maximise equivalence are taken at the task level: carefully developing tasks according to detailed task specifications, examining the means and variances of the scores on each task, and then correlating them with the scores of other sections in TOEFL iBT (Educational Testing Service, 2008: 6-7). The results of such basic statistics, however, are nowhere to be found at the time of writing this thesis. This is far from sufficient to demonstrate the equivalence of tasks.

Methodologically, a crucial weak point in these two studies lies in the fact that

---

[6] A statistical procedure using item-response theory to adjust the scores of different test forms so as to make them interchangeable. It is not suitable for performance tests which include only a few tasks. Kolen and Brennan (2004) offer a comprehensive explanation.

evidence of equivalence is based solely on test scores. As the next section will reveal, score equivalence is necessary but not sufficient for demonstrating test (or task) equivalence. What is more, even evidence of score equivalence is not often made public by ongoing relatively high-stakes proficiency tests, as has been repeatedly criticised by a number of researchers (e.g. Spolsky, 1995; Chalhoub-Deville & Turner, 2000; Weir & Wu, 2006). In fact, though some testing organisations claim that they examine the equivalence of test forms, the actual results of such investigations are not published. This seriously threatens not only the reliability but also the validity and fairness of a speaking test.

2.3.3. Forms of a Test by Different Delivery Modes

A second category of research on equivalence has examined test forms across different delivery modes. The tests which have played a primary role in this line of published research are the Oral Proficiency Interview (OPI) and its tape-mediated version, Simulated Oral Proficiency Interview (SOPI). As more researchers have been involved in the research and development of SOPI, a wider variety of evidence of test equivalence has emerged.

Clark and Li (1986) devised four forms of SOPI in Chinese and administered them to 32 students of Chinese at two universities, as well as four forms of OPI. The ratings were given by two raters (inter-rater reliability was over .90), who produced a very high average correlation of .93 between SOPI and OPI. Similarly, Stansfield (1990) also reported very high average correlations between OPI and SOPI in Portuguese, Hebrew and Indonesian, ranging from .90 to .94, with a high average inter-rater reliability of over .90. Although the numbers of candidates involved in the study were relatively small (n = 10 to 30), Stansfield claimed that SOPI was a valid

substitute for OPI. However, while a high correlation demonstrates linearity between the test scores of two different delivery modes, it does not necessarily mean that the two tests are equally difficult (O'Loughlin, 2001: 19). It is still possible that the candidates performed systematically better on one test than the other.

Kenyon and Tschirner (2000) disproved this possibility by finding no statistical mean difference between the scores converted from ACTFL proficiency levels of the German OPI and SOPI taken by 20 students. Furthermore, a G-study revealed zero score variance explained by test versions, demonstrating no consistent differences in the scores due to the different delivery modes. The two tests again produced a very high correlation of .96. As for rater reliability, the absolute agreement of the ACTFL proficiency levels was 90%. The results of G-study found that there was a larger variance due to raters on SOPI, but its variance for error and interaction between candidates and raters was smaller, which led Kenyon and Tschirner to conclude that SOPI might be both more reliable and economical.

Shohamy (1994), on the other hand, argued against over-reliance on concurrent validity, i.e. establishing test equivalence based on high correlations of scores to justify test substitution. Following a suggestion by van Lier (1989) to examine the speech samples from the OPI in order to find out what kinds of speech activities candidates were actually performing, Shohamy attempted a more comprehensive validation study of the Hebrew SOPI, via a comparison with the Hebrew OPI. Her research compared the two tests by *a priori* analysis of task characteristics such as the topics and functions they were intended to elicit, as well as by *a posteriori* analysis of candidates' linguistic performance in terms of error types and frequencies, use of communicative strategies, and discourse features such as lexical density (i.e. ratio of content words), functions, discourse markers, and so on. The concurrent validity of the

Hebrew SOPI (compared with the OPI) was very high ($r = .89$ to $.95$, $N = 40$). Out of the 40 candidates' performances, 10 transcripts were analysed for each test. The results showed that the SOPI tasks appeared to require more varied topics and functions in the *a priori* analysis. Contrary to this expectation, however, *a posteriori* analysis of the linguistic performance revealed that SOPI actually elicited fewer functions. In addition, there were more self-corrections and paraphrases, and more lexically dense 'literary' texts, which included more content words in SOPI, implying that OPI and SOPI may not be measuring the same construct.

The study by Shohamy (1994) cautioned against how expected responses (in test and task specifications) and actual responses by candidates can mismatch, and called language testers' attention to the need for more comprehensive validity evidence. Yet, three methodological issues should be noted. One is that it is not clear if the Hebrew OPI and SOPI were administered to the same candidates. If not, the differences found between the two tests might be partly attributable to the differences in candidate characteristics, especially with such a small number of transcripts. Second is the lack of reports on the reliability of coding of linguistic performance in terms of lexical density, hesitation, errors, etc., and on the coding schemes. Similarly, third is the lack of information on task types and contents, and whether any controls were attempted across the test forms being compared. Without such information, the credibility of the results may arguably be open to question.

These methodological shortcomings in Shohamy's (1994) study were better addressed in O'Loughlin's study (2001), which also employed a multi-method approach to examine the equivalence of face-to-face and tape-mediated versions of a test, the Australian Assessment of Communicative English Skills (***access***). O'Loughlin administered both tests to the same candidates in a counter-balanced design, describing

the coding scheme for classifying content words and non-content words (although not the reliability of the coding), and utilising matched topics for each type of task (description, narration, discussion, role-play). Because of the matched topics, the functions expected to be elicited were similar in *a priori* analysis of the tasks.

In *a posteriori* analysis, quantitatively, O'Loughlin calculated the difficulty of both tests with multi-faceted Rasch Modelling (MFRM) software, FACETS, in search of evidence of equivalent test scores, which yielded different difficulty values in some administrations. He also reported the existence of rater bias. Parallel form reliability was high (r = .81 and .94 in two administrations). Qualitatively, he attempted to compare evidence of generalisability and substantive aspects of test validity by means of analysing the linguistic performance of the candidates and the processes they underwent while taking the tests. O'Loughlin confirmed the results by Shohamy (1994) via more lexically dense performance in the tape-mediated version of the ***access*** exam. Observation and interviews revealed that the degree of nervousness while taking the tape-mediated version varied among the candidates. Questionnaire results showed that the majority of candidates felt more nervous about the tape-mediated version, and felt it to be more difficult. In conclusion, the face-to-face and tape-mediated versions of the access exam were suspected to have tapped into different constructs of speaking and should not be considered to substitute for each other.

The significance of O'Loughlin's study lies in the triangulated evidence of test equivalence, including the substantive validity evidence of candidates' processes and attitudes towards both delivery modes, which has now been examined by an increasing number of studies on computer-mediated versus face-to-face tests (e.g. Kenyon & Malabonga, 2001; Zhou, 2009; Qian, 2009). As O'Loughlin (2001: 168) noted, "the use of multiple methods ultimately yielded a clearer, more comprehensive answer to the

research question [about test equivalence] than would have been the case if only a single method had been adopted". This is an important implication for this thesis in that evidence of equivalence needs to be triangulated using different methods of data analysis for a comprehensive investigation into parallelness: the scores (or ratings), linguistic performances of candidates, and their test-taking processes.

Whilst O'Loughlin (2001) analysed only linguistic performance at task level, Weir and Wu (2006), whose study is reviewed in the next section, compared three test forms at task level as well as at test-form level, although using different types of evidence from O'Loughlin (2001). The next section discusses equivalence at the task level in order to appreciate further methodological insights to investigate the parallelness of spoken narrative tasks.

2.3.4. Tasks of a Test

Weir and Wu (2006) attempted to establish equivalence at both levels of test form and task for a tape-mediated speaking test in Taiwan, the General English Proficiency Test (GEPT) Intermediate Speaking Test, in which three monologic tasks are given (i.e. reading aloud, short-answer questions, and picture description). Three test forms were investigated, called Forms 1, 2 and 3 respectively, and each of the 120 candidates completed a common test form (Form 2) and another form (either Form 1 or 3). Ratings were given for each of the three tasks separately. Three trained raters were involved in rating a total of 240 tapes, using 60 tapes as a common batch so as to create connectivity in the ratings which would allow the use of MFRM (by FACETS), in addition to correlation, factor analysis and ANOVA for statistical analysis.

Qualitative analysis by Weir and Wu (2006) did not involve analysing any transcripts to compare features of the candidates' linguistic performance due to

practical constraints. Instead, they obtained expert judgements by 12 raters on specifically developed checklists for the three task types in order to investigate whether the difficulty of each task was perceived as the same. The checklists asked if the raters thought the tasks (of the same type in the three different test forms) were the same in terms of:

1) the assumed degree of candidates' familiarity with the topics;
2) the assumed degree of candidates' familiarity with the lexical and grammatical items in the input, as well as in the expected responses from the candidates;
3) the likelihood of the candidates completing the task within the time given.

Each statement was worded as follows: "The lexical items required to describe the pictures are equally familiar to candidates", and the raters were required to agree or disagree.

The three aspects above were intended to tap into the three factors which, according to Skehan's (1998)[7] framework, can affect the difficulty of tasks: code complexity (lexical and syntactic difficulties), cognitive complexity (familiarity and information processing), and communicative demands (time pressure). Weir and Wu assumed that if each task type in the three test forms was regarded as being the same in terms of these three aspects, then they were of the same difficulty, and were therefore equivalent. Additionally, for short-answer question tasks, a function checklist developed by O'Sullivan, Weir and Saville (2002) was added to examine if the expected functions were actually elicited in the candidates' performance (using transcripts of two highest rated samples). The results for parallel-form reliability of each task type were

[7] Skehan's framework (1998) is reviewed in more detail in Section 2.5.1.

low (r = .38 to .59) and were assumed to have been influenced by differences in rater severity (Weir & Wu, 2006: 179). All the task types on Forms 2 and 3 were equally difficult according to MFRM, ANOVA, factor analysis and rater judgement, but not Form 1.

Weir and Wu's study (2006) was innovative in attempting to establish equivalence at the task level, in employing various statistical methods, and in incorporating Skehan's (1998) framework from the field of second language acquisition (SLA) as a theoretical background for analysing task characteristics. However, a shortcoming is the lack of substantive validity.[8] In addition, the generalisability aspect of validity evidence (from the candidates' linguistic performance) was absent except for the functions elicited in short-answer tasks. The evidence for equivalence was largely dependent on the expectations of the raters for the candidates' language knowledge and process, which might not have reflected the candidates' actual performance or the process.

Two related studies, although not directly addressing task equivalence, have demonstrated the importance of examining the candidates' linguistic performance and their perceptions of what they had done and felt during task completion. Both of the studies examined the effects of manipulating three task administration conditions (i.e. the amount of planning time, scaffolding support during planning and response time) using monologic tasks in the IELTS Speaking Test. Horai (2009) analysed the scores and linguistic performance of 100 EFL candidates in terms of accuracy, complexity and fluency. The results showed a significant difference between the scores achieved in the un-manipulated (i.e. IELTS original) and the 'no planning' conditions. In addition, significant differences were found in all the three areas of performance by

[8] Weir and Wu (2006) referred to it as "cognitive validity".

proficiency level. High-proficiency candidates produced significantly more fluent, accurate and complex speech than the low-proficiency group. Furthermore, the fluency of mid-proficiency level candidates (described as 'borderline' group in Horai (2009)) appeared to be most affected by the changes of task administration conditions. These results clearly emphasise the value of investigating not only the scores but also linguistic performance of the candidates at different proficiency levels.

Investigating the effects of the same task administration conditions as Horai (2009), Weir, O'Sullivan and Horai (2009) analysed the IELTS scores using MFRM and the candidates' perceptions of what they went through during task completion using a questionnaire, which included statements such as "I thought of how to satisfy the audiences and examiners" and "I felt it was easy to produce enough ideas for the speech from memory". It was again found that the candidates scored the highest on un-manipulated version of the task. Moreover, the candidates showed significant differences in their perceptions at different proficiency levels. The lower-proficiency group did not find any version to be easier than the others, while the other two ability groups felt that they were able to perform the best on the un-manipulated version.

It is clear from the studies by Horai (2009) and Weir, O'Sullivan and Horai (2009) that a fuller investigation into task parallelness with transcripts of the candidates' linguistic performance and candidates' processes at different levels of proficiency is desirable.

2.4. Summary

Sections 2.2 and 2.3 have reviewed the concept of validity, models of speaking assessment, and previous studies on the equivalence of test forms and tasks. The point has been illustrated that:

1) establishing task parallelness involves seeking validity evidence;
2) evidence should be triangulated;
3) contextual factors of speaking assessment, such as examiners, raters and rating scales, should be strictly controlled for;
4) parallel-form reliability is not sufficient; the MFRM is suitable for handling ratings data which enables catering for rater severity, investigating rater behaviour, and considering task difficulty;
5) SLA frameworks may be useful for describing and investigating task characteristics.

To be more specific, evidence of parallelness should be demonstrated by investigating not only the ratings but also the substantive and generalisability aspects of validity, i.e. candidates' processes (assumed by experts as well as reported by candidates) and linguistic performance (expected and actual). By comparing the expected and actual responses for these aspects, *a priori* and *a posteriori* investigation for validity is achieved.

Moreover, since research into speaking tasks inevitably involves various contextual factors, it is ideal to make use of MFRM so that these factors can be considered when finalising the ratings and the difficulty of tasks. As the task is central to this thesis, describing and examining task characteristics are of crucial importance. Like Weir and Wu (2006), knowledge from the field of SLA should be incorporated. For investigating spoken narrative tasks, this is beneficial because there is abundant research on this task type and on the effects of manipulating certain task characteristics.

From this viewpoint, Section 2.7 reviews and discusses related SLA research

which introduces how speech production has been modelled, what kinds of processes are thought to be involved in it, and how researchers have operationalised and investigated the features of learner language. In so doing, it leads to identifying the research questions and necessary pilot studies for this thesis.

2.5. Operationalisation of the Evidence for Generalisability and Substantive Validity in Spoken narrative Performance

Following the review in the previous section, which clarified the importance of examining generalisability (i.e. candidates' linguistic performance) and substantive (i.e. candidates' process) aspects of validity for spoken narrative tasks, this section seeks knowledge of how these aspects are operationalised and investigated in the field of SLA. Section 2.5.1 reviews theoretical frameworks of speech production for first language (L1) and second language (L2), which provides a rationale for the operationalisation of the variables that are reviewed in Sections 2.5.3 and 2.5.4.

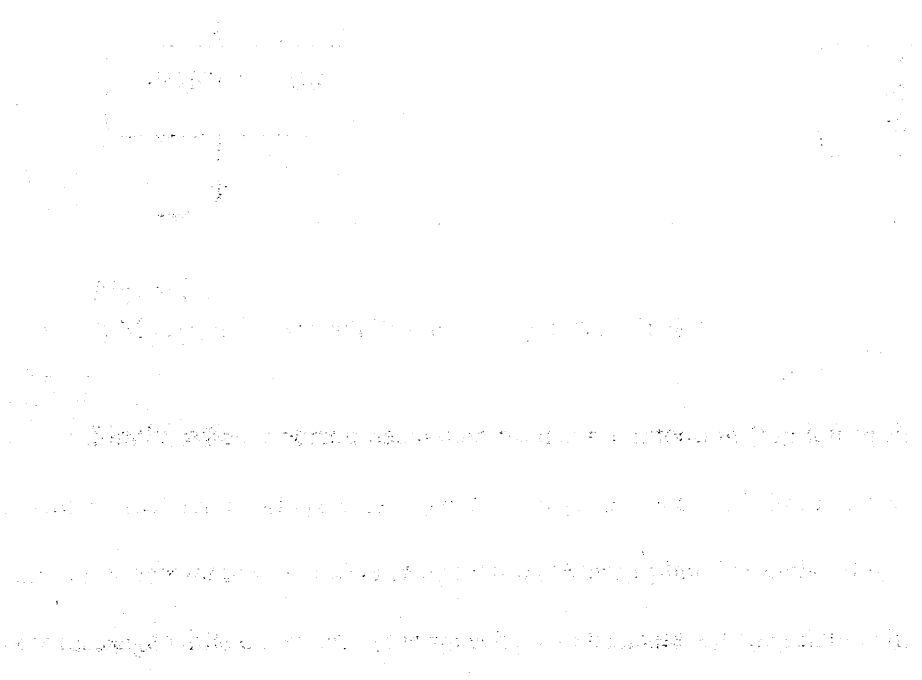
2.5.1. Theoretical Frameworks of Speech Production

2.5.1.1. Speech Production in L1

In order to understand the mechanisms of L2 speech production, we first need to understand how L1 speech operates. Since the 1970s, several models have been proposed for explaining L1 speech production (e.g. Fromkin, 1971; Clark & Clark, 1977; Kempen & Hoenkamp, 1987; Levelt, 1989), though the one most influential and widely accepted in the fields of psycholinguistics as well as SLA is the modular model by Levelt (1989) (de Bot, 1992; Kormos, 2006; Segalowitz, 2010), which compiled significant findings of empirical studies on speech errors and reaction times of normal speakers and speakers with language pathologies. The model has gone through some

revisions (i.e. Levelt, 1993; 1999), and this thesis will base its argument not on the latest version (of 1999) but on the 1993 version, as the model has not been changed fundamentally, and there is benefit in using terms consistent with and relating to discussion of previous studies which cited the older versions (e.g. de Bot, 1992; Poulisse, 1997; Bygate, 2001; Yuan & Ellis, 2003; Skehan, 2009).

Levelt's model (1993) assumed separate autonomous components that are responsible for different aspects of speech production. It comprises three main systems that are involved in the process of speech production: the Conceptualizer, Formulator and Articulator, as illustrated in Figure 2.4.



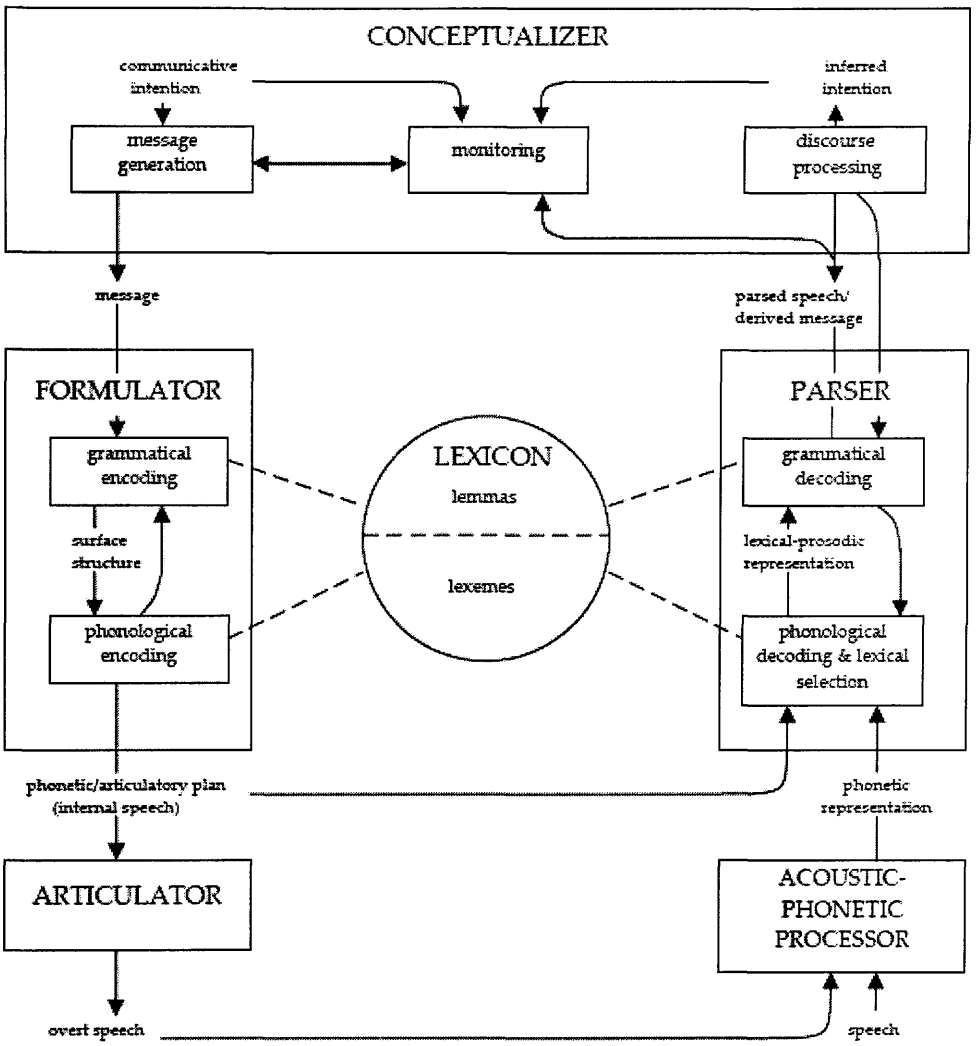


Figure 2.4
A Model of L1 Speech Production by Levelt (1989)

Firstly, when a person has a communicative intention (top left in Figure 2.4), the Conceptualizer is where a *preverbal message* is generated through two stages of planning: *macro-planning* and *micro-planning*. Macro-planning elaborates the content of the message while considering the speech act to be achieved, the relationship with the other speaker(s), level of formality and expected discourse patterns, as well as the content knowledge of the topic. In micro-planning, decisions are made on linguistic propositions such as tense, perspective (i.e. who/what will be the subject) from which

the intention will be told, what information is already known or shared, and the presentation order of information. The final product of these two stages of planning, a preverbal message, is passed on to the second system, the Formulator.

After receiving the preverbal message which is propositionally organised, the Formulator starts *grammatical encoding,* drawing on lemmas (i.e. the meaning and syntax of words) stored in the *mental lexicon* for the necessary lexical items and their syntactic organisation. The resultant *surface structure* is then processed through *phonological encoding*, which utilises lexemes (i.e. morphological and phonological information about words) in the mental lexicon. In this way, this stage produces *internal speech (a phonetic/articulatory plan)* that is ready to be articulated.

The Articulator handles the execution of internal speech. As the generation of internal speech is considered to be slightly ahead of execution, the internal speech needs to be temporarily stored while execution is taking place. This storage system is called the *articulatory buffer*, from which the Articulator retrieves chunks of internal speech and articulates them as *overt speech.*

In order to explain how the high speed of L1 speech production is achieved, Levelt (1989: 20-2) assumed "parallel processing", which enables different chunks of a speech plan to be processed in the three systems at the same time. For example, once the Conceptualizer produces the first chunk of a preverbal message and passes it on, it starts processing the next chunk immediately while the Formulator engages in encoding the first chunk of the preverbal message. In addition, Levelt also suggested that most of the speech production process is largely automatic, which facilitates the high speed of L1 speech. The Formulator, where grammatical and phonological encoding takes place, and the Articulator operate without conscious awareness and require very little attentional resources. The conceptualization process, on the other hand, demands

attention in terms of message generation and *monitoring*.

Importantly, monitoring in Levelt's model is constantly available at each stage of speech production. Firstly, the preverbal message produced in the Conceptualizer is monitored to see if it matches the initial communicative intention. Secondly, the product of the Formulator, internal speech, is parsed and matched against the entries in the mental lexicon for accuracy. Similarly, the overt speech executed by the Articulator is listened to, parsed and again monitored. Hence, the speaker pays attention to both the meaning and well-formedness of an utterance and, if an error is detected, another cycle of message generation is triggered, leading to self-repairs.

Thus far, the mechanisms of L1 speech production proposed by Levelt (1993) have been briefly reviewed, and have explained how and why L1 speech is produced at high speed. In contrast, speaking in the L2 often tends to be more hesitant, less accurate, and to contain shorter clauses. In order to account for this, the next section starts by mentioning the differences between L1 and L2 speech production, and proceeds to introduce two theories of how attention is allocated among different aspects of speech production in L2.

2.5.1.2. Speech Production in L2

The mechanisms of speech production in L2 are considered to be fundamentally the same as those in L1 (de Bot, 1992), though three features differentiate L2 speech production according to Poulisse (1997): incomplete L2 knowledge, a lack of automaticity, and the possibility of code-switching between two languages. Poulisse further stated that Levelt's model (1987) does not need alteration in order to explain the first two features of L2 speech, while explaining the third feature has been the focus of modelling L2 speech production in terms of how the mental

lexicon(s) of bilinguals is(are) structured which enables the (conscious or unconscious) mixing or separating of two languages (Poulisse, 1997: 208).

Although exploring how the mental lexicon is structured and how the stored lexical items are activated and selected in L2 speech has been a very important and attractive area of research (e.g. Paradis, 1987; de Bot, 1992; de Groot, 1992; Poulisse & Bongaert, 1994), it is beyond the scope of this thesis. In the light of attempts to investigate the parallelness of spoken narrative tasks, what this thesis seeks in this section of the literature review are theoretical justifications of the variables to examine different aspects of candidates' linguistic performance, which renders Poulisse's (1997) other two characteristics of L2 speech more relevant.

The first characteristic, incomplete L2 knowledge, may manifest itself in the erroneous use of some lexical or grammatical features in L2 spoken performance (Poulisse, 1997: 205). However, even when L2 speakers have appropriate linguistic knowledge (e.g. of how irregular verbs inflect), they can still make errors. This is closely related to the second characteristic of L2 speech, lack of automaticity. Unlike L1 speech production, the processes which take place in the Formulator and Articulator may not be automatised in the L2, so they might require conscious attention (Kormos, 1999: 312). As a result, the process of L2 speech production is likely to be serial (in contrast to parallel processing in the L1) and, therefore, slow (Poulisse, 1997: 208). Moreover, since attentional capacity[9] is limited (Schmidt, 2001), there is less attention available for monitoring when it is being largely consumed by the encoding and articulatory stages, which eventually leads to overlooking some errors that might otherwise have been corrected.

---

[9] Attentional resources are considered to be limited due to the constraints of working memory where ongoing language processing takes place (Gathercole & Baddeley, 1993).

As can be seen, the allocation of attention plays a crucial role in producing L2 spoken performance. This has attracted an increasing number of studies which have examined the influence of allocated attention under various conditions and constraints (e.g. Foster & Skehan, 1996; Bygate, 1999; Robinson, 2001). This line of research assumes that how L2 speakers allocate attention between different aspects of performance can be manipulated by altering task characteristics and administration conditions (i.e. manipulating task complexity), and thus aims to sequence tasks according to their influence on pedagogical use in order to facilitate L2 interlanguage development. The next section discusses the two leading hypotheses in task-based research, by Skehan (1998) and Robinson (2001, 2003), both of which suggest a model of the task-related factors which affect the linguistic performance of L2 speakers. Reviewing this line of research will give insights into identifying the characteristics that need to be considered when selecting and designing parallel tasks.

2.5.1.3. Task-related Factors affecting L2 Spoken Performance

Among a number of researchers who have explored the task-related factors affecting L2 performance, Brown and Yule (1983: 37-53) were the first to conduct an empirical investigation (Fulcher, 2003: 60) by observing and analysing candidates' linguistic performance. In a series of studies, they identified a number of affecting factors, among which the following appear to be relevant to spoken narrative tasks:

1) The number of objects, characters, and events; the greater the number of elements to describe, the more cognitively difficult the task is.
2) Whether or not the same setting is maintained throughout a story; if it changes, one has to assess the relevant elements in the new setting and the effect on the

story in order to describe and make sense of it, thus it is more cognitively difficult.

3) Whether or not the characters are of the same type, usually in terms of gender; the more similar the characters are, the more complex the referential expressions have to be, thus making it more demanding in terms of precise communication.

4) Whether or not the pictures contain something culturally unfamiliar to the candidates: the task is more cognitively difficult if they have to describe objects that they have never encountered before.

In sum, Brown and Yule's factors involve cognitive load (i.e. 1 and 2), linguistic demands (i.e. 3), and familiarity (i.e. 4). Candlin (1987: 18-19) listed similar factors, but added another factor as to whether or not task structure and task goal are clearly presented to candidates. Drawing on Candlin's classification, Skehan (1996, 1998) proposed a framework for categorising tasks according to task demands which he suggested are determined by the following three dimensions interacting with one another: *code complexity*, *cognitive complexity* and *communicative stress*. Put simply, the three dimensions represent "the language required, the thinking required, and the performance conditions for a task" (Skehan, 1998: 99) respectively. More specifically, code complexity concerns both the syntactic and lexical complexity and variety required to complete the task. Cognitive complexity includes two concepts: cognitive processing and cognitive familiarity. Cognitive processing involves the amount of information processing that a candidate has to do, which can be estimated by, for example, the degree of clarity of structure and concreteness of the information in a task. Cognitive familiarity refers to candidates' familiarity with the topic and task.

Communicative stress indicates the conditions under which the task needs to be performed (e.g. time pressure, required response type).

Skehan suggested that the effects of manipulating these factors should be examined according to aspects of fluency, accuracy and the complexity of candidates' performance. The assumption underlying Skehan's suggestion is that there is a trade-off relationship between fluency and form, and additional competition within the form (accuracy and complexity) of performance. Thus, some tasks can direct candidates' attention to prefer fluency over accuracy and complexity, while others may lead to prioritising accuracy or complexity. Ellis (2003) explained this assumption concisely in relation to the model of speech production by Levelt (1989; 1993) described earlier in Section 2.4:

> [T]here are likely to be trade-offs as L2 learners struggle to conceptualize, formulate, and articulate messages. Attention to one aspect of production is likely to be at the expense of others. For example, L2 learners concerned primarily with what they want to say, i.e. with conceptualizing, may not be able to give much attention to how they say it, i.e. with formulation, with the result that their speech is full of errors. [...] Conversely, L2 learners' attention to accuracy may interfere with their ability to conceptualize, leading to marked disfluency. (Ellis, 2003: 109)

While researchers agree that attentional capacity is limited (Schmidt, 2001: 12), not every researcher accepts the existence of the trade-off relationship between accuracy and complexity that Skehan (1998) suggested. Instead of the single-capacity model of attentional resources which Skehan supports, Robinson (1995, 2001) argued

for multiple resource pools of attention, based on Wickens (1984), who advocated that attention can be divided and directed to different task dimensions at the same time.[10] Accordingly, Robinson suggested that accuracy and complexity do not compete for attentional resources because these two areas of performance do not share the same attentional resource pool, but can rather improve in tandem. Robinson's classification of the task-related factors affecting L2 spoken performance includes three categories, labelled as: *task complexity* (i.e. factors that affect cognitive processing), *task difficulty* (i.e. the factors which candidates bring to the task, such as working memory and motivation), and *task conditions* (i.e. factors relevant to task goals and grouping candidates for interactive tasks).

Among these three categories, task complexity factors have attracted a number of pieces of research into how manipulation of these factors can influence candidates' spoken performance (e.g. Gilabert, 2005; Révész, 2007; Kuiken & Vedder, 2007). This is because Robinson (2003: 56) suggested that they are "the intrinsic cognitive demands of the task", and therefore can be estimated *a priori*, which makes it easier to build research hypotheses.

Task complexity factors can be further distinguished into *resource-directing* factors (e.g. number of elements, spatial location, reasoning of causal events, reasoning of characters' intentions and relationships) and *resource-dispersing* factors (e.g. planning time, prior knowledge). Robinson (2007) argued that Skehan's trade-off relationship is only true in cases where *resource-dispersing* factors are manipulated, and that increasing *resource-directing* factors will lead to greater complexity and accuracy. This is because it will lead candidates to "complexify their speech to meet the

---

[10] Wickens's theoretical models of attention included the SEEV model (i.e. salience, effort, expectancy and value) of selective attention, and the Multiple Resources Model of divided attention to task demands (Wickens, 2007).

increased conceptual and functional demands of the task" (Robinson, 2007: 20), inducing a closer focus on grammar and thus greater accuracy.

The purpose of this thesis is not to prove or disprove either of the two hypotheses by Skehan (1998) or Robinson (2001), but to gain insights into what factors need to be examined *a priori* in order to establish the parallelness of spoken narrative tasks. Therefore, the next few paragraphs in this section aim to summarise the factors suggested by the two hypotheses, according to the recommendations made by Bachman (2002: 469), which were described earlier in Section 2.2.2, in order to distinguish between task-inherent features, the attributes of candidates, and interactions between the two.

Bachman (2002) criticised Skehan (1998) for confounding candidate factors with the effects of the task itself. According to Bachman, only *code complexity* (i.e. the linguistic complexity required to complete a task) is solely characterised by the task. *Cognitive complexity* is a function of the characteristics of candidates and tasks, since the amount of processing required and the degree of task and topic familiarity will inevitably differ from candidate to candidate. The problem may also be applied to Robinson's *task complexity* factors too, although Robinson's classification appears to be more in line with Bachman's suggestions in that it attempts to distinguish between task-inherent features (i.e. *task complexity*) and candidate factors (i.e. *task difficulty*). Nevertheless, *task complexity* factors, especially *resource-dispersing* factors, include features such as prior knowledge, which might differ from candidate to candidate. However, Robinson (2001) justified his classification by stating that task complexity is:

> [...] the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner. These differences in information processing demands, resulting from

> design characteristics, are relatively fixed and invariant (Robinson, 2001: 29, my emphasis).

Based on this argument, this thesis will include the relevant features of Robinson's task complexity as well as Skehan's cognitive complexity, which largely overlap when identifying the task characteristics of spoken narrative tasks.

As for candidate factors, Skehan (1998) did not include them in his framework of task-related factors. On the other hand, Robinson (2001: 31-2) accepted Bachman's view (2002) that candidates differ in their cognitive ability, such as working memory capacity, and attributed candidate factors (i.e. *task difficulty* factors in his terminology) to between-candidate differences in task performance stemming from the different availability of attentional resources. Robinson (2001: 32) further divided candidate factors into *ability variables* (e.g. working memory capacity, cognitive styles) and *affective variables* (e.g. motivation, confidence, anxiety). This suggests the need to control for ability variables and investigate affective variables as well, since they relate closely to the process candidates undergo during task completion.

There is a category of factors in both Skehan's and Robinson's classifications which I have not discussed in the light of criticisms by Bachman, but they are not relevant to examining spoken narrative tasks. Bachman (2002) again regarded Skehan's *communicative stress* as a function of candidate and task characteristics; however, many features in this category relate to interactive tasks (e.g. 'opportunity to control interaction' which depends on a candidate's language ability, affect, etc.), and are therefore irrelevant to spoken narrative tasks which are monologic. Similarly, Robinson's *task conditions* (i.e. factors relevant to task goals and grouping candidates for interactive tasks) are meant for interactive tasks, so they too will not be discussed further.

To summarise, briefly, the discussion so far, Figure 2.5 presents the overlap between the factors suggested by Bachman (2002), Skehan (1998) and Robinson (2001) (straight lines between factors indicate overlaps).

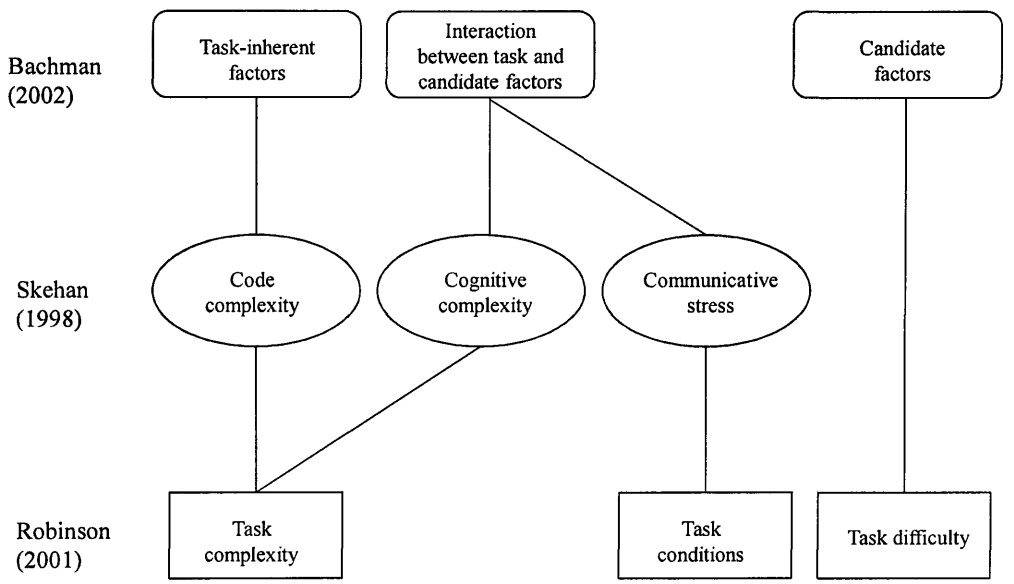


Figure 2.5
Task-related Factors Affecting L2 Spoken Performance

Firstly, Bachman insists that task-inherent factors refer only to the code-complexity factors in Skehan's framework, and that the other two represent interaction between task and candidate factors. However, Robinson argues that task complexity factors, which overlap with Skehan's code complexity and cognitive complexity, are task-inherent and that their effects on performance are relatively fixed. Skehan's communicative stress largely overlaps with Robinson's task conditions, but both are irrelevant to investigating monologic spoken narrative tasks with the strict control over contextual features (i.e. examiner, raters, etc.) applied in this thesis. Finally, Bachman's candidate factors are also specified in Robinson's task difficulty factors. Investigating

the affective variables in candidate factors (i.e. perceived difficulty, anxiety, etc.) means gaining insights into substantive validity (i.e. candidates' processes during task performance); its operationalisation is discussed further in Section 2.5.4.

Hereafter, this thesis regards task complexity factors as being inclusive of Skehan's code complexity and cognitive complexity, and uses this term accordingly. The next section discusses the task complexity factors relevant to spoken narrative tasks in detail. These will be examined for *a priori* evidence of parallelness in Pilot Studies 2 and 5. In addition, this thesis will utilise Robinson's task difficulty questionnaire to examine their affective variables (discussed further in Section 2.5.4). However, they are labelled as 'candidate factors' in this thesis to avoid confusion with the 'task difficulty' which MFRM analysis produces.

2.5.2. *A Priori* Evidence of Task Parallelness: Task complexity Factors

If two spoken narrative tasks are to be proven parallel, they should share the same features of task complexity so that their effects on candidates' performance are expected to be the same. This section discusses relevant task complexity factors, from Skehan (1998) and Robinson (2001), and oversees how they have been operationalised and investigated in previous studies. Due to the research design of this thesis, some features of task complexity are irrelevant; the use of only one task type (i.e. spoken narrative tasks) excludes the factors of information organisation, information type and task familiarity (in Skehan's *cognitive processing*), and the strictly controlled contextual features for task implementation (such as planning time and examiner, as discussed in Section 2.3) render some of Robinson's features irrelevant (i.e. +/- here-and-now, +/- single task, +/- planning time, +/- few steps, +/- perspective taking).

Table 2.1 lists the remaining relevant factors of task complexity as well as their operationalisation by previous researchers.

Table 2.1
Relevant Task complexity Factors for Investigating the Parallelness of Spoken narrative Tasks

	Task Complexity	Operationalisation
Skehan (1998)	Syntactic complexity & variety Lexical complexity & variety Topic familiarity	Brown et al. (2002); Weir & Wu (2006)
Robinson (2001)	+/- few elements +/- intentional reasoning +/- causal reasoning +/- prior knowledge	Kuiken & Vedder (2007) Baron-Cohen (1995) Robinson (2005) Robinson et al. (1995)

Previous studies utilised expert judgements (either by raters or the authors themselves) when hypothesising which task(s) should be more complex according to the relevant task complexity factors. Brown, Hudson, Norris and William (2002) asked two raters to assign a plus or a minus to a variety of 103 tasks based on whether they thought the required range of language was above or below the average ability of the candidate population. Weir and Wu (2006), in their attempt to establish equivalence of picture description tasks in a semi-direct speaking test, asked 12 raters to agree or disagree that the tasks from three test forms were the same in their assuming of candidates' familiarity with the required lexical items, grammatical structures and functions, as well as with the roles of people, locations, objects and events depicted in the pictures.

Similar to an aspect of task complexity from Brown and Yule (1983), which concerns the number of characters, objects and events to be described (explained in Section 2.5.1.3), Kuiken and Vedder (2007) assumed the degree of task complexity to

be high or low according to the number of elements to be included in the candidates' performance in their research into the complexity of writing tasks.

The degree of reasoning that a task demands is also important, and two kinds of reasoning are relevant to spoken narrative tasks: causal reasoning and intentional reasoning. Causal reasoning (Baron-Cohen, 1995) is often required in narratives to explain event relations and to support the interpretation of events by giving reasons, which may elicit logical coordinators and subordinators such as *so*, *because* and *therefore* (Robinson, 2005: 5). Likewise, in order to explain characters' actions, narrating a story may demand reasoning about characters' intentions and beliefs, requiring cognitive state verbs such as *think* and *believe,* which are likely to be followed by subordinate clauses (Robinson, 2005: 5). Therefore, both types of reasoning are considered to affect the syntactic complexity of spoken narrative performance.

Lastly, Robinson, Ting and Urwin (1995) examined the effects of prior knowledge of the topic of a listening task (lecture comprehension) by giving a pre-listening activity to an experimental group of candidates. Comprehension was measured by multiple-choice questions concerning not only the content of the lecture but also the inferences that could be made from the lecture. The experimental group generally scored better than the control group, and significant differences were found for the questions about inferences. This indicates the possibility that, because of the prior knowledge of topics, the experimental group did not need to focus as much attention on understanding the lecture, which gave them an advantage on the questions about inferences. This could also be applicable to candidates' prior knowledge of the topics of spoken narrative tasks, in that it might affect how much attention is directed towards understanding the content of the story, eventually affecting linguistic performance.

In sum, the *a priori* analysis of spoken narrative tasks in this thesis should be judged as being the same in terms of the task complexity factors of the required linguistic complexity and variety, topic familiarity, number of elements depicted in the pictures, required causal and intentional reasoning, and the assumed candidates' prior knowledge of topics. From the next section onwards, this literature review proceeds to discuss the variables for investigating *a posteriori* evidence of task parallelness along with candidates' linguistic performance and processes.

2.5.3. *A Posteriori* Evidence of Generalisability: Linguistic Performance of Spoken Narrative Tasks

As we have seen so far, the current mainstream of task-based research deals with the effects on L2 spoken performance of changing the task complexity factors, based on the frameworks by Skehan (1996; 1998) and Robinson (1995; 2001). Such studies include: manipulating the planning time, having the pictures in front or not, missing out some pictures in a sequence or not, to name but a few (Foster & Skehan, 1996; Mehnert, 1998; Norris, Brown, Hudson & Bonk, 2002; Ortega, 1999; Robinson, 1995; Skehan & Foster, 1999; Wigglesworth, 1997). In the light of the discussion on the allocation of attention during task performance, the most common variables used to capture differences in linguistic performance under the different conditions are those of fluency, accuracy and complexity. These three types of variable are reviewed in Sections 2.5.3.1 to 2.5.3.3.

In addition, since the *a priori* analysis of task complexity factors will include the number of elements (i.e. characters, events, etc.) expressed in the pictures, *a posteriori* analysis of linguistic performance should examine whether these elements are actually mentioned in the narratives told by the candidates. In order to investigate

this, the frameworks of narrative features by Labov (1972) and other relevant studies are reviewed in Section 2.5.3.4.

2.5.3.1. Fluency

For several few decades, various researchers have proposed different definitions of fluency because of its multidimensional nature (Segalowitz, 2010), ranging from the 'speed' or 'smoothness' of speech delivery to factors beyond individual speakers such as the pragmatic acceptability of utterances and listener perceptions (see Koponen & Riggenbach, 2000). Among them, Lennon (2000: 26) offered a working definition of fluency as "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of on-line processing" . Furthermore, in an earlier article, Lennon (1990) suggested that fluency variables could be classified into two aspects: *temporal* variables which deal with the speed of delivery, and *hesitation markers* that represent disfluency phenomena such as repetition and false starts.

In task-based research, various fluency variables are used in different studies. Examples of temporal fluency variables include speech rate (the number of meaningful syllables per minute) (Yuan & Ellis, 2003), mean length of runs (the number of words per pausally defined unit) (Crookes, 1989; Mehnert, 1998; Yuan & Ellis, 2003; Robinson, 1995) and the number of words per C-unit (Robinson, 2001). However, regarding use of the number of words per syntactically defined unit as a fluency variable (e.g. Robinson, 2001; 2007), Koizumi (2005) showed that it is more likely to be measuring syntactic complexity because of its higher correlation with other variables of syntactic complexity. The other type of fluency variables, hesitation markers, include: the number of false starts, reformulations, lexical replacements (Skehan &

Foster, 1999; Iwashita, McNamara & Elder, 2001); the number of pauses per T-unit (Bygate, 2001); the number of unfilled pauses per T-unit (Iwashita, et al., 2001); the number of clauses containing self-repairs per clause (Wigglesworth, 1997).

A number of researchers have attempted to identify variables of fluency that can distinguish between fluent and non-fluent speech (e.g. Lennon, 1990; Towell, Hawkins & Bazergui, 1996; Derwing, Rossiter, Munro & Thomson, 2004; Ejzenberg, 2000; Kormos & Dénes, 2004). Among them, Kormos and Dénes (2004) offered the most recent and relevant validation study of fluency variables; this involved the largest number of participants (N = 16) and the use of computer technology to measure precisely the length of pauses. More specifically, they correlated human ratings of how fluent the speech was and quantified the results of several variables of fluency. Among the temporal variables, speech rate and mean length of runs correlated the most highly with fluency ratings. None of the hesitation markers or frequency of pauses turned out to be valid. Previous studies have also agreed that speech rate and mean length of runs are the best predictors of perceived fluency (e.g. Towell, et al., 1996).

2.5.3.2. Complexity

Complexity is "the extent to which learners produce elaborated language" (Ellis & Barkhuizen, 2005: 139), and is concerned with the syntactic and lexical aspects of performance. Variables of syntactic complexity in previous studies have included: the number of subordinate clause per clause (Wigglesworth, 1997); the number of words per unit (Bygate, 2001 (T-unit); Mehnert, 1998 (C-unit); Ortega, 1999 (pausally defined unit)); the number of words per the number of clauses per C-unit (Skehan & Foster, 1999; Foster & Skehan, 1996; Iwashita et al., 2001; Robinson, 2001; 2007); and the number of subordinate clauses per T-unit (Crookes, 1989; Mehnert, 1998). These

are called *general* variables of syntactic complexity, and can be classified into two types according to what they measure: the length of a chosen unit and the amount of subordination. In addition, Norris and Ortega (2009) recommended examining the amount of coordination (Bardovi-Harlig, 1992), as opposed to subordination, especially in studies involving candidates of lower proficiency. Coordination and subordination are part of *conjunction*, one of the cohesive devices which are said to contribute to coherence[11] (Halliday & Hasan, 1976). Coherence refers to whether the events in a narrative are interconnected and make sense as a whole, and is another important aspect to examine in this thesis because recounting events in a coherent sequence is regarded as essential to narrative performance (Luoma, 2004: 144).

As can be seen, the general variables of syntactic complexity deal with different units of analysis, clauses and subordinate clauses whose definitions often vary between researchers. For the units of analysis, some researchers use T-units as the unit for analysis; however, Ellis and Barkhuizen (2005) recommended using C-units or AS-units because they can take sub-clausal units into account. AS-unit stands for Analysis of Speech Unit, defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn & Wigglesworth, 2000). Moreover, Foster et al. (2000: 362-3) argued that AS-units are more reliable than C-units, since their definition takes intonation and pausing into account and the examples of AS-units in their article offer a consistent classification of false starts, repetitions and self-corrections. Therefore, in this thesis, AS-units are employed where units are involved in the variables. Accordingly, the definitions of clauses and subordinate clauses follow those

[11] This may not be is always the case, as Halliday and Hasan (1976) illustrated. Nevertheless, Carrell (1982: 485) insists that coherent texts are likely to be cohesive.

suggested by Foster et al. (2000: 365-6):

- An independent clause will be minimally a clause including a finite verb. (e.g. *That's right*; *Turn left*; *I take a different way*);
- An independent sub-clausal unit will consist of (a) one or more phrases which can be elaborated into a full clause by means of the recovery of ellipted elements from the context of the discourse or situation (e.g. *Three months* as the answer to a question: *How long will you stay here?*);
- Or, an independent sub-clausal unit will consist of (b) a minor utterance, which will be defined as non-sentences (e.g. *Oh poor woman*; *Thank you very much*; *Yes*);
- A subordinate clause will consist minimally of a finite or non-finite verb element plus at least one other clause element (subject, object, complement or adverbial). Examples include:
    - It is my hope *to study crop protection*
    - And you'd be surprised *how he can work*
    - When I was at university *I specialised in this subject.*

In addition to the general variables of syntactic complexity that have been mentioned so far, Norris and Ortega (2009) recommend using specific variables, such as the semantic properties of verbs (i.e. state, activity, etc.), which will inform researchers of the current stage of L2 acquisition that a candidate is at. However, this information is irrelevant to the purpose of this thesis which is to examine whether two spoken narrative tasks are successful in eliciting parallel linguistic performances.

Another aspect of complexity, mentioned earlier, is lexical complexity, which can be further divided into lexical variety (and density) and lexical sophistication (Wolfe-Quintero, Inagaki, & Hae-Young, 1998). For lexical variety, the following variables are often employed: the ratio of lexical words (Robinson, 1995), type-token ratio (Wigglesworth, 1997; Crookes, 1989; Ortega, 1999; Robinson, 2001; 2007), mean segmental type-token ratio (Yuan & Ellis, 2003), and D value (Kormos and Dénes, 2004). These variables represent in-text lexical complexity. The ratio of lexical words is

supposed to indicate the 'density' of a text because lexical words primarily convey information. Nevertheless, Laufer and Nation (1995: 309) argue that "it does not necessarily measure lexis, since it depends on the syntactic and cohesive properties" of performance. Also, type-token ratio (TTR) has been criticised as being greatly affected by text length (Wolfe-Quintero et al., 1998), and mean segmental TTR is devised to compensate for this disadvantage by segmenting each text into a certain length and averaging the TTR of all the segmented texts. However, Jarvis (2002) showed that D value is more reliable than TTR and mean segmental TTR, using curve-fitting modelling procedures.

Such variables of in-text lexical variety do not reveal how 'rare' or 'advanced' the lexical items used in linguistic performance are. Although lexical sophistication does not appear to have been widely examined in task-based research, it may be rational to examine the degree of lexical sophistication in each task in order to establish the parallelness of elicited word 'levels'. In the field of vocabulary learning research, Laufer and Nation (1995) argued for examining the Lexical Frequency Profile (LFP), that is to say, how many frequent words are observed in a text using frequency-based word lists. This is based on the assumption that frequent words are likely to be learned earlier and, therefore, infrequent words may represent higher 'levels' of words. LFP utilises the General Service List (i.e. the list of the most frequent 1,000 word families and the second 1,000) plus the University Word List (i.e. 550 frequent words in academic texts across all subjects). Laufer and Nation (1995: 319) found that the percentage of words from the most frequent 1,000-word list successfully discriminated between different levels of proficiency, and therefore concluded LFP to be "reliable and valid" for measuring lexical use in writing.

Similarly, Meara and Bell (2001) devised a program called P_Lex which

counts the number of 'difficult' words in a text. Meara and Bell (2001: 9) criticised LFP for not being sensitive enough to capture the characteristics of different texts and for relying heavily on the UWL list and a "not in the list" category where words in a 'typical' text seldom belong. P_Lex uses Nation's word list in which the first 1,000 words are marked as 'easy'. All other words are classified as infrequent and 'difficult'. P_Lex divides a text into 10-word segments and calculates the ratio of 'difficult' words in each segment, eventually producing a single value called lambda to indicate how advanced lexical use in the text is. However, when it faces a "not in the list" word, P_Lex demands that the researcher classify it as 'easy' or 'difficult'. This may seriously threaten the reliability and generalisability of the resulting lambda value.

Another frequency-based word list that is worth noting is JACET8000 (JACET, 2003) which is a collection of eight lists of 1,000 words derived from corpus-based research on English newspapers, textbooks, examinations and exam-preparation books available in Japan. While LFP is chosen for its reliability and widespread use in previous studies of vocabulary learning involving non-Japanese learners, JACET8000 might be expected to be more sensitive to the lexical use of Japanese candidates.

2.5.3.3. Accuracy

Accuracy refers to the extent to which the target language is produced according to its rule system (Skehan, 1996: 23), and there are general and specific variables that can characterize the accuracy of an utterance. The general variables include the percentage of error-free clauses (Skehan & Foster, 1999; Foster & Skehan 1996; Yuan & Ellis, 2003), the percentage of error-free C-units (Robinson, 2001; 2007), the number of errors per T-unit (Bygate, 2001) and the number of errors per 100 words (Mehnert, 1998). Specific variables can vary depending on the target features that the

researchers focus on. Examples include the percentage of correct use of target features such as articles, verb morphology, and word order (Robinson, 1995; Wigglesworth, 1997; Crookes, 1989; Skehan and Foster, 1997; Mehnert, 1998). Ellis and Barkhuizen (2005) suggested that target-like verbal morphology may be suitable for syntactic accuracy for focused tasks that are intended to elicit certain grammatical features.

For general variables, researchers disagree as to which variables are thought valid. Bygate (2001) suggested that calculating the number of errors per unit might be more sensitive because it does not obscure the actual occurrences of errors as counting error-free units does. On the other hand, Mehnert (1998: 86) argued that errors per 100 words may be suitable for relatively lower-proficiency speakers since it does not involve definitions of clauses and units which are often problematic. Since there exists no previous research which has attempted to validate these variables of accuracy, unlike the variables of fluency and lexical complexity, there is no way of knowing beforehand which variables should be used in this thesis. Therefore, it is crucial that this issue is investigated along with comparing the accuracy of performance of two spoken narrative tasks.

Another issue that needs to be mentioned regarding accuracy variables is the problem of defining 'errors' and 'target-like' use. There has been a great deal of effort put into developing a reliable system for error identification and classification in the field of error analysis (e.g. Corder, 1981; James, 1998), and some researchers have applied a detailed error classification framework in second language writing such as Kroll (1990) and Polio (1997). However, such thorough error identification, based on exhaustive lists of errors, is not practical when researching spoken language, as there will be ellipsis, incomplete sentences, and slips of the tongue that even native speakers can make. Thus, it appears that using the very broad baseline adopted by Skehan and

Foster (1997: 195), who regarded language use which is 'nonexistent in English or indisputably inappropriate' as errors, will be useful and practical.

In summary, Sections 2.7.3.1 to 2.7.3.3 have reviewed the variables of fluency, complexity and accuracy that have been extensively used in the field of task-based research for examining the effects of manipulating task complexity factors on L2 performance. Three issues need to be emphasised at this point. Firstly, such previous studies did not establish the evidence of task parallelness beforehand (Weir & Wu, 2006), although some of them used different spoken narrative tasks under different conditions (e.g. Crookes, 1989; Robinson, 1995). This may be due to an unverified assumption that they are parallel so that any differences in performance may be attributed to the differences in conditions, which further emphasises the importance of conducting research on task parallelness in this thesis. A second cause of concern in these studies is that most of the operationalised variables, especially for accuracy and syntactic complexity, are used conventionally without any empirical justification. This seriously threatens the credibility of the research findings and indicates that the variables need to be examined for their sensitivity to capture differences in the quality of performance. Finally, in his recent article which re-examined a series of task-based studies, Skehan (2009) pointed out the lack of native speaker (NS) data in the field and insists on the necessity of obtaining such data in order to discuss meaningfully the performance of L2 speakers. To the best of my knowledge, Foster and Tavakoli (2009) are the only researchers that have collected NS performance and compared it with learner data on the same spoken narrative tasks. The NS performance in the study by Foster and Tavakoli (2009) was used as baseline data to substantiate their claims in relation to the effects of the task complexity factors under investigation, without having to consider the influences by L2 processing. Consequently, it appears very important

that NS data is included in the investigation.

The next section adds another aspect to be examined in spoken narrative performance, although it is not often used in the previous studies that have been discussed so far; this is 'idea units' which also represent the features of narratives. This will complete the review of the aspects of linguistic performance that need to be investigated in order to establish the validity evidence for generalisability.

2.5.3.4. Idea Units

As briefly mentioned in Section 2.2.1, performance that is based on spoken narrative tasks should follow the organisation of a narrative. Luoma (2004: 144) suggested that the requirements for narrative structures are as follows: setting the scene; identifying the characters and referring to them consistently; identifying the main events; and telling them in a coherent sequence. This suggestion corresponds with Labov (1972: 360), who defined a minimal requirement for a narrative as "a sequence of two clauses which are temporally ordered". Specifically, a fully-formed narrative will have the following features: *abstract* (summary of the whole story at the beginning), *orientation* (setting the time, place, characters and situation), *complicating action* (telling all the events in the story), *evaluation* ("indicating the point of the narrative"), *result or resolution* (telling what happened in the end), and *coda* (ending or concluding the narrative) (Labov, 1972: 363-70).

In SLA studies, some researchers have used Labov's broad categorisation to classify the events in narratives (e.g. Viberg, 2001), and others have categorised events according to more detailed units of analysis such as clauses (e.g. Aarssen, 2001). Appel (1984) is one of the researchers who employed a more detailed segmentation of events when analysing spoken narrative performance. To be more specific, Appel investigated

how many events or "idea units" in a story were covered in the candidate's narration, recalled from a written prompt, which she used to compare first and second performances by the same candidate. Ellis and Barkhuizen (2005: 153-154) argued that this variable[12] is best suited for use when the performance is based on pre-determined content, which is suitable for spoken narrative tasks. The definition of an idea unit is "a message segment consisting of a topic and comment that is separated from contiguous units syntactically and/or intonationally" (Ellis & Barkhuizen, 2005: 154). It is also possible to separate *main* idea units which are the essential ideas to complete the story, from *minor* idea units that are not essential but which enrich the story (Ellis & Barkhuizen, 2005: 154).

The advantage of examining such features of a narrative is that it will enable us to judge whether the characters, objects and events shown in the pictures are actually included in the narration. If candidates include fewer objects and events than expected, it is another piece of *a posteriori* evidence for the task complexity of the spoken narrative tasks examined.

2.5.4. *A Posteriori* Evidence of Substantive Validity: Candidate Perceptions

Having reviewed the variables to examine candidates' linguistic performance (i.e. evidence for the generalisability aspect of validity) so far, this literature review now turns to another aspect of validity (i.e. substantive aspect) to investigate candidates' processes of task performance. As explained earlier, in Section 2.4.1, Messick (1995) recommended including candidate perception to explore the process that they go through during performance as a substantive aspect of validity, and suggested the use of verbal protocols for it. However, it is not possible to conduct verbal protocols with

[12] Ellis and Barkhuizen (2005: 153) categorised this variable as being of "propositional complexity".

spoken performance. Instead, alternative measures to facilitate candidates' self-reflection were adopted in previous studies, such as direct observations, questionnaires and post-task interviews (e.g. O'Loughlin, 2001; Fulcher, 1996). Of particular relevance to this thesis is the study by O'Loughlin (2001) on the equivalence of direct and semi-direct tests of English (***access***), presented in Section 2.5.2, in which candidates' processes were explored through asking questions concerning their perceptions of the difficulty of tasks and tests, the sufficiency of preparation time and response time, the degree of stress, as well as their attitudes towards the examiner (direct test version) and recording instruments (semi-direct test version).

Other researchers have attempted to investigate whether or not such candidate perceptions are related to their actual performances, and found a clear relationship between the two in that the candidates who were more proficient displayed more positive perceptions (e.g. Scott & Madsen, 1983; Iwashita & Elder, 1997; Brown, 1993). However, these studies investigated perceptions towards a test, and therefore their results and interpretation cannot be free from possible influence by previous experience with the test (Elder, Iwashita, & McNamara, 2002: 351).

In view of this problem with previous studies, Elder et al. (2002) utilised a questionnaire on candidate perception prepared by Robinson (2001), which was tailored for use at the single-task level, to explore the difficulty of spoken narrative tasks from the Test of Spoken English with different task complexity factors being manipulated. Robinson's questionnaire was devised to investigate the affect variables that candidates brought to the tasks in order to examine candidate factors in his framework of task-related factors affecting L2 performance. Elder et al. (2002) did not find candidate perceptions to be in accordance with either the predicted difficulty with regard to task complexity factors or the scores given to candidates' spoken narrative

performances. Elder et al. (2002: 364) concluded that candidate perception is "a multidimensional phenomenon, resulting from a series of complex and unstable interactions between different task features and different test-taker attributes", and called for more research. This thesis, with its control for task complexity factors and candidates' proficiency levels, may be able to provide a baseline research for this discussion.

Robinson's questionnaire (2001), used in the study by Elder et al. (2002), includes five questions on candidates' perceived difficulty of the task, degree of stress, self-rating of their performance, interests, and motivation towards the task. Robinson's questionnaire is intentionally short and brief, with a 9-point scale, so that it can be given immediately after task completion with minimal disruption on the next task's performance (Robinson, 2001: 41). With its strong advantages, Robinson's questionnaire is expected to provide additional information about other candidate factors such as proficiency levels and background information, which will help further to triangulate the evidence for investigating task parallelness.

2.6. *A Posteriori* Evidence of Score Parallelness

Finally, this literature review comes to the last section on the different types of evidence for task parallelness: scores. As discussed in Section 2.3, previous studies in language testing have looked into score equivalence, traditionally parallel-form reliability, and in more recent studies by MFRM. In order to score a candidate's performance, firstly, there needs to be a rating scale. The definition of a rating scale is presented by Davies et al. (1999), below, as:

> A scale for the description of language proficiency consisting of a series of constructed levels against which a language learner's performance is judged.

> Like a test, a proficiency [rating] scale provides an operational definition of a linguistic construct such as proficiency. Typically such scales range from zero mastery through to an end-point representing the well-educated native speaker. The levels or bands are commonly characterised in terms of what subjects can do with the language (tasks and functions which can be performed) and their mastery of linguistic features (such as vocabulary, syntax, fluency, and cohesion). [...] Raters or judges are normally trained in the use of proficiency scales so as to ensure the measure's reliability. (Davies et al., 1999: 153-4)

This definition highlights several important concepts for constructing or selecting rating scales. Firstly, rating scales must reflect the construct under investigation. Secondly, the levels in rating scales should describe what candidates can do and what kinds of linguistic features they may exhibit. Lastly, raters must be trained in the reliable application of rating scales. In the case of this thesis, the construct, broadly defined, is the ability to produce spoken narrative stories (based on picture sequences). In order to examine task parallelness, the aspects of linguistic performance which need to be investigated were identified earlier in this literature review: fluency, syntactic complexity (including coherence and cohesion), lexical complexity, accuracy, and idea units. Thus, the rating scales should also include these aspects. As constructing such a rating scale from scratch would be unrealistic within the scope of this thesis, it is necessary to use an existing rating scale which, ideally, can also provide materials for rater training, since having benchmark performances is absolutely indispensable for rating spoken performance.

After a thorough search and trialling spoken narrative tasks for the main study (as described in Pilot Studies 2 and 5), the tasks used in the main study did not come from a test provided with rating scales tailored for this task type. Accordingly, it was eventually decided to adopt the Common European Framework of References for Languages (CEFR) in the main study; this has three advantages, as explained next.

Firstly, the CEFR is the comprehensive outcome of active research conducted

in the past 30 years (Council of Europe, 2001); according to empirical research presented by North (2000), the CEFR is well-developed and of high quality. Secondly, the standardised samples of spoken performance illustrating the six CEFR levels are publicly available[13] for training purposes. What is more, the Council of Europe has published a manual which offers guidance on how to train raters to rate performances according to the CEFR (i.e. Council of Europe (2009)). These materials, as well as the rating scales, are often impossible to obtain from existing tests from outside testing organisations. Moreover, the CEFR oral assessment criteria grid (i.e. Council of Europe, 2009: 185) lists Range, Accuracy, Fluency and Coherence for spoken performance which may correspond to the complexity, accuracy, fluency and coherence under investigation. Since this thesis does not handle interactive tasks, the column for Interaction will be replaced by Sustained Monologue (Council of Europe, 2001: 58-9). Revising the grid is supported by the Council of Europe (2009: 53) in order better to suit the rating of performances in the samples. The third justification to use the CEFR for rating scales is the comparability of results with other studies. Since the CEFR has been widely recognised and extensively referred to in language testing and teaching research, the results and implications of the main study will appeal to a broader audience.

The rater training, followed the procedures suggested by the Council of Europe (2009) for 'standardisation', involves two phases: training in rating performances in relation to the CEFR levels (using illustrative samples) and benchmarking narrative performance samples (i.e. my main data samples) to CEFR levels. This process is reported in detail in Chapter 4.

---

[13] A DVD can be sent freely on request, and samples can also be seen on the CIEP website: http://www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/index.php.

2.7. Summary

Following the review of previous studies in the field of language testing in Sections 2.2 and 2.3, the relevant literature in SLA, especially in task-based research, was discussed in Section 2.5 so as to identify how the evidence for generalisability and substantive validity could be operationalised. As a theoretical framework, Levelt's model of L1 speech production (1993) was introduced, followed by discussion relating to the allocation of attention, which eventually led to two hypotheses by Skehan (1998) and Robinson (2001) in task-based research concerning the effects of task complexity factors on L2 speech production. At this point, the literature review identified the task complexity factors that should be investigated *a priori* for task selection in this thesis.

As a part of the *a posteriori* analysis for task parallelness, the generalisability aspect of validity can be examined in terms of the fluency, complexity, accuracy and idea units that emerge in candidates' linguistic performance. It was also decided to obtain ratings data using the CEFR oral assessment grid which is expected to correspond to these aspects. The ratings data will be analysed using MFRM analysis together with rating scales and rater leniency, producing values for the difficulty of the narrative tasks in question. To explore substantive validity, the process which candidates go through during task completion will be investigated by a short questionnaire devised by Robinson (2001) (see Appendix 5).

2.8. Conclusions to the Literature Review

Following the discussion of the aspects of validity, their operationalisation, relevant task complexity factors for task selection, and the selection of rating scales, this literature review finally introduces the research questions, below:

- RQ1. Is the difficulty of the two spoken narrative tasks (to be selected in pilot studies) the same according to MFRM analysis?

- RQ2-1. Are the candidates' perceptions of the two spoken narrative tasks the same?

- RQ2-2. Are the candidates' perceptions of the two spoken narrative tasks the same at different levels of proficiency?

- RQ3-1. Are the performances of the two spoken narrative tasks the same in terms of the linguistic variables?

- RQ3-2. Are the performances of the two spoken narrative tasks the same in terms of the linguistic variables at different levels of proficiency?

- RQ4. How do the linguistic variables correlate with the ratings of spoken narrative performance in the corresponding rating categories?

Each research question presents research gaps that this thesis aims to fill. Firstly, RQ1 with the use of MFRM provides values for task difficulty of the two spoken narrative tasks, taking into consideration candidate ability as well as the contextual factors involved in speaking assessment, as discussed in Section 2.2. Moreover, using CEFR for the rating scales and the procedures for rater training suggested by the Council of Europe (2009) adds to the originality of this thesis as well as to the comparability of the results with other studies. Secondly, RQ2 refers to

candidate perceptions of the two tasks using Robinson's questionnaire (2001), and is expected to provide a baseline study for the discussion of the relationship between candidate perceptions and proficiency. Thirdly, RQ3 investigates task parallelness in terms of the candidates' linguistic performance, with RQ3-1 looking at the whole candidate population, and RQ3-2 considering the effects of different proficiency levels. RQs 1 to 3 aim to provide comprehensive evidence based on scores, candidates' linguistic performance, and candidate perceptions for which, as Section 2.3 demonstrated, previous studies on test and task equivalence have fallen short. Finally, RQ4 addresses the issue of the validity of the linguistic variables commonly used to examine linguistic performance, mainly in the field of task-based research, especially the accuracy and complexity variables as discussed in Section 2.5.3.

Before any of the RQs can be methodologically defined, there needs to be a series of pilot studies in order to select appropriate narrative tasks that are seemingly parallel in terms of the task complexity factors identified in Section 2.5.2, as well as to test the feasibility of the RQs. In the light of this, Chapter 3 describes the series of pilot studies conducted to feed the methodology of the main study.

Chapter 3: Pilot Studies

3.1. Introduction

Following on from the research questions (RQ) established in the previous chapter, Chapter 3 introduces a series of pilot studies to help refine the research methodology. First of all, there need to be pilot studies in order to select appropriate narrative tasks that are seemingly parallel in terms of the task complexity factors identified in Section 2.5.2: linguistic complexity, topic familiarity and prior knowledge, the number of elements drawn in the picture, as well as the demands for intentional and causal reasoning. After such careful selection, investigating task difficulty with contextual factors, as suggested by Bachman (2002), can be realised in RQ1. Additionally, the feasibility of identifying the characteristics of linguistic performance should be examined. Both of these will be explored in Pilot Studies 1 and 2, using a publicly available spoken corpus of Japanese learners of English called the NICT JLE Corpus (The National Institute of Information and Communications Technology Japanese Learner English Corpus). The transcripts data in this corpus comprise over 1,200 interviews from a test of speaking English in Japan, the Standard Speaking Test (SST), which involves spoken narrative tasks. Although the SST uses its own rating scales to rate candidates' performance and not the CEFR assessment grid, the corpus still provides invaluable information about candidates' English speaking proficiency level, some background information, which spoken narrative task was selected and given.[14]

The NICE JLE Corpus does not contain recordings data in its public version;

[14] As the author was trained as an SST interviewer and rater in 2006, actual picture sequences for spoken narrative tasks are to hand.

however, a small sample of recordings can be obtained with permission from the SST administering organisation, ALC Press. Pilot Study 3 will be another feasibility study for fluency variables (in addition to other variables) for RQ3 using these recordings. The linguistic performance of native speakers of English (NS) is also collected and incorporated into the data. Pilot Study 3 further aims to examine the 'sensitivity' of the linguistic variables under investigation, using correlations with SST levels and NS level and tests for mean differences among different levels as a stepping stone for RQ4 in the main study. Adding to the NS data used in Pilot Study 3, Pilot Study 4 examines the NS linguistic performance on the two SST spoken narrative tasks. With the findings and implications from the four pilot studies, Pilot Study 5 involves selecting the spoken narrative tasks for the main study in this thesis. Table 3.1 summarises the aims of the five pilot studies.

Table 3.1
Summary of Pilot Studies

Pilot Study	Aim
1	Feasibility study to analyse linguistic performance with two 'supposedly-parallel' SST spoken narrative tasks;
2	Analysis of task complexity factors of the two SST spoken narrative tasks with expert judgements;
3	Analysis of the 'sensitivity' of linguistic variables (including fluency) with performances of an SST spoken narrative task by Japanese candidates and NS;
4	Analysis of NS linguistic performance on two SST spoken narrative tasks;
5	Selection of two narrative tasks for the main study.

3.2. Pilot Study 1: A Feasibility Study of Linguistic Performance (1): at Two Levels of a Standard Speaking Test (SST)

3.2.1. Purpose

Through this pilot study, I aim to familiarise myself with analysing linguistic performance elicited via spoken narrative tasks, using SST tasks, and to try out several linguistic variables in terms of accuracy, complexity and idea units. This will also be beneficial for estimating and planning an appropriate timeline for the main study. Two different SST Levels are examined in this study since it is considered important to investigate task parallelness in relation to candidates' levels of proficiency.

3.2.2. Data

3.2.2.1. Tasks

Two spoken narrative tasks were derived from the SST, a speaking test administered in Japan. The SST takes the form of a 15-minute simulated conversation with an interviewer and a candidate, and includes a description (of a picture), a role-play and a spoken narrative task. The pictures or topic cards for these tasks are decided on by the interviewer from among several possible choices according to his or her estimation of the candidate's ability. The interview is recorded and rated by two raters who listen for certain rating criteria, give a level for the performance of each task, and finally decide on an overall single level of between 1 (Novice Low) to 9 (Advanced).[15]

---

[15] At the time of writing this thesis, there is no comparability study of SST levels and CEFR levels. However, judging from the SST level descriptors (ACTFL-ALC Press, 2000), SST Levels 1 to 9

The spoken narrative tasks in this pilot study are supposed to be parallel; both sequences of six pictures contain a conflict (ACTFL-ALC Press, 2000: 26) and are to be given to candidates at an estimated level of intermediate or advanced in order to elicit a narrative in the past tense. Yet, it was confirmed by Mr. Hirano, the Head of the Educational System Department at ALC Press, the developer and administrator of the SST, that no research has been done on the difficulty or parallelness of these tasks (2008, personal communication). The two tasks will be called the *train station* task and the *car accident* task (see Appendix 1 for the actual pictures).

3.2.2.2. Transcripts

The transcripts of candidates' linguistic performances from the two SST tasks were derived from the NICT JLE Corpus. The transcripts[16] of 5 candidates who were rated as between SST Levels 4 and 7 and declared their TOEIC scores were chosen for each task, giving a total number of 19 transcripts. This is to allow use of TOEIC scores as another measure of candidates' English proficiency, and to avoid any possible overlap of neighbouring levels as well as ensuring that intermediate levels, for which the SST is designed (ACTFL-ALC Press, 2000: 32), are investigated. The sample transcripts are presented in Appendix 2. The 19 transcripts from the two tasks did not come from the same candidates and, therefore, were not appropriate for examining task parallelness; nevertheless, the corpus provides a valuable resource for trialling the analysis of linguistic performance.

---

appear to correspond approximately to Below A1 to B2/C1 levels.

[16] Except for Level 4 with the car accident task, where there were only 4 transcripts in the first place.

3.2.3. Linguistic Variables

The linguistic variables used are summarised in Table 3.2, below:

Table 3.2
Linguistic Variables in Pilot Study 1

Aspects	Variables	Details
Accuracy	Target-like use of past tense	Percentage of correct use of the past tense in obligatory contexts
	Errors per 100 words	The number of errors divided by the total No. of words divided by 100
Syntactic complexity	Average length of AS-unit	Average number of words in one AS-unit
Lexical complexity	D value	Calculated using the CLAN program on the CHILDES website*
Idea units	Abstract	Summary of the whole story at the beginning
	Orientation	Setting the time, place, characters and situation
	Complicating action	Describing all the characters and events in the story in a presented order
	Evaluation	Indicating the point of the narrative
	Result	Telling what happened in the end
	Coda	Ending the narrative

Note. *CHILDES website: http://childes.psy.cmu.edu/.

Since SST spoken narrative tasks aim to elicit narratives in the past tense, it is appropriate to trial a *specific* variable of accuracy, i.e. target-like use of the past tense, as discussed in Section 2.7.3.3. As a *general* variable of accuracy, the number of errors per 100 words is trialled. For syntactic complexity, the average length of AS-units is counted, and for lexical complexity, a D value is calculated (i.e. lexical variety). The idea units are classified according to Labov's framework of a fully-developed narrative (1972) which contains Abstract, Orientation, Complicating action, Evaluation, Result or Resolution, and Coda. An example of the classification of idea units is shown below, in Table 3.3, for a transcript of a candidate at SST Level 4 on the *train station* task. Slashes indicate AS-unit boundaries.

Table 3.3

An Example of the Classification of Idea Units by Labov (1972)

Abstract	on my way to the office I had the happening last week /
Orientation	at station it was eight o'clock / and it was a usual day /
Complicating Action	but I was in the form / and I waited for the train / when I waited for the train I falled the bag / and there is the man who was next to me / and his arm hit my arm / and I fall the bag / and the bag was the under the form / so I had a trouble / and the train came in the station / and I was surprised / and I thought my bag was broken by the train / but the train left the station / I found the bag was safe / and I called the station clerk /
Evaluation	-
Result	and I got it /
Coda	-

3.2.4. Research Question

Are the two SST spoken narrative tasks parallel in terms of the linguistic variables of accuracy, syntactic and lexical complexity, and the idea units for candidates at SST Levels 4 and 7?

3.2.5. Procedure

All repetitions, fillers and self-corrections, according to the tags in the NICT JLE Corpus, were removed in order to segment utterances into AS-units. False starts were left if they formed a clause before being abandoned or rephrased, since they must be included in the average length of AS-units (Foster et al., 2000: 368). The linguistic variables were identified manually by the author, except for the D value for lexical complexity, which was calculated using the VOCD command in the CLAN program available from the CHILDES Website. For accuracy and complexity, means were compared via a Mann-Whitney U-test. For idea units by Labov (1972), each transcript

was given 1 if it had the unit and 0 if not. Chi-square tests were run on the results to see if different tasks on the same level would elicit the same idea units. All the statistical tests were conducted using SPSS 11.5.

3.2.6. Results and Discussion

Table 3.4 summarises the descriptive statistics for the variables as well as the number of words and TOEIC scores of the chosen transcripts.

Table 3.4
Descriptive Statistics for the Transcripts from the SST Tasks

Variable		Lv. 4				Lv. 7			
		Mean		SD		Mean		SD	
		Train	Car	Train	Car	Train	Car	Train	Car
TOEIC score		586.40	666.25	72.63	153.15	849.00	848.00	57.38	116.97
No. of words		103.40	138.50	23.86	89.70	143.40	181.80	37.63	57.62
D value		47.99	44.66	28.10	19.65	34.35	39.67	7.76	17.45
AS-unit length		7.73	8.59	.67	1.01	10.17	9.90	1.50	1.20
Errors per 100 words		7.93	9.58	2.61	3.88	4.59	3.84	2.80	1.01
Target-like use of past tense		.76	.64	.17	.32	.94	.95	.09	.05
Idea units	Abstract	.00	.00	.00	.00	.00	.00	.00	.00
	Orientation	1.00	1.00	.00	.00	1.00	1.00	.00	.00
	Complicating action	.80	.50	.45	.58	.40	.80	.55	.45
	Evaluation	.00	.00	.00	.00	.00	.00	.00	.00
	Result	1.00	1.00	.00	.00	1.00	1.00	.00	.00
	Coda	.60	.25	.55	.50	.20	.20	.45	.45

Note. Train = train station task; Car = car accident task.

As shown in Tables 3.5 and 3.6, below, at either proficiency level, none of the differences in the linguistic variables of accuracy, syntactic complexity and lexical complexity for the two tasks were statistically significant. Yet, it is still useful to discuss the results as the lack of statistical significance might be due to the small sample size, and the purpose of this pilot study is to test the feasibility of examining linguistic

variables as well as to identify potential issues and suggestions for further research.

Table 3.5
Results for Linguistic Variables of Accuracy and Complexity (SST Lv. 4)

Variable	Task	N	Mean	SD	Z	*p*
Target-like use of past tense	Train	5	.76	.17	-.24	.905
	Car	4	.64	.32		
Errors per 100 words	Train	5	7.93	2.61	-.24	.905
	Car	4	9.58	3.88		
Average length of AS-unit	Train	5	7.73	.67	-1.47	.190
	Car	4	8.59	1.01		
D value	Train	5	47.99	28.10	-.73	.556
	Car	4	32.62	7.76		

Note. Train = train station task; Car = car accident task.

Table 3.6
Results for Linguistic Variables of Accuracy and Complexity (SST Lv. 7)

Variable	Task	N	Mean	SD	Z	*p*
Target-like verb morphology	Train	5	.94	.09	-.11	1.000
	Car	5	.95	.05		
Errors per 100 words	Train	5	4.59	2.80	-.31	.841
	Car	5	3.84	1.01		
Average length of AS-unit	Train	5	10.17	1.50	-.31	.841
	Car	5	9.90	1.20		
D value	Train	5	43.98	17.08	-.94	.421
	Car	5	39.67	17.45		

Note. Train = train station task; Car = car accident task.

At Level 4, it appears that the train station task elicited more accurate use of the past tense and less errors per 100 words. In addition, while the train station task elicited less complex language, as indicated by the average length of AS-units, a wider variety of vocabulary was found as shown by the D value.

At Level 7, interestingly, the patterns changed. For accuracy, the correct use of verbs in the past tense was closer in terms of the difference between the two tasks. For errors per 100 words, the car accident task seemed to produce more accurate performance, which was the opposite at Level 4. Also, the car accident task produced

less syntactically complex utterances with shorter AS-units. D value remained the same and more varied vocabulary was used in the train station task. However, the gap between the tasks was smaller than at Level 4. Although not statistically significant, these changes among linguistic variables at different levels of proficiency are potentially interesting for further investigation with a larger sample. In addition, the need to examine the construct and sensitivity of variables is crucial. It has been demonstrated that, for example, there was a discrepancy between the general and specific variables of accuracy (i.e. errors per 100 words and target-like use of the past tense) at Level 7. How does this discrepancy occur, given that both variables must be indicative of the same construct? Is either of these two variables of accuracy not in line with proficiency levels? This issue will be investigated in Pilot Study 3.

Regarding idea units, the chi-square tests did not produce any statistically significant results either, which may indicate that the two SST narrative tasks were parallel at both levels. The values are listed below for each idea unit in Table 3.7.

Table 3.7
Results for Chi-Square Tests on Idea Units by Labov (1972)

Idea Units		Lv. 4					Lv. 7				
		Train	Car				Train	Car			
		N	*N*	χ^2	df	*p*	*N*	*N*	χ^2	df	*p*
Abstract	Yes	0	0	NA			0	0	NA		
	No	5	4				5	5			
Orientation	Yes	4	2	.900	1	.343	2	4	1.667	1	.197
	No	1	2				3	1			
Complicating action	Yes	5	4	NA			5	5	NA		
	No	0	0				0	0			
Evaluation	Yes	0	0	NA			0	0	NA		
	No	5	4				5	5			
Results	Yes	5	4	NA			5	5	NA		
	No	0	0				0	0			
Coda	Yes	3	3	1.102	1	.294	1	1	.000	1	1.000
	No	2	1				4	4			

Notes. Yes = mentioned. No = not mentioned. NA = not calculated.

As can be seen, the tendency was very uniform across tasks and even across levels that chi-square values for some idea units were not calculated. More specifically, the candidates seem to have told a coherent story by setting the scene and describing the characters (*orientation*), telling most of the events in the story (*complicating action* and *results*) and following the order of pictures presented in front of them. However, out of six idea units, two were not mentioned at all by any of the candidates for either task: *abstract* (e.g. "This story is about... ") and *evaluation* (e.g. "So, this story is a good example of the danger of driving while talking on the mobile."). This may be due to the situation in which the tasks were given to the candidates. It was a testing situation, for an SST interview, and the candidates were told how to begin their story (i.e. "One day last week, ... ") by the SST interviewer. This sentence is not likely to elicit abstract ideas at the beginning of a story. Similarly, the SST interviewer also takes the lead in starting and ending the task (e.g. "Okay, now let's look at another picture for today." and "That's the end of this task. Thank you."), and in moving on to the next stage of the interview, so the candidate may not have had a chance to add any evaluation at the end. If given in a different situation, different patterns might be observed. Moreover, the classification by Labov (1972) could have been too broad to differentiate more detailed narratives from less detailed ones, as most of the candidates' narratives were categorised as *complicating action*. Using smaller units of ideas which consists of a topic and a comment, as suggested by Ellis and Barkhuizen (2005: 154), might be less crude.

3.2.7. Conclusions and Suggestions for Further Research

This feasibility study has successfully familiarised the author with procedures for handling variables and statistical analyses, and also raised several important issues

and suggestions from the patterns observed in the linguistic performances on the two SST tasks. No statistically significant difference appears to indicate that the two tasks are parallel; however, with such a small sample size ($n = 4$ or 5 per level per task), this is not necessarily credible. Therefore, a larger sample size is the first suggestion for further research. Moreover, as Skehan (2009) suggests, collecting the performance of native speakers of English is essential as baseline data. It is also crucial to elicit linguistic performance from the same candidates on the two tasks if parallelness is to be examined; this will be included in the design of the main study. Secondly, examining the sensitivity of linguistic variables is crucial, so that appropriate interpretation of results can be made. In addition, it is desirable to have recordings with transcripts so that fluency can be investigated; this can give a better idea of one's spoken performance. What is more, having recordings will lead to more accurate segmenting of performance into AS-units. In this pilot study, as noted above, tags in the NICT JLE Corpus were used to identify self-repairs, repetitions and false starts. However, they may not be reliable as there were much manual tagging and transcribing involved (Izumi et al., 2003: 34), and no evidence for reliability when tagging is publicised. Self-repairs and false starts are often hard to distinguish from each other,[17] so it was best to classify them myself in a consistent way. Regarding the idea units, it was suggested that smaller units of ideas may be more useful to capture the differences between more detailed narrations and less developed ones. The sensitivity of the variables, including those of fluency, will be investigated in Pilot Study 3. Finally, this pilot study highlights the need for caution in situations where tasks are given, and the way instruction is given before

---

[17] For instance, I came across an utterance classified as a self-repair in the NICT JLE Corpus: *However he the car didn't hit Steve.* Is this *he* a self-repair, or can it be classified as a false start? Compiling such examples and deciding which category they fall into, I will be able to classify these phenomena and segment performance into AS-units in a consistent and reliable manner.

administering tasks. It will be ideal, in the main study of this thesis, to let candidates start and end the story without interrupting them or leading the process, as SST interviewers are supposed to do because of it being a testing situation.

3.3. Pilot Study 2: Expert Judgements on the Two SST Tasks

3.3.1. Purpose

This pilot study aims to collect expert judgement on the two SST spoken narrative tasks, the *train station* task and the *car accident* task used in Pilot Study 1. While Pilot Study 1 trialled examining the linguistic performance elicited by the two tasks, it is also indispensable to have objective judgement on the relevant task complexity factors of these tasks, as discussed in Sections 2.3.4 and 2.5.2. In order to do this, the experts must be given both tasks and then interviewed. Therefore, a 90-minute workshop was held with the Second Language Learning and Pedagogy Research Group at Lancaster University on 23 October 2008 involving 15 participants (including the author). During the workshop, participants were given both tasks and asked to share their opinions about them.

3.3.2. Participants

The 15 participants included 4 lecturers (2 Hungarian, 2 British), 10 MA students (4 Chinese, 1 Bangladeshi, 2 Japanese, 3 British), and the author, all from the Department of Linguistics and English Language at Lancaster University. The majority of the participants had had experience of teaching English as a foreign language.

3.3.3. Procedures

After a brief introduction about the aims of the workshop (i.e. trialling two SST narrative tasks, observing elicited performance and sharing opinions about the tasks), the participants were asked to get into groups of three and allocate the roles of narrator, listener, and observer within the group. While the narrator prepared for one minute by

looking at the first task (the *train station* task), the observer (and listener) also looked at the task and jotted down what they thought would be elicited by (i.e. primarily functions). When the narrators started, listeners listened and observers took notes as to what was actually elicited. Afterwards, the groups were asked to discuss what was expected beforehand and what was actually elicited within their group. For the second task (the *car accident* task), the same procedures were repeated. When both tasks had been discussed within each group, sharing between all groups followed.

3.3.4. Research Questions

The discussion was conducted based on the questions below:

Q1: What kinds of expressions and functions were elicited?

Q2: Were there any problems with the pictures in the tasks?

Q3: Did you think they were equally difficult?

The three questions were deliberately vague, so that the participants could freely raise issues or comments. Q1 was intended to collect evidence for the lexical and grammatical items required by the tasks, which relates to the code complexity of the tasks as advocated by Skehan (1998). It also asked about the functions expressed in the pictures, following the example of Weir and Wu (2006) who asked for expert judgement on familiarity with the required functions. Q2 was devised to investigate if there was anything unfamiliar or difficult to understand in the pictures, and was expected to elicit responses about familiarity, prior knowledge or reasoning that the participants had brought to or found in the tasks, as suggested by Robinson (2001). These two questions were set up as stepping stones for Q3, which was most important for collecting expert

judgement on the parallelness of the tasks.

3.3.5. Results and Discussion

From the first question, concerning what was elicited by the tasks, it became clear that the two tasks could not be considered parallel. Regarding the functions elicited by the two tasks, one group reported that the car accident task elicited argument, a different speech act from the train station task. Similarly, another group answered that the expressions used in the two tasks were quite different because the car accident task elicited a lot of passive forms which was the case in the train station task. Relating to these points, it was agreed that the functions used in the two tasks were not the same; the car accident task requires arguments and justification for both men who appear in the sequence (e.g. "It's your fault because you were talking on your mobile and weren't looking carefully.", "Your car broke the tail light of my scooter. I want compensation."), whereas the train station task tends to elicit complaints only from the owner of the briefcase (e.g. "Why did you do that?", "Look what did you do to my briefcase!"), often without justification or elaboration. This is because the man who elbowed the owner of the briefcase does not seem to argue and, in the next picture, the owner's attention is held by the briefcase as the train comes in.

For Q2, a native speaker of English reported a lack of prior knowledge concerning train stations. She mentioned that the train station task requires prior schematic knowledge about the train station with which she was not very familiar; she could not instantly come up with words such as *railroads* and *platform*. The SST is usually given in Tokyo, Japan, where the public transportation system is well developed, and the two tasks were assumed to refer to public places which most candidates would be familiar with. However, it became clear that if candidates are from suburban areas or

the countryside, where they usually use a car to travel around, then they may be disadvantaged.

Responses to Q3 further questioned the parallelness of the two SST narrative tasks, and several groups stated that the car accident task seemed more difficult. One participant found the task hard to narrate without using "I"; she started off with a third-person character but later found it difficult to sustain because she had to describe another man in detail. Although the train station task has other characters as well, they play rather minor roles. Thus, the prominence of characters in the sequences is different in the two tasks. Also, another group suggested that the car accident task makes much more cognitive demands of participants as it requires deciding what is not in the pictures (especially between Pictures 5 and 6, and how the story ends in Picture 6), expressing arguments with justification, and deciding whose fault it was and what the causes were.

The groups also disagreed on whose fault it was in the car accident task. Some said the car accident task was easier because they could see clearly who was at fault for the accident (i.e. the man on the scooter), while others were somewhat confused with how the two vehicles came to crash into each other (which direction was the car driver going? why did the scooter rider drop his mobile phone?) and therefore had difficulty in deciding who was to blame. This point suggests that there appears to be a difference in the clarity of the wrongdoer-sufferer relationship in the two tasks. In the train station task, the wrongdoer-sufferer relationship is not strongly demonstrated because the briefcase proved to be intact after all and the man who elbowed the owner of the briefcase simply appeared to be absorbed in conversation. In the car accident task, there is clear damage which will necessarily establish a wrongdoer-sufferer relationship. While the scooter rider should not have been talking on his mobile, it is not quite clear

how the crash was caused, and therefore there is difficulty in deciding whose fault it was. Therefore, the disagreement as to whether the car accident task was easier or not suggests that it may be more ambiguous than the train station task, which may mean that they are not parallel.

3.3.6. Conclusions and Suggestions for Further Research

In summary, the contributions made during the discussion suggest that while both tasks deal with a public place and vehicles which were thought to be highly familiar, candidates from urban areas may be familiar with both trains and cars, whereas those from suburban areas, where public transport is not so frequently used, might be disadvantaged. In addition, the train station task elicited complaints from one character only, often without justification, while the car accident task requires arguments with justification for both characters. This leads to a difference in the clarity of the wrongdoer-sufferer relationships and the prominence of the characters in the two tasks, and to the realisation that passive forms are required only in the car accident task. An important implication for the selection of tasks for the main study is that there are different degrees of 'seemingly-parallel'. The two SST tasks under investigation, although designed to be parallel in there being a common setting in a public place and a conflict depicted in the pictures, were not actually parallel in terms of the familiarity, prominence and relationships between the characters, and in the expressions and functions elicited. The tasks for the main study must therefore be selected according to a more rigorous standard and judged as more 'seemingly-parallel' than the SST tasks were.

Another point which the author noticed about the performance elicited was that native speakers elaborated much more on the characters' emotions and explaining the

background (e.g. "The briefcase had really important documents for me, so I was astonished when it fell on the tracks".). This emphases the importance of using smaller idea units, as suggested in Pilot Study 1, rather than the broad classification offered by Labov (1972) when examining how detailed narratives are.

3.4. Pilot Study 3: A Feasibility Study of Linguistic Performance (2): Investigating the Sensitivity of Linguistic Variables in an SST Task

3.4.1. Purpose

As discussed in Pilot Study 1, linguistic variables should be examined for their sensitivity to capture differences in the quality of linguistic performance at different levels of proficiency. In addition to linguistic performance by candidates at SST Levels 4 to 9, native speaker performance is also collected and analysed. A 'sensitive' variable is defined as a variable which correlates highly with proficiency levels (including NS level) and discriminates between speakers at different proficiency levels. This is to replicate the correlation studies with human ratings (as were performed by Kormos & Dénes (2004)) and tests for significant differences between proficiency levels (as in Laufer & Nation (1995)). The purpose here is not to compare the two SST narrative tasks, so this pilot study uses only one of the SST spoken narrative tasks so that linguistic performance by candidates at different SST Levels on the same task can be examined. It was decided to use only the car accident task in this study, as this appears 'more difficult', and might therefore be better suited to examining higher-level (i.e. SST Level 9) candidates. The linguistic variables in this study include those of fluency and lexical sophistication which were not included in Pilot Study 1.

3.4.2. Data

3.4.2.1. Japanese Candidates Data

This pilot study attempts to examine the aspect of fluency, so not only the transcripts but also the corresponding recordings were obtained and analysed. Firstly,

24 transcripts with TOEIC scores ranging from SST Levels 4 to 9[18] were identified in the NICT JLE Corpus. Then, a request to obtain the corresponding recordings was sent to ALC Press; this was accepted on the condition that the results should be reported back later. Descriptive statistics for the number of transcripts at each SST level and their TOEIC scores are shown in Table 3.8. As can be seen, the TOEIC scores are generally in line with the SST Levels.

Table 3.8
Descriptive Statistics for the TOEIC Scores of the 24 Transcripts

SST Lv.	N	Mean	SD	Min	Max
4	4	661.3	149.8	450	795
5	5	726.0	110.8	580	860
6	4	816.3	60.2	735	880
7	5	848.0	117.0	640	920
8	4	837.5	55.6	800	920
9	2	962.5	10.6	955	970

3.4.2.2. Native Speakers Data

Although the NICT JLE Corpus contains a small 'native corpus' in which several English native speakers perform SST tasks, there was insufficient native performance of the narrative task for this study. Therefore, 5 native speakers of English studying at Lancaster University participated in this study: 2 linguists and 3 non-linguists (one of whom was a former English teacher). Meeting them one by one in a quiet room, they were asked to look at the task and then narrate a story. There were no temporal limitations for the preparation or storytelling. Their narration and responses were recorded and transcribed for the purposes of analysis.

[18] Judging from the level descriptors, SST Levels 4 to 9 correspond approximately to CEFR Levels A2 to B2/C1.

3.4.3. Linguistic Variables

3.4.3.1. Fluency, Accuracy, Syntactic Complexity and Lexical Complexity

The linguistic variables for fluency are added to the analysis in this pilot study: i.e. speech rate and mean length of runs, both of which are validated in the research by Kormos and Dénes (2004). Moreover, the percentage of error-free clauses and the number of errors per AS-unit are added for accuracy, so that a more comprehensive study of the accuracy variables is possible. Other newly-added variables are the average number of subordinate clauses per AS-unit for syntactic complexity, and the percentages of frequent words as calculated by a Lexical Frequency Profile (LFP) (Laufer & Nation, 1995) and the JACET8000 Vocabulary List (JACET, 2003) for lexical complexity (i.e. lexical sophistication).

3.4.3.2. Idea Units

Regarding the idea units, smaller units of ideas, with a topic and a comment, are employed in this pilot study, as suggested by Ellis and Barkhuizen (2005). Analysis with smaller units than those used by Labov (1972) might produce different results from those seen in Pilot Study 1. Methodologically, this variable refers to how many events in the stories are covered in the candidates' narrative, which Appel (1984) used for comparing first and second performances by the same candidates. The events which occur in the car accident task are shown in Table 3.9. The shaded cells represent the 'main' idea units, which all the NS mentioned in their story. The others are treated as 'minor' idea units.

Table 3.9

Idea Units in the Car Accident Task

1	A guy was driving a car
2	Which ... [car's description (e.g. *he recently bought*)]
3	He wanted to go to ... [stating purpose]
4	He was in a hurry ... [his state]
5	Another guy was riding a scooter
6	He was talking with a girl on a mobile phone
7	He was not concentrating on the road
8	At a corner, they hit each other
9	The rider's mobile phone was broken
10	It hit the wing mirror of the car
11	The car was okay
12	They got off/out of their vehicles
13	They got angry
14	The rider complained about the broken scooter (tail light)
15	The rider complained about the broken mobile phone
16	The rider requested compensation
17	The driver insisted that it was the rider's fault
18	Because the rider was not careful enough
19	The police were called
20	Because they could not resolve the argument
21	The driver explained what happened and insisted the rider was talking on a mobile phone
22	The rider also insisted / gave up
23	The policeman took notes
24	The policeman understood / took the side of the driver
25	The policeman went back to report
26	They were asked to go to the police station
27	The driver drove off or left
28	The rider called a repairman
29	The rider's scooter was taken away by a truck
30	The repair costs would be dealt with by ... [whomever]

3.4.4. Research Questions

1. For SST levels 4-9 and native-speaker level, which variables correlate highly with the levels, and discriminate between them?

2. If the variables do not correlate highly or discriminate between the levels, how can this be explained?

3.4.5. Procedure and Analysis

Twenty-four transcripts with TOEIC scores were extracted from the NICT JLE Corpus, checked for precision with recordings, and modified where necessary. The NS performances were recorded and transcribed. All repetitions, fillers and self-corrections were removed. False starts (which were abandoned or rephrased afterwards) were identified by the author and preserved for segmenting performances into AS-units. Performance on the car accident task was separated out from the corresponding 24 recordings of SST interviews and analysed for fluency.

Linguistic variables were identified manually by the author, except for D value, coverage rates of the vocabulary in the LFP, and the JACET8000 List for lexical complexity, which were calculated using existing programs. All linguistic variables were correlated with SST Levels using Spearman's rho coefficients. For the number of idea units mentioned in the story, each transcript was given 1 if it mentioned an event and 0 if it did not, and the total numbers of idea units were correlated. Then, in order to examine which variables discriminate between SST Levels, a Kruskal-Wallis Test and later a *post hoc* test LSD were used. All the variables used in this pilot study are summarised in Table 3.10.

Table 3.10
Linguistic Variables in Pilot Study 3

Aspect	Variable	Details
Fluency	Mean length of runs	Average number of syllables produced in utterances between pauses of 0.25 seconds and above
	Speech rate	Total number of syllables produced in a narration divided by the amount of total time required to produce the speech sample (including pause time) expressed in seconds
Syntactic Complexity	Average length of AS-unit	Average number of words per AS-unit
	No. of subordinate clauses per AS-unit	Average number of subordinate clauses per AS-unit
Lexical Complexity	D value	Calculated by CLAN program on the CHILDES website at: http://childes.psy.cmu.edu/
	LFP 1, 2, 3	Percentage of words listed in the Lexical Frequency Profile Vocabulary Lists 1, 2 and 3
	JACET 8000 Vocabulary List Lv. 1-8	Percentage of words listed in the JACET Vocabulary Lists 1 to 8; these lists are based on the British National Corpus as well as the vocabulary frequently found in English textbooks, newspapers, tests and magazines available in Japan.
Accuracy	Percentage of error-free clauses	Percentage of clauses not containing any error in relation to the total number of clauses
	No. of errors per AS-unit	Number of errors divided by the total number of AS-units
	Errors per 100 words	Number of errors divided by the total number of words produced divided by 100
	Percentage of target-like use of the past tense	Percentage of verbs in the past tense in obligatory contexts (i.e. where past tense verbs are required)
Idea Units	No. of idea units expressed	Total number of main ideas that are necessary to complete the story and minor ideas that enrich the story

3.4.6. Results and Discussion

3.4.6.1. Descriptive Statistics

Descriptive statistics for the variables across different levels of proficiency are presented in Table 3.11.

Table 3.11

Descriptive Statistics for the Variables across Different Levels

		SST lv. 4		SST lv. 5		SST lv. 6		SST lv. 7		SST lv. 8		SST lv. 9		Native Speakers	
		M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
	Total no. of words	161.00	101.94	193.00	56.86	197.50	96.18	222.40	70.37	202.00	70.77	179.50	92.63	220.40	145.34
Fluency	Mean length of runs	3.25	.21	3.40	.36	3.82	.60	5.47	.43	7.15	1.73	10.30	1.57	15.65	12.23
	Speech rate	1.04	.16	1.23	.23	1.70	.44	1.99	.14	2.22	.46	2.22	.36	2.72	.35
Syntactic Complexity	AS-unit length	8.87	.61	9.72	2.55	10.10	.76	10.44	1.13	12.01	2.11	12.55	4.55	10.22	1.89
	Subordinate clauses per AS-unit	.10	.05	.28	.11	.34	.12	.38	.18	.44	.25	.35	.26	.24	.08
Lexical Complexity	D value	32.80	12.07	34.72	10.61	29.24	8.73	36.54	14.56	41.15	12.12	30.40	8.33	45.59	27.96
	LFP 1	76.85	4.25	78.45	2.15	83.59	2.91	80.84	2.52	81.56	1.91	78.39	7.51	78.87	3.95
	LFP 2	12.39	2.29	12.81	2.42	9.93	3.25	11.32	1.56	9.32	2.01	12.58	3.96	10.56	1.61
	LFP 3	1.00	1.19	1.70	1.58	.53	1.07	1.21	.31	2.00	1.59	1.93	2.72	1.47	1.66
	Out of LFP	9.77	4.56	7.03	.88	5.96	3.04	6.63	1.69	7.13	1.88	7.11	.83	9.10	2.26
	JACET8000 Lv.1	76.65	5.40	76.41	3.02	78.18	4.88	79.63	2.88	78.70	1.61	80.96	.80	74.67	5.29
	JACET8000 Lv.2	10.52	4.08	10.28	2.32	9.07	2.38	9.79	2.82	7.85	.81	5.22	1.84	7.98	2.49
	JACET8000 Lv.3	3.54	2.86	3.58	2.17	2.37	2.02	2.78	1.39	3.83	1.64	2.61	.92	1.99	1.27
	JACET8000 Lv.4	.81	.94	.93	.90	1.57	.61	.37	.51	.00	.00	2.50	2.00	2.01	1.19
	JACET8000 Lv.5	2.49	.60	2.48	.84	2.84	1.57	1.93	.91	2.97	.32	2.50	2.00	3.37	1.36
	JACET8000 Lv.6	1.94	1.60	1.25	1.20	.85	1.13	.90	.56	1.59	1.19	2.61	.92	2.17	1.30
	JACET8000 Lv.7	1.20	1.02	.97	.96	.25	.51	.78	1.12	.73	.84	.98	1.39	1.11	.93
	JACET8000 Lv.8	.00	.00	.86	.85	.55	.64	.69	.67	1.23	1.62	1.09	1.54	.12	.27
	Out of JACET List	.79	1.59	1.83	2.08	3.47	1.68	2.03	1.33	2.03	.84	.98	1.39	3.07	1.42
Accuracy	% of error-free clauses	53.33	25.39	67.34	11.71	76.13	4.47	80.16	9.73	67.43	6.55	94.70	1.35	100.00	.00
	Errors per AS-unit	.54	.31	.46	.17	.42	.11	.33	.13	.57	.13	.08	.00	.00	.00
	Errors per 100 words	5.43	3.26	4.02	1.65	3.43	1.07	2.82	1.02	4.06	.81	.64	.33	.00	.00
	% of target-like use of past tense	69.98	25.51	78.15	22.02	88.50	11.32	95.44	5.01	78.32	13.46	96.43	5.05	100.00	.00
Idea Units	No. of main idea units	4.75	.96	5.20	.45	5.00	.82	4.40	1.14	5.25	.96	5.00	.00	6.00	.00
	No. of minor idea units	5.00	2.00	6.60	2.88	5.00	2.00	8.80	3.03	5.00	.82	6.00	1.41	7.80	4.55

3.4.6.2. 'Sensitive' Variables

Table 3.12, below, summarises the results from the Spearman's rho tests, the Kruskal-Wallis tests, and the *post hoc* LSD tests. With *post hoc* LSD tests, the levels that showed a significant difference in their respective means are listed.

Table 3.12

Results for Correlations with and Discrimination between the Different Levels

Aspect	Variables	Spearman's		Kruskal-Wallis		*post hoc* LSD
		r	*p*	$\chi^2(6, 29)$	*p*	Discriminated between[3]
Fluency (Temporal)	Mean length of runs	.913**	.000	24.35*	.000	4-NS, 5-NS, 6-NS, 7-NS, 8-NS
	Speech rate	.894**	.000	22.67*	.001	4-6, 4-7, 4-8, 4-9, 4-NS, 5-6, 5-7, 5-8, 5-9, 5-NS, 6-8, 6-NS, 7-NS, 8-NS
Syntactic Complexity	AS-unit length	.464*	.011	8.96	.176	
	Subordinate clauses per AS-unit	.264	.166	12.68*	.048	4-6, 4-7, 4-8
Lexical Complexity	D-value	.156	.418	2.63	.853	
	LFP 1	.105	.588	9.51	.588	
	LFP 2	-.271	.154	7.15	.154	
	LFP 3	.122	.528	3.04	.528	
	Out of LFP	.086	.656	7.63	.656	
	JACET8000 List Lv.1	.015	.938	5.64	.464	
	JACET8000 List Lv.2	-.384*	.040	8.31	.216	
	JACET8000 List Lv.3	-.216	.260	4.72	.580	
	JACET8000 List Lv.4	.202	.293	13.89*	.031	4-9, 5-9, 6-8, 7-9, 7-NS, 8-NS
	JACET8000 List Lv.5	.235	.220	5.33	.502	
	JACET8000 List Lv.6	.222	.247	6.98	.322	
	JACET8000 List Lv.7	.007	.969	3.29	.771	
	JACET8000 List Lv.8	.038	.843	6.28	.392	
	Out of JACET List	.244	.203	7.24	.299	
Accuracy	% of error-free clauses	.660**	.000	19.78**	.003	4-6, 4-7, 4-9, 4-NS, 5-9, 5-NS, 6-NS, 7-NS, 8-9, 8-NS
	Errors per AS-unit	-.553**	.002	17.75**	.007	4-9, 4-NS, 5-9, 5-NS, 6-9, 6-NS, 7-8, 7-NS, 8-9, 8-NS
	Errors per 100 words	-.638**	.000	17.72**	.007	4-7, 4-9, 4-NS, 5-9, 5-NS, 6-9, 6-NS, 7-NS, 8-9, 8-NS
	% of target-like use of past tense	.528**	.003	14.40*	.025	4-7, 4-NS, 5-NS
Idea Units	No. of main idea units	.368*	.049	9.91	.129	
	No. of minor idea units	.155	.422	6.04	.419	

Note. *Significant at .05 level.

**Significant at .01 level.

[3]Numbers indicate the SST levels. NS=native speakers.

For RQ1, the 'sensitive' measures, those with high correlation with and discrimination between the levels, were: fluency variables (i.e. the mean length of runs and speech rate) and accuracy variables (i.e. the percentage of error-free clauses, errors per AS-unit, errors per 100 words, and percentage of target-like use of past tense). Both fluency variables showed a very high correlation (r = .894 and .913), and the accuracy variables had moderately high correlation ($|r|$ = .528 to .660).

However, when the LSD columns were closely examined, it became clear that most of these measures were only able to discriminate between a limited number of levels. Mean length of runs, one of the fluency variables, only discriminated between SST levels and native speakers. It could not differentiate between Japanese learners of English. The same applies to the percentage of target-like use of the past tense, which discriminated between even fewer levels (SST levels 4, 5, 7 and the native speakers). If they can only discriminate between distant levels, such as the lower-level candidates and the native speakers, these measures may only poorly capture differences in learner performance. The other three accuracy variables discriminated between more pairs of levels; each had 10 pairs listed out of 20 possible pairs. Still, they seldom succeeded in differentiating adjacent learner levels (i.e. SST levels 4-5, 5-6, 6-7, etc.), especially at lower levels. Compared to the rest of the measures discussed above, speech rate, the other fluency measure, discriminated more often between levels that were closer to each other. Together with its high correlation (r = .894), speech rate may be considered the 'most sensitive' variable.

It is clear now that most of the variables that more or less satisfied the two conditions of 'sensitivity' actually failed to demonstrate good discriminating power, so the patterns that they display across the levels should be examined according to RQ2: why did they not discriminate well?

The patterns shown are based on the means that were introduced in Table 3.1 for each variable. Firstly, Figure 3.1 displays the patterns for the fluency variables.

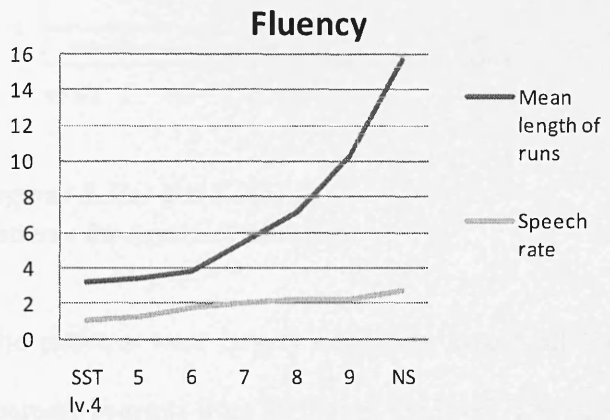
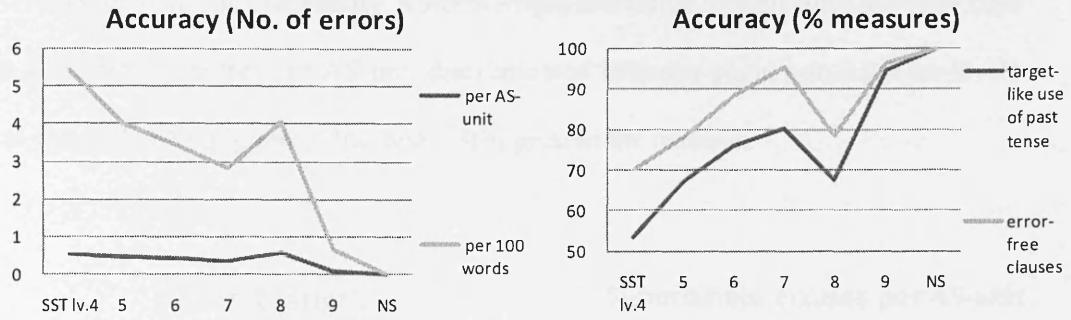


Figure 3.1
Patterns for Fluency Variables

While the speech rate increased steadily from SST level 4 to NS level, the mean length of runs showed a drastic increase between SST level 9 and NS level. It is probable that this caused the relatively small differences between the other levels to be non-significant. This may suggest that, even if a measure correlates very highly, it does not necessarily guarantee its 'sensitivity' for distinguishing between the different levels of performance by the Japanese candidates.

Accuracy variables had moderately high correlations, but they did not discriminate between SST levels, either. Figures 3.2(a) and 3.2(b) present the patterns.



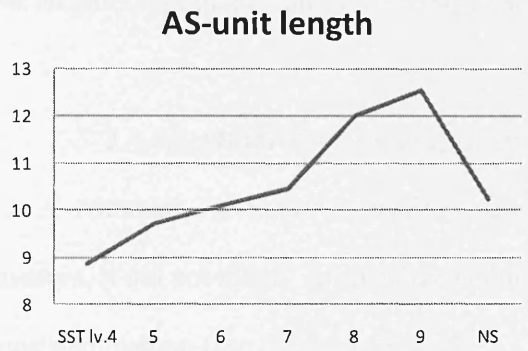
Figures 3.2(a) and 3.2(b)
Patterns for Accuracy Variables

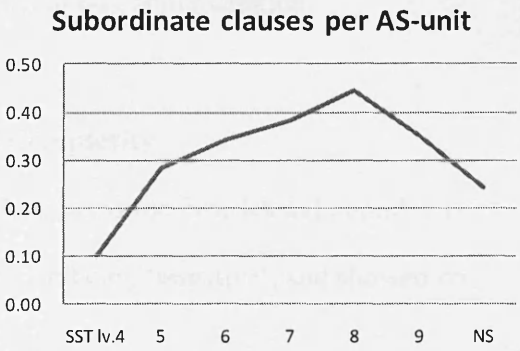
The patterns were largely consistent across all four variables. There was a steady decrease in errors from SST level 4 to level 7; however, at level 8, there was an increase in errors. This is surprising, as one might assume that the higher the candidate's level, the fewer errors they would make in their performance. One explanation is that, judging from the larger means for syntactic complexity variables at SST level 8 in Table 3.11, the SST level 8 candidates might have attempted to use more complex structures than the lower level ones but failed to use them accurately. It is possible that up to SST level 7, candidates may tend to avoid trying out new structures or items and prefer to speak with the ones that they are familiar with and confident in using. Although exploring this assumption would be of interest, it is beyond of the scope of this thesis as it would require scrutinising structures and error types with a larger sample.

3.4.6.3. Other Variables: Syntactic Complexity

The rest of the variables were not proven 'sensitive' according to the operationalisation in this pilot study. Some variables satisfied only one of the two conditions for being 'sensitive', and others did not satisfy either of them. The patterns are examined as to why they could not satisfy the conditions in the next two sections.

As for syntactic complexity, AS-unit length correlated significantly ($r = .464$) and the subordinate clauses per AS-unit discriminated between some non-adjacent levels (i.e. 4-6, 4-7, 4-8). Figures 3.3(a) and 3.3(b) present the patterns.





Figures 3.3(a) and 3.3(b)
Patterns for Syntactic Complexity Variables

AS-unit length displayed a steady increase among the Japanese candidates, but not with the NS. Subordinate clauses per AS-unit showed a very similar pattern, except that they start to decline at SST level 9.

It is quite interesting that the NS performance was less complex than that of higher-proficiency candidates (i.e. SST levels 7-9) according to this variable. One possible explanation for this is the differences in conditions under which the task was given. Compared to the SST candidates, who were under pressure to prove their language proficiency within a limited time, the NS performed the task with no limits on planning or presentation time. This suggests that the conditions of task administration should be controlled for all candidates in future research.

Alternatively, the less complex performances by the NS could be attributed to the task requiring narration, which might not encourage individuals to use complex language in the first place. What is more, it might be that, contrary to our intuitive

expectations, NS do not usually produce syntactically more complex speech than high-level candidates, whether they are given the task in the same situation or not. NS may be more prone to being 'economical' with language, with little intention to produce complex speech in most situations. This issue deserves further investigation, and again, this should be examined under the same conditions of task administration.

3.4.6.4. Other Variables (2): Lexical Complexity

Although Jarvis (2002) justifies using D value as the best lexical complexity measure, it did not satisfy either of the conditions for being 'sensitive', and showed no consistent pattern (see Figure 3.4).

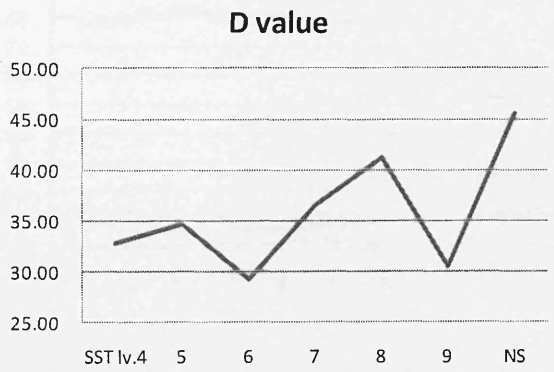
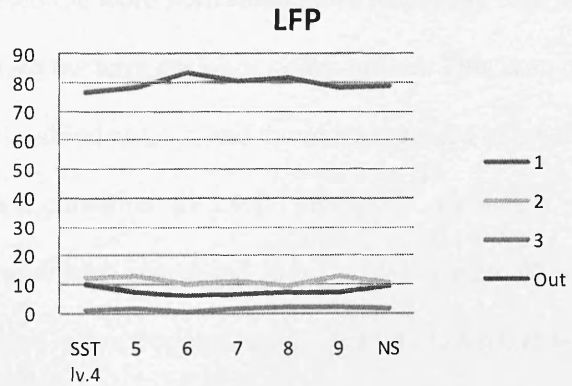


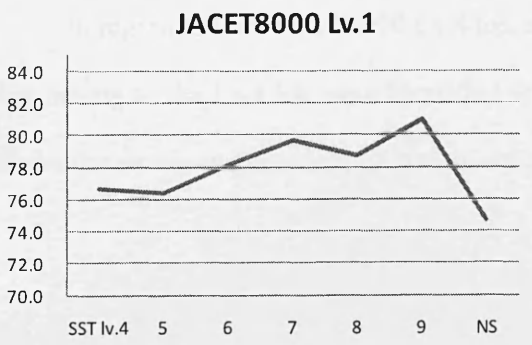
Figure 3.4
Pattern for D value

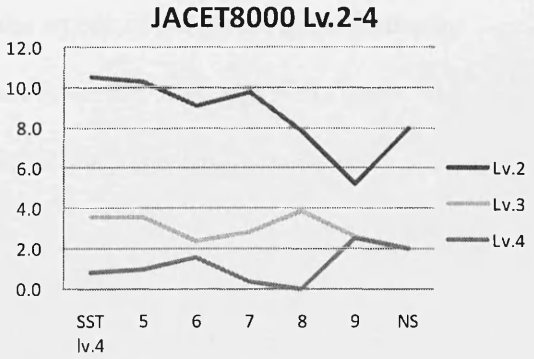
Two reasons might explain this result. Firstly, D value is meant to be used for measuring lexical variety. Thus, it may not be suitable to apply it to spoken performance on a narrative task because the content is largely predetermined and the vocabulary range cannot be expected to vary as much as with tasks with more freedom to produce a wider variety of content. Secondly, the time limit for the interviews could have influenced some SST candidates. Since the narrative task is given in the last stage of an SST, there can be different degrees of urging by the interviewer, depending on how much time is

left. If, for example, SST levels 6 and 9 candidates (who scored lowly on D value) had to finish telling the story quickly, then they might not have been able to demonstrate fully the vocabulary range that they possessed. The NS, who told a story to the author with no time limit, might have been able to demonstrate their vocabulary range more fully. This issue needs to be examined with new sets of data, obtained under the same conditions and with no time limit for narration.

The frequency-based variables of lexical sophistication presented rather flat patterns across the levels, as shown in Figures 3.5(a) to 3.5(c).







Figures 3.5(a), 3.5(b), and 3.5(c)
Patterns for Lexical Complexity (Vocabulary Lists)

The figures suggest that the Japanese candidates and the NS used more or less similar

levels of vocabulary according to LFP. This is in line with the discussion earlier about D value; since the content is pre-specified, the vocabulary range is decided by the task to some extent, thus leading to the use of similar vocabulary across different levels. However, JACET8000 drew somewhat different patterns. Its Lv.2 list had a moderate negatively significant correlation, and its Lv.4 list discriminated between some levels. In order to find out why these phenomena were related to a particular level of vocabulary, it was decided to examine the lists of actual words observed with their frequencies.

By scrutinising JACET8000 lists, it was shown that lower-level SST candidates used the word *policeman* more frequently than higher-level candidates who more often used the term *police* or *police officer*. This is an unexpected result because *policeman* is classified as Lv.2, and therefore regarded of lower frequency, whereas *police* and *officer* are classified as Lv.1. Therefore, according to the JACET8000 lists, lower-level candidates succeeded in using 'less frequent' words, whereas higher-level candidates used 'more frequent' words, leading to a negative correlation. It calls for caution in that results can be hugely influenced by such slight differences in the words used.

In regard to the JACET8000 Lv.4 list, all the words in the narrative performance that belong to the Lv.4 list were identified at each level (i.e. SST and NS). Table 3.13 shows the words and numbers of their occurrences in the transcripts at each level.

Table 3.13

JACET8000 Lv. 4 Words Used at Each Level

SST Lv.	Words	Occ.	SST Lv.	Words	Occ.
4	due	1	8	–	–
	running	1	9	accuse	2
	waiting	1		gay	1
5	clash	1		illegal	1
	coming	2	NS	clash	1
	fixed	1		coming	5
6	coming	1		compensation	1
	insurance	4		insurance	1
	let's	1		let's	1
	spite	2		resolve	1
7	negotiate	1		ridiculous	2
	used	1		smash	2

Note. Occ. = No. of occurrences.

Table 3.12, presented earlier, indicates that the JACET8000 Lv.4 list discriminated between SST levels 4-9, 5-9, 6-8, 7-9, 7-NS, and 8-NS. The numbers of occurrences appear different between SST levels 7-NS, 6-8, and 8-NS. There are hardly any differences at SST levels 4, 5, and 9. However, as the numbers of transcripts differed (i.e. SST lv.4 = 4; lv.5 = 5; lv.9 = 2), the resultant percentage of JACET8000 Lv.4 words was larger at SST level 9.

This, again, raises questions about using such word lists to identify which levels of words speakers are able to produce during narration. In addition to the discussion on predetermined vocabulary range by task, there is an issue of the selective use of words by learners. The SST level-8 candidates in this study did not use any JACET8000 Lv.4 words, but it does not necessarily imply that they did not have any lexical knowledge of them. The same applies to SST level-7 candidates who did not use many words from JACET8000 Lv.4. In sum, rather than expecting to find meaningful differences in lexical use among different proficiency levels with these variables, it would be more sensible to analyse narrative performance qualitatively. For example, we might explore

if there are any differences in the expressions used to describe the same characters, items or events in the story at different levels.

3.4.6.5. Other Variables (3): Idea Units

The last variable that is discussed here is idea units: the numbers of main and minor idea units. As plotted below in Figure 3.6, most of the SST candidates covered more than four main idea units out of six (by NS performance), which means that even lower-level learners could convey the essential events of the story to some extent. The minor idea units showed more variation.

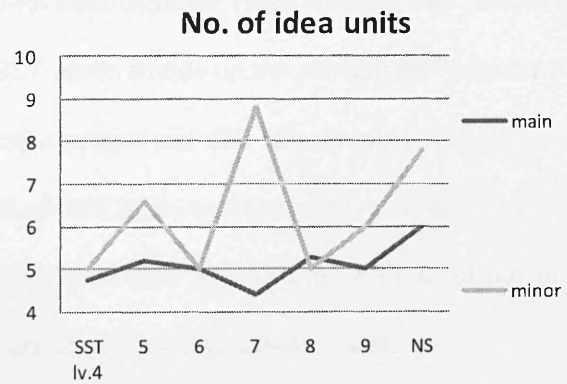


Figure 3.6
Patterns for Idea Units

The numbers of idea units closely relate to how much the candidates talked. Judging from the means of the numbers of words in Table 3.11, SST level-8 candidates talked less than level-7 ones did, which explains why the level-8 candidates produced lower numbers of minor idea units. As the time available for narration may well influence the number of words produced, the patterns of this measure emphasise again the importance of task administration to standardise the conditions for every speaker.

3.4.7. Conclusions and Suggestions for Future Research

To sum up, this pilot study has found that the only 'sensitive' measure having high correlation with proficiency levels and high discrimination between many levels is speech rate. Among the rest of the variables, some either correlated highly with the SST levels but could not discriminate, or could discriminate to some extent but did not correlate highly with the levels. Others satisfied neither of the conditions and thus were not 'sensitive'.

The major limitation of this study lies in the fact that it uses SST levels as a reference measure for the quality of Japanese learners' performance on a narrative task. As explained in Section 3.2.2.1, an SST level is an overall rating for an entire interview with three different types of tasks. The narrative task is only one of them. Although the SST raters decide on the provisional level for the performance on each task type, they are averaged out and not revealed within the final SST level. So, it is possible for a candidate to do well (or poorly) on the narrative task but poorly (or well) on the other tasks, and then for the final SST level not to reflect specifically on the quality of performance on the narrative task.

What is more, SST raters use analytic scales (explained briefly in Section 3.1.1.) for different aspects of the performance on each task, but these 'sub-levels' are not revealed either. Therefore, SST levels cannot provide information on how the candidate is profiled in different aspects of their performance. By employing SST levels as the reference measure for correlation, this study implicitly presupposed that there was a linear increase in complexity or decrease in errors as the levels go up, which is not always the case in second language research (Fulcher, 2003: 103). Thus, it is highly desirable to rate the narrative separately and then to use those ratings, rather than SST levels which are decided after considering performance on other tasks in the interview.

In addition to having ratings solely based on narrative performance, three suggestions are made for future research. Firstly, the task should be given under the same conditions for every speaker. It may be especially important to allow speakers to talk as much as possible with no test-like pressure or time limit for narration, so that the measures for syntactic complexity and idea units can be fully explored without the possible interference of pressure or time. The second suggestion is to run more qualitative analyses, especially for lexical complexity, rather than relying on word lists for meaningful information about differences in how the story is told. Lastly, the design of this study needs to be replicated but with a larger sample size in order to verify whether the patterns observed in this study can be generalised. Variables for accuracy and syntactic complexity may benefit the most from this suggestion.

Although this pilot study has its limitations, its contribution is unique in that it has systematically and empirically examined various variables for their sensitivity. The main study should include a similar design for RQ4, but with a larger sample size and ratings based solely on narrative performance. Stricter controls on task administration are indispensable, as these will facilitate building on the conclusions drawn from this pilot study, which is likely to have important implications for the field of language testing and task-based research.

3.5. Pilot Study 4: Native Speaker Performance and Perceptions of the Two SST Tasks

3.5.1. Purpose

This pilot study examines the linguistic performances by and perceptions of native speakers of English (NS) on the two SST narrative tasks. According to the expert judgements in Pilot Study 2, the two tasks were not parallel due to differences in the prominence among and relationships between the characters, and it appeared that the car accident task was more difficult. By comparing NS performance on and perceptions of the two tasks, this pilot study aims to investigate whether such differences in task design would elicit different linguistic performances by (in terms of some of the variables used in Pilot Study 3) and perceptions of the NS, who are free from L2 processing load, and to collect baseline data for the two SST tasks.

3.5.2. Data

The same 5 native speakers of English who participated in Pilot Study 3 (with the train station task) also performed the car accident task: 2 linguists, 1 former English teacher, and 2 non-linguists. They were met, one by one, in a quiet room, and asked to look at the train station task and then narrate a story, and then do the same for the car accident task. There were no limits on duration for preparation or narration. After completing both tasks, they were asked if they thought either of the tasks seemed more difficult and, if yes, why. Their narrations and responses were recorded and transcribed for analysis. The sample transcripts of a native speaker are shown in Appendix 3.

3.5.3. Research Questions

Do the two SST spoken narrative tasks elicit parallel performances from the NS in terms of linguistic variables? Do the NS perceive the two tasks to be equally difficult?

3.5.4. Linguistic Variables

In this pilot study, the mean length of run (fluency) and the two frequency-word lists (i.e. JACET8000 and LFP) were excluded from the analysis because they were not proven to be very informative in Pilot Study 3. The remaining variables included speech rate, AS-unit length, the number of subordinate clauses per AS-unit, D value, and the numbers of main and minor idea units.

3.5.5. Procedures and Analysis

The same procedures used in Pilot Study 3 were repeated to segment performances into AS-units. Once all the linguistic variables had been calculated for the 10 transcripts (i.e. 2 tasks for each of 5 native speakers), Wilcoxon signed-rank tests for related samples were conducted to see if there were any differences between the two tasks. The NS perceptions were collected and summarised.

3.5.6. Results and Discussion

3.5.6.1. Syntactic Complexity and Reasoning

The results of the Wilcoxon tests are shown in Table 3.14. Statistically significant difference was found with the number of subordinate clauses per AS-unit.

Table 3.14
Results for Wilcoxon Signed-rank Tests

	Variable	Task	*N*	Mean	SD	*Z*	*p*
	No. of words	train	5	225.40	156.37	-.132	.893
		car	5	220.40	145.34		
Fluency	Speech rate	train	5	2.71	.56	-.135	.893
		car	5	2.72	.35		
Syntactic complexity	Average length of AS-unit	train	5	8.71	1.19	-1.214	.225
		car	5	10.22	1.89		
	No. of subordinate clauses per AS-unit	train	5	.08*	.08	-2.032	.042
		car	5	.24*	.08		
Lexical complexity (variety)	D value	train	5	51.79	13.10	-.405	.686
		car	5	45.59	27.96		
Idea Units	Main idea units	train	5	6.00	.00	.000	1.000
		car	5	6.00	.00		
	Minor idea units	train	5	7.20	3.90	-.948	.343
		car	5	8.40	6.11		

Note. *Significant at $p < .05$ level.

The reason why this variable revealed a difference is partly because subordination is more likely to be used in the car accident task due to the need to describe the two main characters in detail, unlike in the train station task. So, references such as *the man who was driving the car* tend to appear frequently, increasing the number of subordinate clauses. This may relate to the contention made in Pilot Study 2 concerning the prominence of the characters; in the car accident task, the prominence of both wrongdoer and sufferer was clearer. Moreover, the car accident task appeared to have required more reasoning, eliciting *because*-clauses such as "the rider wasn't pleased because he thought he was driving along the right side of the road". This, again, is related to the expert opinion of Pilot Study 2 that this task involved justifying and explaining, whereas the train station task did not. This may be because both characters in the car accident task look angry and argue loudly, while in the train station task, the man who elbowed the owner of the briefcase does not seem so upset by being accused,

and thus not often eliciting any argument between the characters. In addition, brief interviews with the NS afterwards confirmed that the car accident task was perceived to be more difficult and confusing in that it requires deciding what happened between Pictures 5 and 6 (i.e. where the police were called and the scooter was towed away), which also conforms to the opinions raised about Pilot Study 2.

Therefore, it appears that the higher reasoning demands and clear wrongdoer-sufferer relationship in the car accident task led to eliciting more subordination from the NS. Such reasoning is part of Robinson's (2007) task complexity factors, and this finding seems to be in line with his (and Skehan's (2009)) that more complex tasks may elicit more complex language. Together with the findings in Pilot Study 2, it is clear that the two SST tasks are not parallel in terms of the difficulty perceived by the NS, expert judgements, and syntactic complexity of the elicited performances.

3.5.6.2. Idea Units

There was an issue which arose in the course of analysis in this pilot study concerning the variable of idea units. Identifying the idea units was conducted following the recommendations of Ellis and Barkhuizen (2005), and those idea units which all the NS included in their narration were regarded as the 'main' idea units. All the other idea units were listed and classified, and treated as 'minor' idea units. There were 6 main idea units in each task, corresponding to the number of pictures in the sequences. However, the numbers of minor idea units differed; while the car accident task included 24 minor idea units, the train station task had only 17. Table 3.15 lists the idea units for the train station task (see Table 3.8 in Pilot Study 3 for the car accident task). The shaded cells represent the main idea units.

Table 3.15

Idea Units in the Train Station Task

1	A man went to the train station
2	The man got down onto the platform
3	The man was waiting for the train
4	Another man who was talking to his friend (unintentionally) pushed his arm
5	The man's briefcase fell onto the track
6	The man panicked
7	because in the briefcase he had some important papers
8	The other man apologised/said it wasn't his fault
9	The other man's friend was looking puzzled
10	They were arguing
11	He thought of getting down onto the track himself
12	But he was unsure whether it was safe
13	The train came in
14	The man was appalled
15	He was worried about what his boss would say
16	The briefcase was intact
17	The man was very happy / somewhat embarrassed
18	Everyone else was looking round smiling
19	A station staff got down to the rail
20	The man thought he could do that himself
21	The station member of staff gave his briefcase back
22	The man who pushed his arm still didn't apologise
23	The man hurried to work

The average number of minor idea units in the narration was 7.2 out of 17 in the train station task, and 8.4 out of 24 in the car accident task. A larger number of minor idea units in the car accident task may be due to the differences in the amount of reasoning and extra functions (i.e. justifying and explaining) that the characters have to perform to make sense of the pictures. Thus, the difference in the number of minor idea units is in line with the expert judgements and NS perceptions; however, comparisons of raw numbers cannot be meaningfully conducted unless the denominators are the same. Thus, it is indispensable that the number of idea units is examined at the stage of task selection for the main study.

3.5.7. Conclusions and Suggestions for Further Research

This pilot study has revealed that the two SST narrative tasks appear to have elicited parallel performances from the NS in terms of fluency and lexical complexity. However, the number of subordinate clauses per AS-unit revealed a significant difference. This can be interpreted as a very strong tendency, considering that it was found with such a small sample size (N = 5) and with a non-parametric test. It was confirmed by the brief interviews afterwards that the NS felt the car accident task was more difficult, given the time gap between the pictures and the reasoning required due to the relationships between the characters. It was shown that, together with the expert judgements (Pilot Study 2), collecting NS perceptions can add to the understanding of what the tasks require; therefore, these should be collected in the main study too.

Investigating the two SST tasks via Pilot Studies 2, 3, and 4 demonstrated that there is a clear need for 'more similar' tasks for the main study, in terms of the functions expressed by each character and the prominence of and relationships between the characters. The next pilot study searches for and trials two pairs of seemingly 'more' similar tasks than the SST tasks, out of which a pair is selected for the main study, after collecting approximate numbers of idea units, subordinate clauses and characters, as well as expert judgements and narrators' perceptions of task difficulty.

3.6. Pilot Study 5: Selecting the Narrative Tasks for the Main Study

3.6.1. Purpose

This pilot study aims to select a new set of narrative tasks for the main study. Pilot Study 2 showed that the SST narrative tasks were quite different; they elicited different functions, leading to different relationships between characters and giving rise to differences in perceived difficulty among the candidates. Thus, in order to minimise such differences, it is desirable for the narrative tasks to have relationships between characters, functions and storylines that are as similar as possible. This pilot study examines the performance of two pairs of narrative tasks in terms of their similarities to help choose a pair to be used in the main study.

3.6.2. Research Questions

1. Does each pair of the tasks elicit the same number of objects, characters, events and subordinate clauses in narration by the same narrators?
2. Is each pair of tasks parallel in terms of participants' perceived difficulty?

3.6.3. Tasks

Both pairs of tasks were taken from Hill (1960) and are those which Ortega (2005) used in her research. Ortega is one of the few researchers who reveals where the tasks used in her study were chosen from. Hill intended the narrative tasks in her book to "be used with students beginning preparation for the Cambridge Lower Certificate examination,[19] or students who have done three or four years of English along efficient

[19] This is currently called the First Certificate in English (FCE) in Cambridge examinations

lines" (Hill, 1960: 2) for oral and writing compositions. In addition, the vocabulary intended to be used in each task in her book is all in the General Service List of English Words[20] (Hill, 1960: 2). Therefore, the narrative tasks in Hill's book can be expected to be parallel in terms of target levels of English proficiency and vocabulary levels.

The two pairs of tasks from Hill (1960) were selected for this pilot study because each pair shares a similar storyline. Using tasks that are as similar as possible was recommended by Pilot Study 2 for the parallelness of their respective content. Moreover, they also appear similar in terms of the factors listed by Brown and Yule (1983: 37-53) that affect candidates' performance and which are relevant to narrative tasks requiring description and narration: 1) the number of objects, characters and events; 2) whether or not the same setting is maintained throughout the story; 3) whether or not the characters are of the same type, usually in terms of gender; 4) whether or not the pictures contain something culturally unfamiliar to candidates. It was decided to use Brown and Yule's factors because of their high specificity regarding the characteristics in the pictures.

3.6.3.1. Tasks 1 and 2

The first pair of tasks, Tasks 1 and 2 (as shown in Figures 3.7 and 3.8), share a similar story of two kids playing a trick on their mother. In Task 1, a mother does the laundry, hangs it outside, and goes back inside. Some time later, two children find a man who is selling balloons near the house so they buy a balloon to play a trick with it. The girl paints a face on the balloon and the boy takes a shirt from the drying laundry. Then, they attach the shirt to the balloon with a face to make it look like a man. They take it to the window of the room where their mother is, and she is very surprised to see a strange man suddenly appear by her window.

(UCLES, 2001b).

[20] A list of 2,000 most frequent words developed by Michael West in 1953 (Bauman, n.d.).

In Task 2, a mother is sitting reading a book in a chair in front of the fireplace in a room with a baby sleeping in a basket at her feet. After some time she falls asleep, and then two children come into the room and find both the mother and the baby sleeping. They decide to play a trick. The girl quietly picks up the baby, and the boy brings a ball with a grinning face and places it in the basket instead of the baby. When the mother wakes up, she is very surprised and screams because her baby's head has changed into a strange grinning ball.

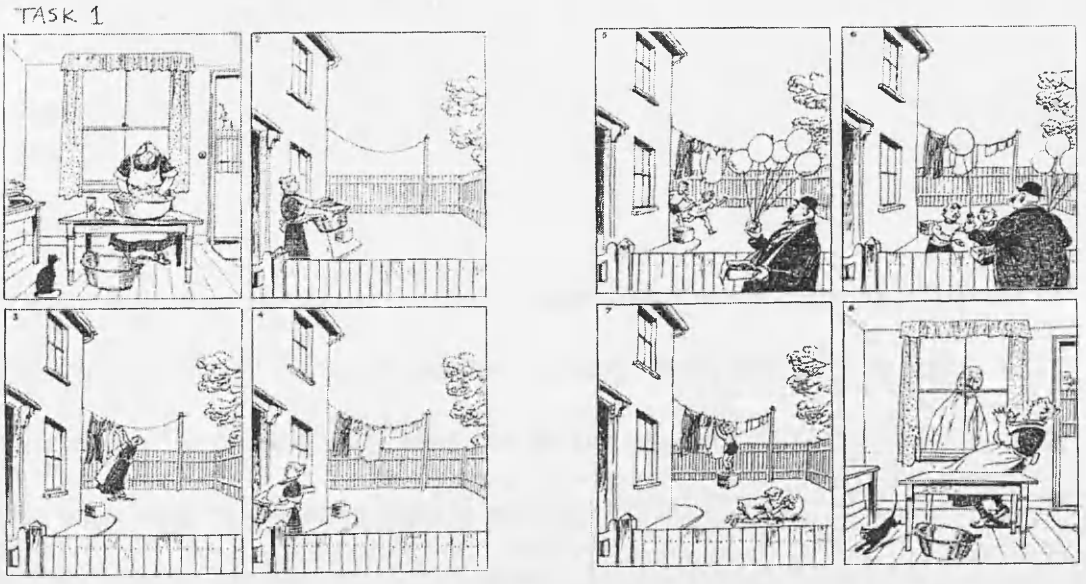


Figure 3.7
Task 1

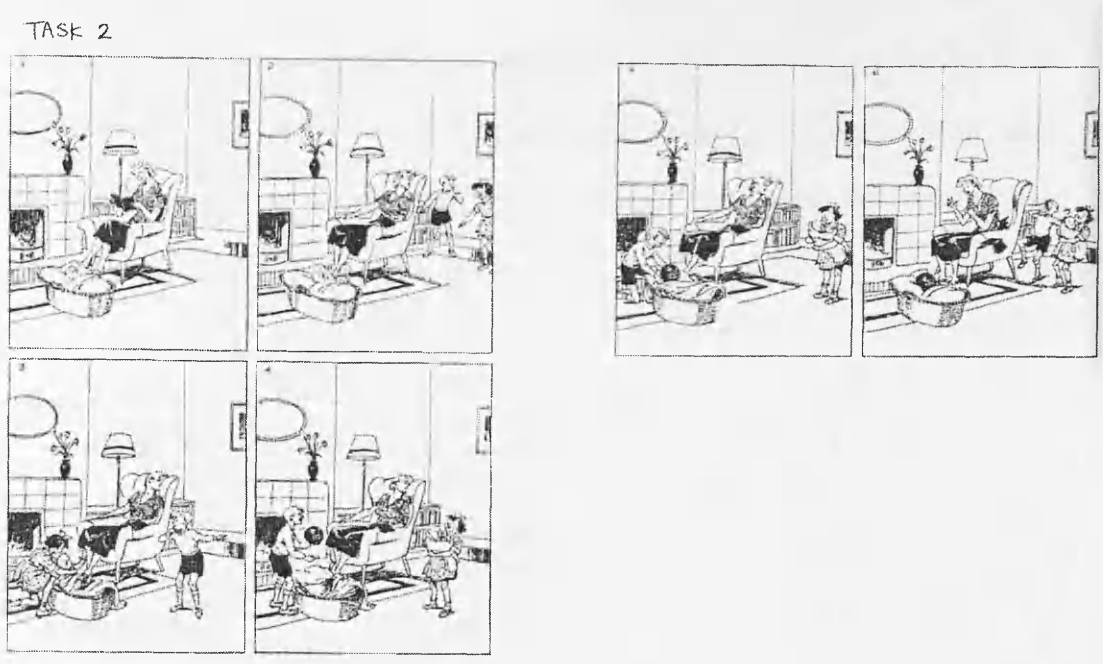


Figure 3.8
Task 2

To summarise, these two tasks appear to have approximately the same number of objects (laundry, balloon, shirt, window / book, chair, basket, ball), three main characters (a mother and two children), and the same number of events (the mother was doing something; two children came in and played a trick; the mother was surprised). Also, the pictures in both tasks do not appear to contain anything culturally unfamiliar to participants because they depict events in a normal family house.

The only aspect in which the two tasks do not match are the changes in the setting. While the same setting was maintained in Task 2 (in the living room), it switched between inside and outside the house in Task 1. Brown and Yule (1983: 44) suggest that if a task contains a change of setting, it is more cognitively difficult because the narrator has to assess the relevant elements in the new setting and the effect on the story. Therefore, Task 1 might be perceived as being more difficult by participants. Alternatively, the setting changes within the same house or property, where the room

appears to be adjacent to the garden, so this might not affect the cognitive load as much as Brown and Yule (1983) suggest. This clearly needed to be investigated in this pilot study. Table 3.16, below, summarises the similarities of Tasks 1 and 2 according to the criteria suggested by Brown and Yule (1983).

Table 3.16
Summary of Similarities between Tasks 1 and 2 according to Brown and Yule (1983)

Criterion	Tasks 1 and 2
No. of objects, characters, events	Same
Setting	Changes in Task 1; possibly more difficult
Type of characters (gender)	Same
Cultural familiarity	Same

3.6.3.2. Tasks 3 and 4

The other pair of tasks, Tasks 3 and 4 (as shown in Figures 3.9 and 3.10), also shares a similar storyline in which a couple (male and female) lose an important possession and later find it unexpectedly. In Task 3, a boy and a girl go out to a toy shop. They do not intend to take their dog with them, so they stop and wave it away when it follows them on their way, and the girl forgets to pick up the purse that she has put on the ground. She notices that she does not have the purse in the toy shop, but has no idea where she has left it. Then, their clever dog comes in with the purse and they are pleasantly surprised.

In Task 4, a woman and a man are fishing from a boat when the woman accidentally drops her bracelet into the sea. She cannot catch it before it sinks under the water. She starts crying just as the man catches a fish. They get back to the beach where they are camping and start cooking the fish. Then, the woman finds her bracelet in the stomach of the fish, which must have swallowed it. She becomes cheerful again and the

man is happy too.

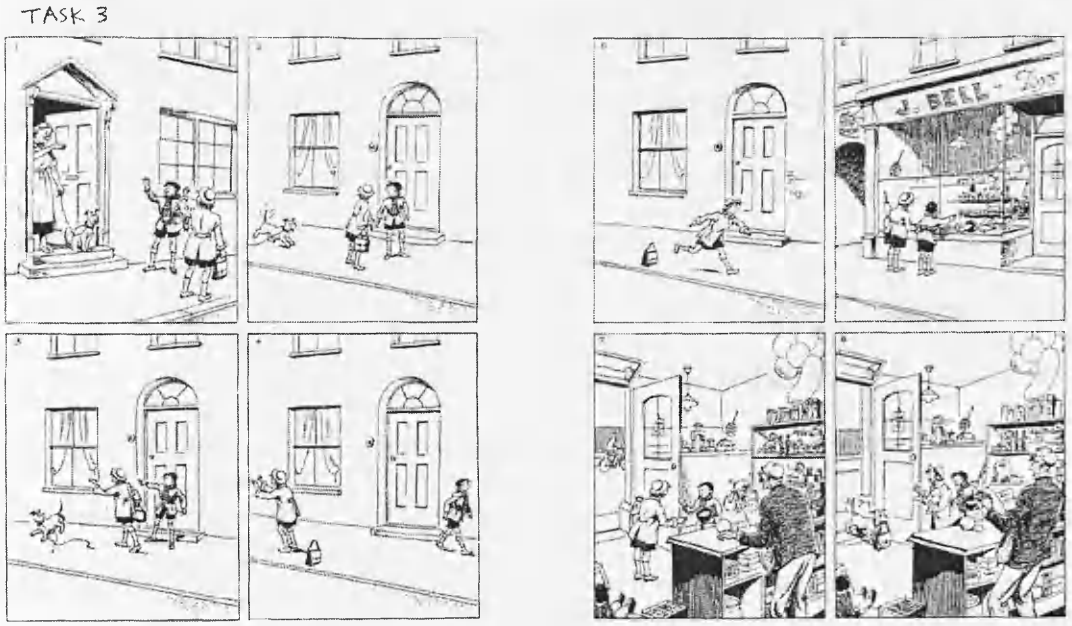


Figure 3.9
Task 3

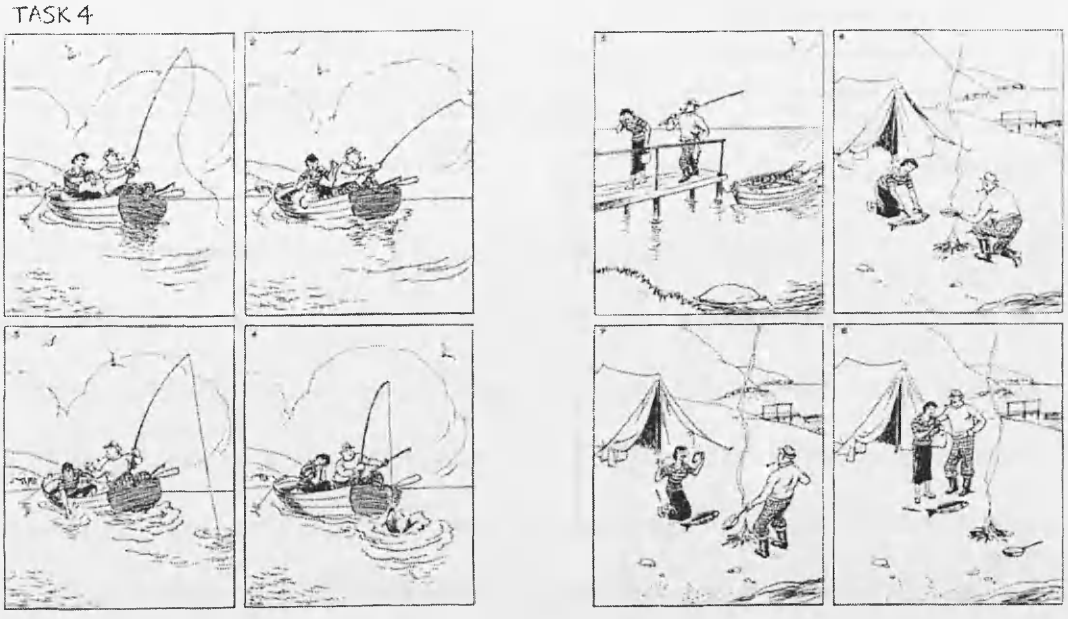


Figure 3.10
Task 4

Compared to Tasks 1 and 2, Tasks 3 and 4 have less in common; the locations (street; sea) and the relationship of the couple (brother/sister; partners) are not the same. However, in terms of Brown and Yule's (1983) criteria, they are quite similar. They appear to have the same number of objects (house, purse, dog in Task 3; boat, bracelet, fish in Task 4), two main characters (a boy and a girl; a man and a woman), and the same number of events (they are doing something; the female loses an important possession; they later find it unexpectedly). The type of characters, in terms of gender, is also the same, although the ages are different (children; adults). Both of them contain changes in the setting, and do not appear to depict anything culturally unfamiliar. Therefore, these two tasks are also expected to be very similar in difficulty according to Brown and Yule (1983), as shown in Table 3.17, below.

Table 3.17
Summary of Similarities between Tasks 3 and 4 according to Brown and Yule (1983)

Criterion	Tasks 3 and 4
No. of objects, characters, events	Same
Setting	Changes in both tasks
Type of characters (gender)	Same (for gender)
Cultural familiarity	Same

Thus, these two pairs of similar tasks were selected based on Brown and Yule's (1983) criteria. The procedures for this pilot study were designed to investigate participants' perceived difficulty and whether these tasks could actually elicit the same numbers of objects, characters and events. This was because even if things are drawn in the pictures, it does not necessarily mean that they will be mentioned in narration, as suggested by Pilot Studies 2 and 4 about the prominence of characters. Also, from the implications of Pilot Studies 2 and 4, it was decided to investigate whether the tasks elicited similar grammatical structures (i.e. similar amounts of subordination).

3.6.4. Participants

This pilot study took place with the Language Testing Research Group at Lancaster University on 10 March 2009. It involved 9 attendees: 3 members of staff (1 English, 1 Hungarian and 1 Belgian) and 6 PhD students (1 English, 3 Japanese, 1 Korean and 1 Iranian) in the Department of Linguistics and English Language.

3.6.5. Procedure

The 9 participants were divided into 3 groups of 3, each with 1 member of staff and 2 PhD students. For one pair of tasks, one PhD student would perform them (with 1-minute planning time for each task) while the other student silently listened and the member of staff took notes. For the other pair of tasks, the roles of narrator and listener were exchanged between the PhD students (the note-taking role of the member of staff remained the same).

The members of staff were provided with worksheets for note-taking which was aimed at keeping a record of the duration of narration, numbers of objects, characters and events, and subordinate clauses for each task. For the number of events, they were asked to write down the verbs that were used in narration so that what was told in the narration could be traced[21]. For the number of subordinate clauses, they noted the clauses with *because* and relative pronouns, both of which seemed quite likely to be elicited by the tasks.

The order of presentation was the same for all three groups: Task 1 before Task 2, and Task 3 before Task 4. After each pair of tasks was performed, the group discussed

[21] The number of verbs was expected to roughly correspond with the number of idea units.

which task they thought was more difficult. The opinions of the groups were later shared with the whole group. The analysis was based on the note-takers' notes and the opinions shared during discussion.

3.6.6. Results and Discussion

3.6.6.1. Tasks 1 and 2

For Research Question (RQ) 1, the number of objects, characters, events and subordinate clauses that were elicited in the two tasks appeared to be more or less the same for each narrator. The notes on the performance on Tasks 1 and 2 taken by the note-takers are summarised in Table 3.18, below.

The narrators of this pair of tasks were Greg, Felice and James (all pseudonyms). All of them mentioned 5 characters in Task 1 (i.e. mother, (two) children, man, ghost), and 4 in Task 2 (i.e. mother, baby, (two) children). Also, the numbers of clauses with "because" and relative pronouns did not differ greatly between the two tasks (0 to 2 clauses for each).

Table 3.18

Summary of Performance Elicited in Each Narrator's Story on Tasks 1 and 2

Narrator	Duration (min:sec)		Verbs (i.e. Events)		Characters		Objects		"because" clauses		Relative clauses	
	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2	T1	T2
Greg	02:38	01:41	do, wash, see, go, take, peg, walk, run, buy, come, surprise (10)	loves, gone, come, hold, bring, put, look, were, surprise (8)	woman, children, man, ghost (5)	Mrs, baby, girl, boy (4)	bowl, powder soap, cat, basket, shirt, balloon, moustache, face, box, washing (10)	chair, fire, book, cot, ball (5)	-	because she's tired (1)	which is (1)	which he decided (1)
Felice	01:45	02:45	passed, offer, finish, hand, play, paint, tie, come back, go out, start, cooking, look, see, wash, surprise (15)	read, sleep, have, fall asleep, sleep, come back, notice, wake up, tell, be quiet, hold, go, wake up, schold, bring up (15)	man, mother, children, ghost (5)	mother, baby, children (sister, brother) (4)	balloon, laundry, washing, washing rope, funny man's face (5)	book, sofa, idea, ball, basket (5)	because she was surprised (1)	because the mother is tired; because they talked loud (2)	man who passed by (1)	-
James	02:30	01:20	washing, watching, went out, wants to, is, hanging, went back to, are walking down, bought, try to, make, draw, was surprised (13)	is sitting, is reading, are sleeping, came, say, is cuddling, wake up, find, was surprised, looks like (10)	lady, man, children, ghost (5)	lady, baby, children (girl, boy) (4)	cat, clothes, washing line, house, balloons (5)	chair, cot, ball (3)	-	-	-	-

Notes. T1, T2 = Task 1, 2.

Numbers in parentheses () in each cell indicate the total number of elements in question.

The number of objects showed some discrepancies between the narrators; while Felice and James mentioned 3 to 5 objects in both tasks, Greg referred to 10 objects in Task 1, twice as many objects as he did for Task 2. The objects he mentioned included *bowl, powder soap, cat, basket, shirt, balloon, moustache, face,* and *box* in Task 1, and he reported being amused by the old-fashioned way of washing shown in Task 1, therefore explaining the first two pictures in detail.

As for the number of verbs, which was aimed at briefly capturing the number of events in this pilot study, all three narrators used approximately the same numbers of verbs in both tasks. Greg used 10 verbs in Task 1 and 8 in Task 2; Felice used 15 verbs for each task; and James used 13 verbs in Task 1 and 10 in Task 2. The reason that Task 1 elicited slightly more verbs might have been due, as raised in the discussion, to the differences in the numbers of pictures in the two tasks. Task 1 contains 8 pictures whereas Task 2 has only 6. This might have led to a fuller narration of Task 1. This suggests that the number of pictures in the tasks should be controlled for in the main study.

The duration of narration differed by 50 to 70 seconds between the two tasks. For Greg and James, Task 1 took longer, which was not surprising considering that it had more pictures. In addition, the note-taker in James's group reported that James hesitated for a long time while he struggled to come up with the word "hanging". However, Felice narrated for longer on Task 2; the note-taker in her group mentioned that she hesitated much more, though she explained the setting in more detail and expressed more personal opinions in Task 2. Also, there were more complex sentences and more connectives than in Task 1. Felice, a Japanese student, suggested that she was more relaxed with Task 2 because she already knew what to do after narrating Task 1, and also recommended that the planning time be longer than 1 minute, as she found it

difficult to make up a story for Task 1 within that short time frame. This suggests that if this pair of tasks was to be used for the main study, then the planning time should be longer than 1 minute and the order of task presentation should be counter-balanced for each participant so that a practice effect will not interfere.

To sum up and answer RQ1, the pair of Tasks 1 and 2 elicited about the same number of characters, events and subordinate clauses from the same narrators. The fact that Greg performed somewhat differently on the two tasks, with a stronger interest towards Task 1, indicates the role of candidate characteristics (Bachman and Palmer, 1996: 64). It might have been Greg's personal characteristics (in terms of experience with an old-fashioned way of washing) and his language knowledge as an English native speaker that affected his performance, enabling more detailed descriptions (with a larger number of objects) for Task 1. It is inevitable that candidate characteristics will come into play, so it is important thing that this does not affect performance enough to lead to different scores. It would not be possible, in practice, to ask all candidates in the main study about their interests and past experiences; however, candidates' teachers might be able to provide relevant information in the main study.

For RQ2, two of the narrators, Greg and Felice, reported that Task 1 was more difficult because it seemed to have a more complex structure. According to Greg, it has two parts where he had to stop and start again, and which converge in the last picture. In addition, Felice said it was more difficult because it was the first task. This was in accordance with the prediction of Brown and Yule (1983) who suggest that Task 1 might be more difficult because it involves a change in setting. However James, a Korean native speaker, said that he felt both of them were equally difficult. Thus, in terms of perceived difficulty, the two tasks were parallel for only one narrator. This discrepancy needs to be investigated further with a larger sample in the main study.

A major issue raised in the discussion was an unclear instruction that was given in this pilot study: "Please narrate a story based on this picture sequence". The note-takers reported that the tense choices in narration differed widely between narrators. One narrator, James, also said that he did not know what to do with the task at first and kept wondering if he should use the past tense or not. An instruction such as "Please tell me *what's happening* in this picture sequence" could elicit the present tense or, if it is said, "Please tell me *what happened*", it could cue the past tense.

In addition, it was suggested that more detailed instruction about what to include in narration would be desirable in order to elicit longer performances which would bear a variety of analyses (e.g. idea units, D value (requires at least 50 words for calculation)). Therefore, it was decided to revise the instructions for the main study following on from these suggestions.

3.6.6.2. Tasks 3 and 4

After the narrators in each group had finished the two tasks, another three narrators took a turn to narrate Tasks 3 and 4: Cheryl, Matt and James. Although the numbers of objects and subordinate clauses appeared comparable between the two tasks (see Table 3.19 below), it became clear that they could not be compared properly for parallelness because the pictures were found to be problematic.

Table 3.19

Summary of Elements Elicited in Each Narrator's Story in Tasks 3 and 4

Narrator	Duration (min:sec)		Verbs (i.e. Events)		Characters		Objects		"because" clauses		Relative clauses	
	T3	T4	T3	T4	T3	T4	T3	T4	T3	T4	T3	T4
Cheryl	01:57	02:32	was, saw, followed, run, wave, started, reached, found, buy, came (9)	fishing, found, dropped, feel, started to cry, caught, cook, surprise, cheered (10)	boy (brother), sister, dog, mother, shop owner	couple (wife, husband) (2)	handbag, road, toy shop, ball, window, purse (7)	sea, bracelet, fish, land, tent, pan, fire, stomach (8)	because he didn't see the dog off (1)	because the bracelet was her favourite (1)	where the purse was (1)	which is a bit cruel (1)
Matt	02:40	02:00	talk, shopping, say, hold, leave, follow, ask, leave, go home, persuade, decide, continue, walking, try, leave, accompany, forget, look, choose, go, remember, figure out, see,	go fishing, manage, catch, take a nap, feel tired, come back, anchor, make fire, cook, prepare, discover, tell, cook, interested (14)	girls, mother, dog (4)	man (1)	bag, shop, window (3)	fishing rod, boat (2)	-	-	dog who brought the bag back (1)	-
Marie	02:00	01:50	go, holding, following, catch up with, found, went to see, forgot, playing, find, picked up, look (11)	go out, fishing, drop, trying, pick up, stop, crying, cooking, cutting, waiting, look, find, camping (13)	children (boys), mother, dog (4)	couple (boyfriend, she) (2)	house, chair, street, toys, bag (5)	watch, fish, boat (3)	(7) *No example noted*	(4) *No example noted*	-	-

Notes. T3, T4 = Task 3, 4.

Numbers in parentheses () in each cell indicate the total number of elements in question.

In Task 3, most of the participants mentioned that it was hard to understand Pictures 3 to 5 (i.e. the children trying to wave the dog away; the purse being on the ground while the girl was still waving; and her starting running without it) in terms of why the purse was put on the ground in the first place and then forgotten. Accordingly, the narrators reported that they felt it difficult to narrate a coherent story in Task 3. Furthermore, Task 4 revealed itself to be even more problematic. One of the narrators, Matt, a Persian native speaker, did not recognise the bracelet in the pictures and so told a story without it. Also, he was not sure whether the other person on the boat (without a fishing rod) was a man or a woman, so he avoided mentioning her. Therefore, Matt's story turned out to be quite different from those of the other two narrators. As can be seen in Table 4, Matt produced a much shorter narration in terms of duration and number of verbs (i.e. events) in Task 4 (40 seconds shorter than Task 3, with 10 less verbs). The problem of the bracelet in Task 4 not being very visible was mentioned by other participants too. In addition, another narrator, Marie, a Japanese native speaker, said that it was hard not only to see the bracelet in the first two pictures, but also to make sense of how the bracelet was returned at the end. She also thought that the 1-minute planning time was too short.

Even with the problematic pictures set aside, Tasks 3 and 4 were not regarded as parallel. In the discussion, it was agreed that they were different in terms of the emotions expressed by the characters, which was not observed with Tasks 1 and 2. In Task 3, both children were surprised and happy about the dog bringing the forgotten purse. However, in Task 4, there was an emotional gap between the woman and the man. The woman was so sad that she started crying, while the man looked somewhat concerned but was probably happy with his catch.

Thus, it was clear that Tasks 3 and 4 were not only problematic but also not as

similar as Tasks 1 and 2 when examined closely. Consequently, it was decided to use Tasks 1 and 2 in the main study.

3.6.7. Conclusions and Suggestions for the Main Study

Although the two pairs of tasks were selected based on their similarities, according to the criteria of Brown and Yule (1983), it was found out that Tasks 1 and 2 were much more suitable for use in the main study. Tasks 1 and 2 elicited approximately the same numbers of characters and subordinate clauses, and clearly appeared worthy of further investigation into the numbers of objects and events though with revised clearer instructions and a larger sample of candidates. Moreover, the differences in perceived difficulty of Tasks 1 and 2 by the narrators were partly related to task familiarity; therefore, a practice task should be given to the participants in the main study and changing the order of presentation made part of the administration procedure. In addition, Matt's struggle with Task 4 cautioned against the danger of underestimating the difficulties or unfamiliarity that the candidates might face in understanding the pictures. Since the tasks by Hill (1960) depict an incident at a Western home, the experts in this pilot study, who have been teaching and studying in the UK for a number of years, might have overlooked what Japanese candidates would struggle with. Therefore, it will be beneficial to collect expert judgements from Japanese teachers in the main study.

3.7. Summary

This series of five pilot studies has raised a number of issues and suggestions for the main study. Pilot Study 1 provided the necessary familiarisation with the analysis of linguistic performance using several variables. Pilot Study 2 involved conducting a group discussion in which expert judgements of relevant task complexity

factors of the two SST tasks were collected. This raised an important point in that the degree of similarity between the pictures and tasks can vary, and differences in the prominence of and relationships between characters might result in quite different linguistic performances.

Pilot Study 3 trialled a larger number of variables, including those of fluency, for analysing linguistic performance. The samples included those by Japanese candidates of SST as well as native speakers of English. It was found that only speech rate correlated highly and discriminated between different levels of proficiency, and that using vocabulary lists such as LFP or JACET8000 was not useful. Other variables, including lexical variety, syntactic complexity, accuracy and idea units, need to be examined further with a larger set of data and obtained under the same task administration conditions.

Pilot Study 4 analysed the NS performance in terms of fluency, syntactic complexity, lexical complexity and idea units, as well as NS perceptions of the two SST narrative tasks. It was again confirmed that the car accident task was perceived as more difficult. The need to distinguish characters and to express the *justifying* and *explaining* functions in the car accident task resulted in the larger number of subordinate clauses per AS-unit and minor idea units. Having found that the NS perceptions help understand the requirements of the tasks, it was decided to also investigate them in the main study.

After collecting some evidence of the characteristics of linguistic performances elicited as well as the narrators' perceptions and expert judgements, Pilot Study 5 identified an ideal, more 'seemingly-parallel' pair of spoken narrative tasks by Hill (1960) for the main study. The main study will use this pair of tasks to address the research questions, which are presented and explained in the beginning of the next

chapter (Methodology).

Chapter 4: Methodology

This chapter explains the methodology used in the main study of this thesis. First the research questions are presented, then follows a description of data collection from the Japanese candidates, Japanese teachers and native speakers of English. Then, the procedures for obtaining ratings data, including rater training, are explained. Finally, the methods of analysis are introduced, corresponding to each of the eight research questions.

4.1. Research Questions

In light of the literature reviewed in Chapter 2 and the results of the pilot studies described in Chapter 3, seven research questions are established, as listed below. The research questions can be broadly divided into four parts, investigating the difficulty of the two spoken narrative tasks by MFRM analysis (RQ1), the perceptions of the candidates, expert judgements by Japanese teachers and perceptions of the native speakers of English (RQ2), the linguistic performances of the candidates (RQ3), and the validity of the linguistic variables (RQ4).

- RQ1. Is the difficulty of the two spoken narrative tasks (by Hill, 1960) the same according to MFRM analysis?

- RQ2-1. Are the candidates' perceptions of the two spoken narrative tasks the same?

- RQ2-2. Are the candidates' perceptions of the two spoken narrative tasks the

same at different levels of proficiency?

- RQ2-3. Do Japanese teachers judge the two spoken narrative tasks to be parallel for the candidates in terms of the relevant task complexity factors?

- RQ2-4. Do English native speakers perceive the two spoken narrative tasks to be equally difficult?

- RQ3-1. Are the performances of the two spoken narrative tasks the same in terms of the linguistic variables?

- RQ3-2. Are the performances of the two spoken narrative tasks the same in terms of the linguistic variables at different levels of proficiency?

- RQ4. How do the variables of fluency, accuracy and complexity correlate with the ratings of spoken narrative performance in the corresponding rating categories?

Firstly, RQ1, with the use of MFRM, provides values for task difficulty of the two spoken narrative tasks, taking into consideration candidate ability as well as the contextual factors involved in speaking assessment. Using CEFR for the rating scales, and the procedures for rater training suggested by the Council of Europe (2009), the ratings of the spoken narrative performances are expected to provide better variables for correlation than the SST Levels in the pilot studies. Details of the process of rater training and major ratings are described in Section 4.5. Secondly, RQ2-1 and RQ2-2

refer to candidates' perceptions of the two tasks elicited using Robinson's questionnaire (2001), and are expected to provide evidence for task paralellness in terms of candidate perceptions as well as a baseline study for the discussion of the relationship between candidate perceptions and proficiency. RQ2-3 and 2-4 aimed to collect expert judgements and NS perceptions of the two tasks devised by Hill (1960). The responses to these two research questions were expected to provide information about the assumed familiarity of the candidates and any issues with the pictures. Thirdly, RQ3 investigates task parallelness in terms of the candidates' linguistic performance, with RQ3-1 looking at the whole candidate population, and RQ3-2 considering the effects of different proficiency levels (based on the ratings given to candidates' performances). Finally, RQ4 addresses the issue of the validity of linguistic variables, and is expected to build on the results of Pilot Study 3. The next chapter will describe the research design of the main study and the methods of analysis for each research question. The processes of trial rating, rater training, and major rating using the CEFR assessment grid are also discussed in detail.

4.2. Data from the Japanese University Students

Data collection from the university students for the main study comprised four components: the candidates' level of English proficiency, language learning background, spoken narrative performance, and perceptions of tasks. This section describes the candidates, procedures and instruments used in this phase of the data collection.

4.2.1. Candidates

The data collection took place at the Tokyo University of Foreign Studies

(TUFS) in Japan for four weeks between 1-26 June 2009. TUFS specialises in foreign languages, it has 26 different language majors. Students admitted to this university are thought to be at the highest level for English reading and writing proficiency in Japan (Daigaku Hensachi Juku, n.d.).

The candidates were recruited via posters and flyers distributed on campus. They were 8 males and 57 females, of which 57 undergraduates and 8 postgraduates. There were 16 differing language majors[22] among the undergraduates. The average age was 20.8 (SD=3.9). All candidates were Japanese native speakers who had learnt English as a first foreign language.

The CEFR levels of candidates' English proficiency assigned by a multiple-choice format Oxford Quick Placement Test (explained in the next section) were B1 (n = 15), B2 (n = 31) and C1 (n = 19). This is an objective indication of candidates' English levels, for which TOEIC scores were used in the pilot studies. The average age for starting to learn English was between 10 and 11, with the C1 candidates starting slightly earlier than the others. The candidates had learned English for about 10 years, again with C1 candidates being the longest. In addition, the length of stay in English-speaking counties (if any) and TOEIC scores (if available) showed a similar increasing pattern from B1 to C1, although only a partial population of candidates provided this information. The basic information about their English learning backgrounds is summarised in Table 4.1, below.

[22] English, German, Italian, Spanish, Russian, Czech, Chinese, Korean, Mongolian, Filipino, Thai, Vietnamese, Cambodian, Burmese, Urdu and Persian.

Table 4.1

Summary of the Candidates' English Learning Backgrounds

Information	B1		B2		C1	
	M	SD	M	SD	M	SD
Age starting to learn English[1]	11.00	2.83	10.61	2.97	10.21	2.72
Years of learning English[2]	9.75	6.74	9.69	3.37	10.64	3.37
Years of stay in English-speaking countries[3]	.53	.25	2.27	2.05	2.44	2.55
TOEIC Score[4]	661.92	124.29	781.18	79.95	901.56	71.20

Notes. [1]In Japan, the starting age for junior high school (equivalent of middle school), where English becomes a compulsory subject, is 12-13.
[2]Up to the time of data collection for the main study.
[3]The percentages of candidates who had stayed in an English-speaking country were 20% (B1), 22.6% (B2) and 36.7% (C1).
[4]77% of the candidates took the TOEIC after they started university.

4.2.2. Instruments

The data include the responses of the candidates via the following four instruments:

1) Oxford Quick Placement Test (QPT)

2) Spoken narrative tasks

3) Robinson's (2001) task-difficulty questionnaire for each task

4) A questionnaire on their language learning background

4.2.2.1. Oxford Quick Placement Test

The Oxford Quick Placement Test (QPT) is claimed to be able to estimate candidates' levels of English proficiency in a short time and with resulting scores claimed to link to the Association of Language Testers in Europe's (ALTE) framework and the Common European Framework for References (CEFR), according to the QPT manual from UCLES (2001: 5). The QPT claims to assess reading, vocabulary and

grammar (ibid.: 3). This study used its pen-and-paper version since it was impossible to reserve a large computer room that the computer-based version requires.

The pen-and-paper version is in multiple-choice format and consists of two parts, with 40 items in Part 1 and 20 more items in Part 2. Part 2 includes items which are "incrementally more difficult than those in Part 1" (UCLES, 2001: 35), so it is recommended that it be given if candidates are considered mostly to be ALTE Level 3 or above (i.e. CEFR levels B2 or above). According to levels estimated for the students by teachers at TUFS who are familiar with CEFR, it was decided to give both parts of the QPT. The scores represent the numbers of test items answered correctly out of 60, with ALTE and CEFR levels being assigned accordingly.

4.2.2.2. Spoken Narrative Tasks

Tasks 1 and 2, which were trialled in Pilot Study 5, were modified and used in the main study. Two pictures were removed from Task 1, and the picture numbers were modified accordingly, as well as being made more visible. The revised picture sequences in the tasks are shown in Appendix 4 (Tasks 1 and 2 will henceforth be called Tasks A and B). The instructions were revised, as explained in the Procedures section below, and the order of presentation was reversed for each candidate so as to minimise any order effect. In addition, a practice task, which was also selected from Hill (1960), was given before the main tasks so that the candidates would better understand what they were required to do and would not be too nervous about the main tasks. The practice task depicts several characters and household objects, and has an understandable ending in the same way that the two main tasks do.

4.2.2.3. Robinson's Task Difficulty Questionnaire

After completing each of the two main tasks, the candidates were asked to complete a task difficulty questionnaire (Robinson, 2001) which enquired into the perceived difficulty, anxiety, self-rating of performance, interest, and enjoyment. Using a 9-point Likert scale, the questionnaire was intentionally brief so that the candidates could move on to the next task quickly after undertaking the previous task and with minimal distraction (Robinson, 2001: 41):

1. How difficult was this task?
2. How nervous were you when doing this task?
3. How well do you think you did this task?
4. How interesting did you think this task was?
5. Would you like to do more tasks like this?

Robinson explains how "responses to these five items assessing overall perception of task difficulty, ratings of stress, perceived ability to complete the task, interest in task content, and motivation to complete these and other tasks like them, were used to assess learner perception of task difficulty" (Robinson, 2001: 41). The questionnaire used in this study is shown in Appendix 5.

4.2.2.4. Language Learning Background Questionnaire

This questionnaire was designed to collect information on some of the characteristics of the candidates (e.g. age, gender) and their language learning background, especially their English learning, such as the age at which they started learning English, the reasons they studied English, and whether or not they had been to English-speaking countries. The descriptive statistics are provided above in Table 4.1.

4.2.3. Procedure

All the instruments except the QPT were administered via a one-to-one interview with the author. The QPT was given to a number of candidates to complete at the same time. Lunchtime was chosen for administering the QPT so that it would be convenient for the students to come to the university without having to worry about getting up early or being late for their out-of-hours activities, such as clubs and work. A classroom with 30 desks and chairs in a lecture building had been booked during lunchtimes for the four weeks of data collection, and the students who decided to take part were able to contact the author with the date on which they wished to take the QPT and have their individual interview.

The one-to-one interviews lasted approximately 30 minutes and were conducted on campus in a quiet room with a desk and chairs where the candidate and the author (i.e. interviewer) could talk alone. Japanese language was used during the interviews except during the candidate's narration of the tasks in order to make sure that all the candidates understood the instructions and questionnaires.

Each candidate was first asked to sign a consent form confirming that they agreed to be recorded and for their data to be stored and used for research. Then, he or she was shown a practice task and given instructions to narrate a story with as much detail as possible. Then, two minutes would be allowed for studying the task and planning what to say. The instructions given translate as below. "*Please narrate a story for this picture sequence explaining who did what and how it turned out. There is no time limit for your narration. Ideally, I would like you to talk for two minutes or longer, though two minutes may seem rather long. So, you can decide on the details freely, such as the situation, the relationships between the characters, and the reasons why certain characters did what they did. Please try to talk as much as possible.*"

It was decided to encourage participants to narrate for two minutes or longer based on the findings from Pilot Study 5. Similarly, two minutes' preparation time was allowed because, in Pilot Study 5, Japanese candidates said that one minute was insufficient. Also, one student who took part in the early stages of data collection at TUFS advised that two minutes' preparation was appropriate as it made it easier to get a better sense of the time allowed for the task if the time for preparation and (ideal) narration were the same. During the interviews, a clock was placed in front of the candidates so that they could check the time that had elapsed if they wanted to.

After the practice task was completed, the interview moved on to one of the main tasks but with shorter instructions: "*As you just have done with the previous sequence, please narrate a story for this picture sequence including as much detail as possible*". Another two minutes for planning were allowed. The candidates who narrated for much less than two minutes in the practice task were asked to include more details in the next two tasks.

The questionnaire on their language learning background was given after completing the second task, before the interview ended. The candidate was compensated with a payment of ¥1,000 (approximately £7), for his or her cooperation in this study, at the end of the interview.

4.3. Data from the Japanese Teachers of English

In order to obtain judgements from experts who were more familiar with English-teaching contexts in Japan, the two Japanese teachers who had taught the majority of the candidates at TUFS were recruited. Both teachers hold a postgraduate degree in TEFL and Applied Linguistics, and had over 5 years of teaching experience in

senior high schools and universities in Japan. They responded to a modified version of Weir and Wu's (2006) Checklist of Difficulty where they were asked to answer yes or no to statements about the assumed familiarity of candidates with the characters, locations, events, objects, the lexical and grammatical knowledge to express them, functions of describing and narrating, and the sufficiency of details and visibility of the pictures in the tasks. Weir and Wu's (2006) checklist originally included 10 statements, but the last one was removed for this thesis as it was about the assumed sufficiency of the time provided to complete the task. This is because, as there was no time limit for the candidates to complete their narration, it was redundant. The remaining 9 statements are listed below:

1. The roles of people in the pictures are equally familiar to candidates.
2. The locations in the pictures are equally familiar to candidates.
3. The events in the pictures are equally familiar to candidates.
4. The objects in the pictures are equally familiar to candidates.
5. The objects and events to be described by candidates are equally visible in the pictures.
6. There are enough details in the pictures for the candidates to complete the task.
7. The lexical items required to describe the pictures are equally familiar to candidates.
8. The grammatical structures required to describe the pictures are equally familiar to candidates.
9. The language functions required to describe the pictures are equally familiar to candidates.

To establish parallelness, it was desirable that the two teachers agreed with (i.e. answered yes to) all statements. However, when they disagreed, the reasons were pursued further.

4.4. Baseline Data from the English Native Speakers

The two main tasks were also performed by a small number of English native

speakers in order to obtain baseline data. Eleven British undergraduate students were recruited, each of whom was majoring in modern languages[23] at Lancaster University, so their profiles were similar to the Japanese candidates in terms of their age (M = 19.7, SD = 1.83), major (foreign languages), and education level (university).

The same procedure for eliciting narrative performance, including the practice task, was followed, as described in the previous section, except that the native speakers were allowed to take as much or as little planning time as they wanted. This was because it was found in the early stages of the interviews that the native speakers felt uncomfortable with two minutes' planning time which some felt was unnecessary. Each interview took no more than 15 minutes and £3 was paid for their cooperation. After they had completed the two tasks, they were asked if they thought the two tasks were equally difficult. If they answered no, they were asked to provide reasons or explanations. This was to examine whether the two tasks were parallel in terms of the perceptions of the NS, who, unlike the Japanese candidates, were free from the constraints of linguistic resources.

4.5. Ratings Data for the Spoken Narrative Performances

This section describes how the rater training was conducted and how the ratings data for the candidates' narrative performance were obtained. Rater training serves a vital role in ensuring the reliability of the ratings given to performances. It is usually the case that meetings are held where raters are first introduced to the rating criteria and illustrative samples of performance (Luoma, 2004: 177). Then, they are asked to rate several performances independently at different levels, then invited to a

[23] These students were learning one or two foreign languages amongst: French, German, Italian and Spanish.

discussion to reveal their ratings in order to eliminate discrepancies between them (McNamara, 2000: 44). Such meetings are intended to moderate understanding of the rating criteria and level descriptors so that different raters are very likely to give the same ratings to a performance.

In order to ensure the quality of rating, it is essential to use rating scales that are appropriate for the intended purpose. Since the spoken narrative tasks used in the main study did not come from a test provided with existing rating scales tailored for this task type, it was decided to adopt the Common European Framework of References for Languages (CEFR) for the main study.

The rater training followed the procedures referred to by the Council of Europe (2009) as 'standardisation', which involves two phases: training in rating performances in relation to CEFR levels (using the illustrative samples), and benchmarking the narrative performance samples (i.e. my main data samples) to CEFR levels. Both phases were administered during a 3-hour rater training session with 70-90 minutes for training and 90-110 minutes for benchmarking. Since not all 9 raters who participated in the main study could attend the session on the day originally planned, two more sessions with the same content were arranged.

The following sections explain how the raters were selected and the procedures for training and benchmarking, and report on issues raised during these processes in detail.

4.5.1. Raters

Nine raters participated in both training and benchmarking. Due to time constraints, the CEFR familiarisation phase, which the Council of Europe (2009)

recommends be arranged before standardisation, was not conducted. As all raters were members of the Language Testing Research Group and/or the Second Language Learning Research Group at Lancaster University, where the CEFR is often cited and discussed, it was hoped that familiarisation would occur as standardisation proceeded. The 9 raters included 5 native speakers of English, 2 Japanese L1 speakers, 1 Arabic L1 speaker, and 1 Polish L1 speaker. All raters, except for one English native speaker, had at least 2 years experience of teaching English as a foreign language, and 5 raters had previously been trained as examiners or raters for some tests of spoken English.

4.5.2. Training with the CEFR Illustrative Samples

4.5.2.1. Selection of Samples

Illustrative samples of the CEFR levels were derived from a DVD published by the Centre International d'Études Pédagogiques (CIEP) in 2008. The DVD includes audio-visual recordings of spoken performances by French teenagers on monologic and interactive tasks at 6 CEFR levels.

The samples at C2 level were not used in the training because there appeared to be no C2 level candidates in my main data. Also, the visual recordings were not shown to the raters because my main data consisted only of audio recordings. Finally, although the Council of Europe (2009: 43) strongly advises showing samples for both task types during rater training, it was decided to use only samples of monologic tasks, as my main tasks were narrative, i.e. monologic. Considering the fact that the levels of the illustrative samples were decided based on performances on both task types, this decision risked training raters with samples that might be unrepresentative of the CEFR levels. Then, raters would not be able to see the whole picture of why a person was at a certain level, even though they might have demonstrated stronger performance on a

monologic task but appeared weaker on an interaction task (or vice versa).

In order to minimise the risk of using unrepresentative samples, two measures were taken. Firstly, the illustrative samples to be used in rater training were chosen after consulting ffrench (n.d.) who lists comments to explain why each of the illustrative samples is assigned to a particular level. Only those samples that appeared to have more comments on examples for monologic tasks were selected, so that the assigned level could be more or less accurately estimated from the monologic performance only. The selected illustrative samples were as follows (in order of presentation in the sessions): Amélie (B1); Tiennot (B2); Camille (A2); Clara (A1); Charlotte (C1).

Using the selected illustrative samples, a pilot rater training session was then held in a Language Testing Research Group meeting at Lancaster University on 20 October 2009. This was to observe whether introducing only the monologic performances from these samples could lead to a reasonable estimation of the assigned levels. Twelve members participated in the pilot rater training session, 8 PhD students and 4 members of staff from the Department of Linguistics and English Language. The session followed the same procedures with the same rating scales as explained in the following sections. It was found that, overall, the members were able to estimate the levels with reasonable precision from the monologic performances.

However, one issue did arise; they found it difficult to distinguish between A1 and A2 levels. Some members mentioned that Camille, who was at A2, demonstrated much weaker performance in terms of accuracy than the A1 sample. Also, Camille's fluency broke down at the end of the monologic task, which also appeared to have led to a wrong estimation of her level being A1. What is more, some members noted that the A1 sample, Clara, seemed to perform much better than their expectations for A1 level. It might be because Camille and Clara both demonstrated monologic performances that

were more or less at the same level, but showed their strengths and weaknesses in the interaction task.

It would have been better to replace Clara (A1) with the other A1 sample on the DVD, Tifaine, who appeared to be weaker than Clara and could have been more representative of A1 for the monologic task. Yet, it was decided to keep Clara's sample because ffrench (n.d.) fails to explain Tifaine's performance due to a misprint (at the time of writing this manuscript). Having no comments or explanations suggests a severe lack of accountability, which might be even more problematic in the main rater training discussions. Thus, it was decided to use all the samples piloted in the Language Testing Research Group for the main rater training.

4.5.2.2. Rating Scales

Three rating scales were adopted from the Council of Europe (2009) though with some changes so as to tailor them to my narrative samples: the global oral assessment grid, the oral assessment criteria grid, and the 'plus level' descriptors. The global oral assessment grid (i.e. Table C1 in: Council of Europe, 2009: 184) was used, except for the last row in the table which states, "Use this scale in the first 2-3 minutes of a speaking sample to decide approximately what level you think the speaker is". This was not relevant because the selected illustrative samples and my samples would only last for about 2 minutes.

The oral assessment criteria grid (i.e. CEFR Table 3; Table C2 in Council of Europe, 2009: 185) was modified as my data did not contain interaction tasks. Thus, the column for level descriptors in 'interaction' was deleted. Instead, a column for 'sustained monologue' was inserted as this suited the rating of performances better for the narrative tasks used in the main study. The level descriptors in 'sustained

monologue' about description and narration were inserted from the subscale tables in the CEFR: 'overall oral production' and 'sustained monologue: describing experience' in Council of Europe (2001: 58-59). Revising the grid is supported by the Council of Europe (2009: 53) in order to match the performances in the samples.

Similarly, the column for 'interaction' was replaced with 'sustained monologue' for the 'plus level' descriptors which were adopted from the Council of Europe (2009: 186 (Table C3)). The Council of Europe (2001: 59) only list A2+ descriptors in 'sustained monologue: describing experience', so these were inserted to A2+ and the other cells were left empty in the revised 'plus level' grid. The three rating scales used in the rater training are shown in Appendix 6 (Tables 1-3).

4.5.2.3. Procedures

The raters were first asked to read the level descriptors in the rating scales carefully and to pay close attention to what was different between the neighbouring levels. Next, they were provided with the rating sheets (see Appendix 7), listened to a sample once, and then noted the overall CEFR level of their initial impression. The sample would be repeated again, followed by detailed rating of the different aspects of performance (i.e. range, accuracy, fluency, coherence and sustained monologue). Finally, an overall considered judgement was made and noted. After the raters presented their ratings to the whole group, the standardised CEFR level of that sample was revealed. A discussion ensued, moderating the judgements towards agreement, whilst sharing any problems or difficulties they experienced with rating. The same procedures were applied to each of the 5 illustrative samples.

4.5.2.4. Results and Issues

Since the illustrative samples were only partially presented for the reasons explained above, it is important to note how the raters judged the performances on the monologic tasks and what difficulties they faced in the main sessions too. The issues raised here included some problems which did not emerge in the pilot session.

The difficulty of distinguishing between A2 and A1 samples was again raised. Also, 3 out of 10 raters disagreed strongly with Tiennot (B2) being at B2 level because they did not think his fluency was better than that of Amélie (B1); they thought he had more unnatural flow with unnatural phrases and that his errors were more noticeable. In order to persuade them, the comments by ffrench (n.d.) were used to emphasise that Tiennot was able to sustain himself "at a fairly even tempo", unlike Amélie who had a long silent pause in the middle of her monologue, and that naturalness of flow was not part of the rating scales for fluency. Still, this might have been caused by the monologic samples being unrepresentative. For accuracy, it was highly probable that this was the case. The difference between B1 and B2 levels is whether the errors cause misunderstanding or not. Amélie made an error which confused the listener in the monologic task (Ffrench, n.d.: 11), but perhaps it did not differentiate her accuracy rating from that of Tiennot.

The raters managed to arrive at the assigned levels in the end, with "suitable agreement (maximum spread equal to one and a half levels) (Council of Europe, 2009: 52)" in each cell. The detailed analyses of monologic performances of the CEFR illustrative samples are presented below in Table 4.2.

Table 4.2

Detailed Analyses of the Monologic Performances of the French Pupils

Name	Overall (Assigned)	Range	Accuracy	Fluency	Coherence	Sustained monologue
Amélie	B1	B1+/B2	B1	B1	B1	A2/B1
Tiennot	B2	B1+/B2	B1/B2	B1/B2	B2	B1+
Camille	A2	A1/A2	A1/A2	A1/A2	A1/A2	A2
Clara	A1/A2	A1/A2	A1/A2	A1/A2	A1/A2	A1/A2
Charlotte	C1	B2/C1	B2+/C1	B2/C1	C1	B2/C1

The next section describes the next phase of standardisation: benchmarking with local samples, i.e. my main data for narrative performances. It turned out that the raters felt it was easier to rate them because the tasks were exactly the same for all the samples; however, other problems emerged in this phase. These will be explained in Section 4.5.3.4.

4.5.3. Benchmarking with the Japanese Samples

4.5.3.1. Selection of Samples

A total of 7 samples were selected. Five samples came from Task A, which were thought to cover from A1 to B2/C1 levels according to the author's impression. The 2 samples from Task B were expected to represent B2 and A1 level performances with seemingly maximum or minimum idea units expressed in the narrated story (i.e. a very detailed/short narration).

4.5.3.2. Rating Scales

The same rating scales as in the training phase were used.

4.5.3.3. Procedures

The same procedures as in the training phase were repeated. What was different from the training phase was that there was no predetermined standardised rating for the samples to be taught, so the level of a sample had to be agreed upon and decided among the raters in the session. Therefore, the discussion took much longer in this benchmarking phase. When the sessions ran out of time, the remaining samples that needed benchmarking were discussed via email. The results and issues are introduced in the next section.

4.5.3.4. Results and Issues

The detailed analyses of the 7 narrative benchmarking samples are summarised in Table 4.3, below.

Table 4.3
Detailed Analyses of the Narrative Performances of the Japanese University Students

Task	ID No.	Overall	Range	Accuracy	Fluency	Coherence	Sustained monologue
A	41	A2/B1	B1	A2/A2+	A2/B1	A2/B1	A2+/B1
	22	B2/B2+	B1+/B2+	B1+/B2	B2/C1	B2/B2+	B1+/B2+
	71	A1+	A1/A2	A1/A2	A1/A2	A1	A1
	40	B1/B2	B1/B2	B1+/B2	B1+/B2	A2+/B1+	B1+/B2+
	50	B1+/B2	B1/B1+	A2+/B1+	B1+/B2	A2+/B1+	B1/B2
B	02	B1+/B2+	B2	B2/B2+	B1/B2	B1+/B2	B1+/B2
	63	A1/A1+	A1/A2	A1	A1	A1/A2	A1/A2

The first thing to note is that all the raters agreed to have an A1+ level which the initial rating scales did not unclude. When rating Candidate 71, they felt that she was not a 'straight' A1, but not quite 'good enough' to be rated as A2. As it was expected that there would be more performances like that of Candidate 71, it was immediately decided to create an A1+ level. Dividing a level in this way is recommended by the Council of Europe (2001: 31-32) if it is necessary in order to meet the demands of the

samples at hand.

There are only a few cells that have only one agreed level, although the raters did manage to agree within the acceptable variability limit of one and half levels. The sources of this variability appeared to be the following three issues: restrictions imposed by the tasks, vagueness of the descriptors, and raters' tendencies toward severity/leniency.

Firstly, the raters found it difficult to apply some of the descriptors to the narrative performances that they were listening to. Many of the CEFR descriptors illustrate performances that are task-dependent and/or topic-dependent, which are not usually elicited by spoken narrative tasks. For example, in the oral assessment criteria grid, the B1 descriptor in Range says, "Has enough range to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events" (my emphasis). Also, C1 in Sustained monologue is described as "Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion" (my emphasis). The narrative tasks used in the main study do not require such performance, resulting in the differences in the stances that the raters took towards non-matching descriptors. Some resorted to estimating a sample's level by comparing it to the other samples which were rated based on more relevant descriptors. Others did not attempt to supplement or imagine what was not written in the descriptors; thus, different samples received the same ratings because the relevant features were present in a limited number of cells in the rating scales.

Secondly, some descriptors were found to be problematic because they hardly differed between levels. For example, the descriptors for coherence are presented below, with my emphasis:

A1: Can link words or groups of words with very basic linear connectors like "and" or "then".

A2: Can link groups of words with simple connectors like "and", "but" and "because".

In the case of my spoken narrative tasks, it is less likely that connectors such as *but* and *because* are used unless the candidates chose to (and were able to) incorporate more minor ideas such as reasons and background events in the story. Therefore, it becomes unclear what is different between the descriptors for A1 and A2. In addition, the difference between "basic connectors" and "simple connectors" is hard to grasp. These issues, relating to how the descriptors are phrased again resulted in different stances from the raters, as described above.

Thirdly, the tendency of the raters towards severity/leniency appeared to have influenced the variability in the ratings. Japanese raters tended to be lenient because they were more sympathetic towards Japanese learner performances, while some of the experienced raters were inclined to adhere to their own impressions of the CEFR levels, or to project standards from different rating scales that they had previously been trained on.

4.5.4. Major Rating

Once the rater training workshops were finished, the raters were given a CD-ROM with all 65 candidates' narrative performances of the two tasks so that they could undertake ratings in their free time using the other rating materials supplied during the workshop (i.e. the rating scales, rating sheets and tasks). Timekeeping sheets were also provided so that the raters could be paid according to the hours they worked

on rating performance. The raters were paid £11 per hour. The deadline to submit the ratings was 28 February 2010, which meant that they had about 3 months to complete their work.

4.6. Methods of Data Analysis

4.6.1. Research Design

The research design of the main study can be summarised as shown below in Figure 4.1.

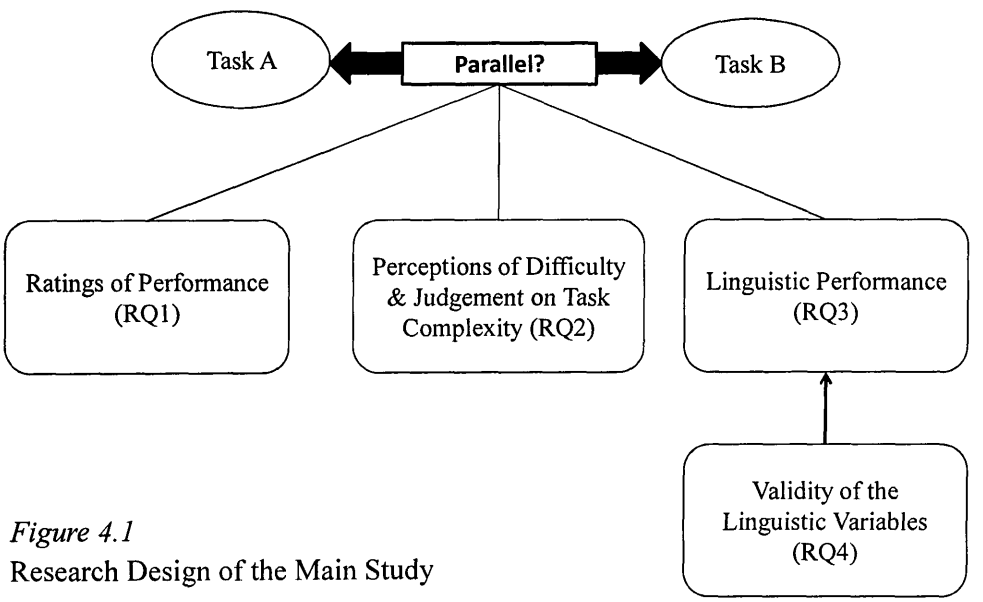


Figure 4.1
Research Design of the Main Study

The parallelness of the two tasks was examined in terms of the ratings (i.e. score parallelness), candidate perceptions (i.e. evidence of substantive validity), expert judgements and NS perceptions of task complexity (i.e. further evidence of a priori analysis of task characteristics), and linguistic performances (i.e. validity evidence of generalisability). The validity of the linguistic variables fed into the interpretations of

the results of the linguistic performances.

More specifically, for RQ1, the ratings that the raters submitted (before the deadline for the major rating, as noted in Section 4.5.4) were used and analysed via the MFRM method to yield values of task difficulty for the two spoken narrative tasks. The ratings data that were adjusted by MFRM were also used for RQ4 as correlates of the linguistic variables. For RQ2, Robinson's Task Difficulty Questionnaire (2001) was administered to the Japanese candidates in order to obtain their perceptions. In addition, a modified version of Weir and Wu's Checklist of Difficulty (2006) was given to the Japanese teachers for their expert judgement of relevant task complexity. Moreover, a brief subsequent interview with the 11 NS provided further information about the complexity of the two tasks. RQ3 was aimed at investigating the characteristics of the linguistic performances elicited by the two tasks using the linguistic variables. The analysis for RQ4 was expected to provide crucial information about the nature of the linguistic variables and to give insights into the interpretations of the results of analysis for RQ3.

4.6.2. MFRM Analysis of Task Difficulty, Candidate Ability and Fair Average Ratings (RQs1, 3 & 4)

Answering RQ1 required obtaining a single measure of task difficulty for each of the two tasks under investigation, which would take into account candidate ability and rater variability. Research suggests that it is inevitable that there will be different ratings for the same performance if multiple raters are involved, due to varying degrees of rater severity even after rater training (Lumley and McNamara, 1995). To compensate for rater variability, using a multi-faceted Rasch measurement (MFRM) technique is recommended (Fulcher, 2003: 142; Council of Europe, 2009: 99); this aims to take into

account rater variability when calculating candidates' ability levels. Therefore, it was decided to use FACETS in the main study, this is an MFRM program that has been widely utilised in the field of language testing research.

MFRM is an extension of the Rasch model derived from Item Response Theory. Before Linacre (1994), earlier Rasch models only considered the candidate and test items (i.e. two 'facets') when calculating the probability of a candidate getting an item correct. Linacre (1994) extended the model in order to cater better for the needs of performance testing, e.g. testing speaking or writing that involves raters. The underlying theory of MFRM is that the probability of a candidate receiving a rating is equal to the candidate's ability plus other facets, such as task or topic difficulty, and the severity/leniency of the raters (McNamara, 1996: 118-121). By employing MFRM, it is possible to estimate candidates' ability and compare task difficulty more accurately without having to depend on which candidate happened to be rated by which rater (Linacre, 1994: 1). If the difficulty of two tasks calculated by MFRM analysis yields no difference, then the two tasks can be assumed to be parallel in terms of the ratings of linguistic performance by the candidates (for RQ1). Moreover, the estimation of candidates' ability was used to allocate a CEFR level to each candidate for RQ3 so that the linguistic performances could be examined according to their different levels of (speaking) proficiency.

The output from FACETS can also be used to estimate each candidate's agreed ratings in each of the rating categories on the rating scale. When FACETS calculates candidates' ability measure, taking rater severity into account, it also outputs an adjusted measure called 'fair average'. This measure is calculated assuming that the rater severity has a zero measure (i.e. there is no difference in rater severity between the raters), thus adjusting raw ratings for severe and lenient raters (Linacre, 2009: 223).

Utilising the candidates' fair average ability measure, the varied raw ratings could be summarised reliably into one single rating for each of the rating categories. This was extremely useful for answering RQ4, where an agreed rating of each candidate's performance in each rating category was indispensable for calculating correlations with the values of the linguistic variables. Once the fair average measures for all the rating categories for every candidate had been calculated, they were labelled with the corresponding CEFR levels; the range of fair average measures that each CEFR level covered was also indicated by FACETS. In this way, a single rating in each of the rating categories for each candidate's performance on the two tasks was obtained.

4.6.3. Perceptions by Candidates and NS and Expert Judgements of the Tasks (RQ2)

For RQ2-1 and 2-2, the candidates' responses to the 9-point Likert scale questionnaire by Robinson (2001) were tested for mean differences using paired-sample *t*-tests and PASW 17.0. Regarding RQ2-3 and 2-4, because the sample sizes were very small and the data involved the responses from brief interviews, the responses were only qualitatively summarised and discussed.

4.6.4. Linguistic Performances on the Tasks (RQ3)

The linguistic variables used for RQ3 are summarised in Table 4.4, below.

Table 4.4

Variables of Fluency, Accuracy, Complexity, Coherence and Idea Units

Aspect	Variables
Fluency	Speech rate
Accuracy	No. of errors per AS-unit No. of errors per 100 words % of error-free clauses
Lexical Complexity (Lexical Variety)	D-value (calculated by CHILDES program)
Syntactic Complexity	AS-unit length No. of subordinate clauses per AS-unit
Coherence	Ratio of coordination (calculated by Coh-Metrix)
Idea Units	No. of main idea units No. of minor idea units

For fluency, speech rate was used because of what was learnt from the findings of Pilot Study 3, in which the mean length of runs did not discriminate well between different levels of proficiency. Hesitation markers, such as false starts and repetitions, were not investigated as those were not proven to be valid according to Kormos and Dénes (2004). For accuracy, general variables of accuracy were employed: percentage of error-free clauses, errors per 100 words, and errors per AS-unit. A specific variable of accuracy, the percentage of correct use of the past tense was trialled in Pilot Studies 1 and 3; however, it was decided not to use this variable. This was because the aim of the main study was not to examine candidates' mastery of the past tense, and verbs in the present tense were not regarded as errors if the candidate started telling the story in the present tense. For lexical complexity, D value was investigated again. Pilot Study 3 demonstrated the use of vocabulary lists, such as LFP or JACET8000, to be unreliable, since percentages could easily change because of a few words which belong to different lists. Although D value did not correlate highly with or discriminate between different levels of proficiency (i.e. SST Levels), this variable was chosen for the main study, given that the ratings would be awarded specifically to spoken narrative performances (unlike overall SST Levels), and therefore the results might differ from those in Pilot

Study 3. Moreover, as the two tasks for the main study were expected to elicit similar lexical items involving a home with a mother and children, examining whether lexical variety would yield no difference was considered worthy. The variables of syntactic complexity included AS-unit length and the number of subordinate clauses per AS-unit in order to build on the findings of Pilot Study 4, which discussed NS performance and task complexity.

As another variable which closely relates to syntactic complexity, as well as cohesion and coherence, the ratio of coordination was used. As noted in Section 2.7.3.2, Bardovi-Harlig (1992) recommends the use of this variable over variables of subordination, especially for candidates at lower levels of proficiency where complex language use may be rare. The formula suggested by Bardovi-Harlig (1992) for this ratio uses the numbers of independent clauses which are connected by coordination, clauses and sentences. The number of sentences is subtracted from the number of clauses, and the value obtained is used to divide the number of independent clauses which are connected by coordination, which then yields the ratio of coordination in decimal form. Although this formula appears convincing, identifying the independent clauses which are connected by coordination seemed overwhelming, on top of the manual identifications of errors, AS-units and subordinate clauses for other linguistic variables in the main study. Therefore, it was decided to utilise an automated Web-based text-analysis tool called Coh-Metrix (McNamara, Louwerse, Cai and Graesser, 2005). It produces a number of indices that represent the linguistic and discourse cohesion of a text in a chosen genre and, among them, there is an index called "positive additive connectives". This index examines the incidence of connectives including (but not limited to) coordination devices such as *and, also, next* and *but.* The 'narrative' genre was chosen for this study.

To examine how detailed the narration was, the idea units were identified by using the baseline data from the 11 English native speakers (as explained in Section 4.3), following the recommendation of Ellis and Barkhuizen (2005). It was decided that an idea unit would be a 'main' idea unit if it was told by all 11 native speakers who participated in the baseline data collection, as explained above, leaving the other idea units as 'minor' ones. The actual idea units for both tasks are reported in the next chapter (Results).

All the spoken narrative performances were transcribed. One set of transcripts was segmented into AS-units manually. Another set of transcripts was created for analysing fluency and lexical variety; these were not segmented into AS-units, however all the unfinished words, fillers and Japanese words that would not be recognised as English words were removed, and therefore not counted for calculating speech rate or D values. For accuracy and syntactic complexity variables, all the errors, AS-units, clauses and subordinate clauses were manually identified by the author. Intra-coder reliability of coding AS-units, clauses and subordinate clauses was calculated by simple agreement on 10 transcripts for each task (out of 65 transcripts each). The total of 20 transcripts were chosen based on randomly generated ID numbers using Microsoft Excel, and the agreement reached 90%. To ensure accuracy when counting errors, the errors in 10 transcripts from the Japanese candidates were identified by an English native speaker, who had experience as a TEFL teacher, and by the author. The inter-coder reliability was .945 and was thus considered satisfactory. Subsequently, errors were identified by the author alone.

Methodologically, all the variables were then analysed for mean differences using PASW 17.0. As it turned out that there were an insufficient number of candidates at B2/B2+ and NS levels (assigned based on MFRM analysis) for conducting

parametric tests, Wilcoxon's signed-rank tests were used at these levels. Paired-sample *t*-tests were used for A2/A2+ and B1/B1+ levels in order to examine whether the linguistic performances yielded any differences between the two tasks.

4.6.5. Validity of Linguistic Variables (RQ4)

RQ4 was devised to examine whether the linguistic variables actually reflect human recognition of how fluent (or accurate, etc.) narration is which, in this study, was operationalised as correlation with the fair average ratings (calculated by MFRM analysis). The rating scales that were used by the raters are shown in Appendix 6 (see Table 2 for the 5 rating categories), and the correspondence between variables and rating categories is summarized below in Table 4.5.

Table 4.5
Corresponding Variables and Rating Categories

Aspect	Variables	Rating Category
Fluency	Speech rate	Fluency
Accuracy	No. of errors per AS-unit No. of errors per 100 words % of error-free clauses	Accuracy
Lexical Complexity	D-value (Lexical Variety)	Range
Syntactic Complexity	No. of words per AS-unit No. of subordinate clauses per AS-unit	Range
Coherence	Positive additive connectives (by Coh-Metrix)	Coherence
Idea Units	No. of main idea units No. of minor idea units	Sustained Monologue

In the CEFR assessment grid (in Appendix 6, Table 2), the descriptors in Fluency mention pauses, hesitation markers, such as false starts and reformulation (A2), as well as the smooth flow of language (C1) and even tempo (B2). According to the findings by Kormos and Dénes (2004) and Pilot Study 3, it was decided to investigate only speech

rate, a temporal variable of fluency. Although this variable does not directly represent hesitation markers, it was hoped that this rating scale would function well in the main study, based on the fact that the raters were able to reach agreement during rater training.

The majority of the descriptors in Accuracy seem to refer to the frequency or degree of accuracy, for example: "Shows a relatively high degree of grammatical control." (B2), "Uses relatively accurately a repertoire of frequently used "routines" … " (B1), and "Uses some simple structures correctly" (A2). The degree of seriousness (i.e. gravity) of errors is only mentioned at B2 level, where a candidate "does not make errors which cause misunderstanding… ". Thus, the frequency-based variables of accuracy (i.e. the percentage of error-free clauses, errors per 100 words and errors per AS-unit) were expected to be in line with the ratings in this category.

The Range category includes descriptors about the range and complexity of performance, for example: "has a sufficient range of language to be able to give clear descriptions… " (B2), " …using some complex sentence forms… " (B2), "uses basic sentence patterns… " (A2), and "has enough language to get by, with sufficient vocabulary to express him/herself… " (B1). For syntactic complexity, the number of subordinate clauses per AS-unit would correspond to the descriptors for the use of complex forms. AS-unit length was also expected to correspond, as more complex forms are likely to extend AS-unit length. For lexical complexity, with highly specified content of the stories for the tasks in the main study, D value might not fully disclose candidates' knowledge of vocabulary. However, it was expected that correlation studies with the ratings in this category would reveal whether it was sensible to examine the lexical variety of performance on spoken narrative tasks.

For Coherence, as noted in Section 4.5.3.4, descriptors at A1 and A2 levels are

quite similar in their mentioning only cohesive devices such as *and, then* and *because*. A2+ descriptors also state: "Can use the most frequently occurring connectors to link sentences in order to tell a story... ". Descriptors about coherence (i.e. organisational patterns) start to appear at B2 level. Thus, it was thought that the incidence of positive additive connectives (i.e. the amount of coordination with *and, next*, etc.) might correlate negatively with the CEFR levels, as the use of simple connectives such as *and* might be expected to decrease as the levels go up. If this variable did not correlate highly, then it would also bear useful implications about assessing the performance on spoken narrative tasks in this category.

Lastly, the rating category of Sustained Monologue includes descriptors about the text genre at each level (i.e. description (A2, B1, B2, C1) and narrative (B1, C1)) and the degree of detailed descriptions (at B1, B2, and C1). The variables of idea units were aimed at indicating how detailed the narrative stories were, and it was assumed that this variable might correspond well with the ratings in this category. The correlation coefficients and significance were calculated using Pearson's product-moment correlation using PASW 17.0.

Chapter 5: Results

This chapter reports the results of the analyses for the research questions and brief interpretations of them. Each section in this chapter corresponds to a research question and starts by describing the data and their preliminary analysis (e.g. order effect) before presenting the main results. The results for RQ1 comprise the MFRM analysis of the difficulty of the two spoken narrative tasks, based on the overall CEFR ratings as well as the ratings in the five rating categories. The following section describes the results of *t*-tests for candidate perceptions towards the tasks, both for the whole candidate population (RQ2-1) and for different levels of proficiency based on the MFRM analysis of candidate ability (RQ2-2). In addition, following on from suggestions raised in Pilot Study 5, RQ2-3 explores expert judgements by Japanese teachers on the two tasks, regarding task complexity factors, based on the Checklist of Difficulty by Weir and Wu (2006). Two Japanese teachers of English at TUFS, who were involved in teaching the majority of the candidates, completed the questionnaire. Moreover, the results for RQ2-4 present answers to a brief interview about the perceived difficulty of the tasks by the eleven native English speakers who performed them both. RQ3 (3-1 and 3-2) describes the linguistic performances for the two tasks by the candidates, again for the whole sample and for different levels of proficiency. NS baseline data are also presented. Finally, RQ4 aims to conduct a validation study of the linguistic variables used in RQ3. Discussion of the findings will be presented in Chapter 6.

5.1. Difficulty of the Two Spoken Narrative Tasks Calculated by MFRM Analysis

5.1.1. Data

The data are analysed in order to answer Research Question 1, "Is the difficulty of the two spoken narrative tasks the same according to MFRM analysis?"; they are obtained from the performances of 65 participants on the two narrative tasks, as rated by seven raters. It was decided to use only the ratings data of seven of the ten raters who initially participated in the rater training, as two dropped out and one turned out to be mistiffing (i.e. not using the rating scale consistently), with an infit square value of 1.76 which deviates from the acceptable range of 0.4 to 1.2 (Wright & Linacre, 1994: 370).

Analyses were first conducted on the overall CEFR ratings, and then on the ratings in the other five rating categories as allotted to the performances by the same individuals. Here, the overall CEFR ratings refer to the ratings of Considered Judgement (CJ), which the raters arrived at after considering all five categories (i.e. Range, Accuracy, Fluency, Coherence and Sustained Monologue) for each spoken narrative performance. The CEFR levels were converted into numerical values to facilitate statistical analyses, starting with 1 for Below A1, 2 for A1, and so on up to 10 for C1.

It was confirmed by a two-way mixed-design analysis of variance (ANOVA) that there was no task-order effect on the average ratings by the seven raters for either task tasks, as summarised in Table 5.1, below.

Table 5.1

ANOVA Results for the Average Ratings by Order and Task

Rating Category	Order	Task A		Task B		*N*	Order*Task		
		M	*SD*	*M*	*SD*		*df*	*F*	*p*
Considered Judgement	A→B	5.43	.90	5.67	.89	32	1	.067	.797
	B→A	5.79	1.45	6.06	1.38	33			
Range	A→B	5.46	.87	5.67	.84	32	1	.005	.944
	B→A	5.82	1.39	6.02	1.32	33			
Accuracy	A→B	4.96	.91	5.17	.88	32	1	.000	.982
	B→A	5.49	1.40	5.71	1.34	33			
Fluency	A→B	5.45	1.16	5.71	1.02	32	1	.055	.816
	B→A	5.86	1.64	6.15	1.53	33			
Coherence	A→B	5.21	.93	5.74	.86	32	1	.079	.779
	B→A	5.50	1.46	5.99	1.24	33			
Sustained Monologue	A→B	5.41	.90	5.67	.90	32	1	.539	.466
	B→A	5.78	1.39	6.14	1.40	33			

5.1.2. Considered Judgement (CJ) Ratings

5.1.2.1. Examining the Rating Scale

To answer RQ1 first required a close examination of the use of the CJ rating scale by the raters in order to ensure the credibility of the ratings. This can be done by using four features from the *category statistics* calculated by FACETS (Bond & Fox, 2007: 227). One of the features to be looked at is the *average measure,* which indicates the average candidate-ability measure that each level represents. This should increase as the level goes up (Bond & Fox, 2007: 223). The second feature, called *threshold*, shows the lowest candidate-ability measure that a level is most likely to be assigned (Linacre, 2009: 167). Like the average measure, this feature should increase monotonically across the levels. The third feature is the *outfit mean square* of each level which estimates the fit to the model; if it is larger than 2.0, collapsing the level to an adjacent

level should be considered (Linacre, 2004: 268).

The rating category statistics are summarised in Table 5.2. The average measures show a steady increase across the levels, as do the thresholds. The outfit mean-square values were less than than 2.0.

Table 5.2
Summary of Category Statistics for the CJ Rating Scale (10 CEFR Levels)

Level	Average Measure	Threshold	Outfit MnSq
Below A1	-4.64	None	.90
A1	-3.47	-7.11	1.00
A1+	-2.50	-3.51	.80
A2	-1.71	-3.12	.90
A2+	-.62	-1.58	1.00
B1	.63	-.28	1.00
B1+	1.69	1.67	1.00
B2	3.00	2.80	1.10
B2+	4.10	4.63	1.10
C1	5.16	6.50	.80

The only concern with these data was that the distances between the thresholds of A1+ (-3.51 logits) and A2 (-3.12 logits) were rather small (0.39 logits). This might be regarded as problematic as a distance less than 1.4 logits means that the two levels might not be distant enough to represent two distinct levels (Bond & Fox, 2007: 224). However, Linacre (2004: 274) states that 1.4 logits is required when the rating scale has 3 levels and that, as the number of levels increases, the required distance decreases.

The CJ rating scale had 10 levels (from Below A1 to C1), and all the other distances between the adjacent levels were large enough, i.e. more than 1.0 logits, which Linacre (2004: 274) states is the minimum distance required for a rating scale with 5 levels. In addition, collapsing the A1+ and A2 levels would inevitably increase the standard errors of measurement when estimating candidate ability, task difficulty, and rater severity. Therefore, it was decided to leave the CJ rating scale as it was.

5.1.2.2. Estimates of Candidate Ability, Task Difficulty and Rater Severity

As described in the previous chapter, FACETS estimates candidate ability, task difficulty and rater severity, based on the ratings data. All the estimated values are expressed on a logit scale which is assumed to have equal intervals. The mathematical model for estimating the CJ ratings data was as follows (adapted from Linacre, 2009: 12):

$\log(P_{nmijk} / P_{nmijk-1}) = B_n - A_m - C_j - F_k$, where:

P_{nmijk} = probability of candidate *n* receiving a rating of *k* for Task *m* from rater *j;*
$P_{nmijk-1}$ = probability of candidate *n* receiving a rating of *k-1* for Task *m* from rater *j;*
B_n = ability of candidate *n*;
A_m = difficulty of task *m*;
C_j = severity of rater *j*;
F_k = difficulty of rating (i.e. level) *k*.

From the ratings data inputted, FACETS produced estimates that fit the mathematical model above. The results are summarised in a variable map called a 'ruler', shown below as Figure 5.1.

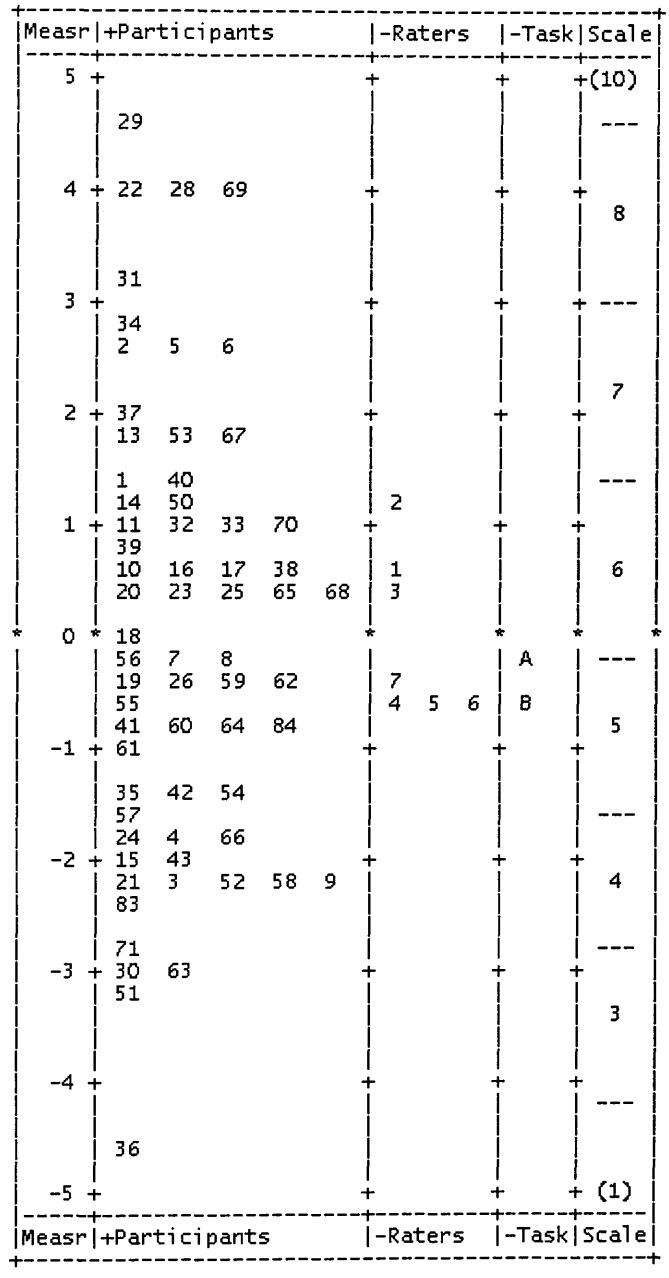


Figure 5.1
FACETS Ruler with CJ Rating Scale

In the first column, *Measr* represents the logit scale on which all the variables are placed. It is centred around zero and larger values represent higher ability or greater difficulty. The second column shows where the candidates (with ID number), according to their ability, fall on the logit scale. The higher they are placed, the more able they are.

According to Figure 1, in my data, the most able candidate is Candidate 29, and the least able one is Candidate 36.

In the third column, raters are positioned according to their severity; the higher they are on the logit scale, the harsher they are. The rater measurement report output by FACETS (summarised below in Table 5.3) shows that Rater 2 has the highest logit value of 1.1 and is therefore the harshest. The separation reliability[24] was 0.98, which indicates that the raters differed significantly in their severity The infit mean squares indicate that the seven raters fit well with the model; all of them fall in the range 0.7 to 1.3 (Bond & Fox, 2007).

Table 5.3
Fit Statistics for Rater Measurement

Rater	Measure	S.E.	Infit Mnsq
2	1.10	.11	.83
1	.57	.11	1.29
3	.44	.11	1.20
7	-.30	.11	.74
5	-.51	.11	.77
4	-.63	.11	1.03
6	-.68	.11	.93

The fourth column displays the difficulty of the tasks. Surprisingly, the two tasks are not positioned horizontally, which means different levels of difficulty, this despite the efforts to ensure parallelness *a priori* in Pilot Study 5. The FACETS statistics show Task A had difficulty of -0.14 logits, while the difficulty of Task B was -0.54 logits, and this difference is statistically significant (χ^2 (1, 455) = 24.7, $p < .01$). Thus, although the difference in logit values for the two tasks was small (0.40 logits),

[24] The separation reliability calculated by FACETS indicates how reliably the analysis is separating the raters into different levels of severity, and thus is different from traditional inter-rater reliability. Traditional inter-rater reliability using a Spearman-Brown Prophecy Formula (Henning, 1987: 82-83), yielded a coefficient of .957 for Task A and .949 for Task B.

Task A was more difficult than Task B, and with statistical significance. However, this statistically-significant difference needed to be complemented by how it would affect the ratings given to the performances on both tasks in order to evaluate how significant the difference was in practice, as statistical significance does not always indicate meaningful difference. This will be discussed in the next section.

5.1.2.3. Effect of Task Difficulty Difference between Tasks A and B

This section discusses how the significant difference between the difficulty of Task A (-0.14 logits) and Task B (-0.54 logits) affects the ratings (i.e. levels) that candidates would receive.

Based on the MFRM mathematical model introduced above, the levels that would be assigned to a candidate can be identified by calculating $B_n - A_m - C_j$, where B_n = ability of candidate *n*, A_m = difficulty of task *m*, C_j = severity of rater *j* (Linacre, 2009: 154-155). It gives a logit value which can fall into one of the ability ranges that the levels represent on the rating scale.

The ability ranges are indicated in the fifth column of the FACETS ruler; this represents the CJ rating scale (1 = Below A1; 10 = C1) with the horizontal dashed lines between the levels indicating transitional points where the probability of receiving a given level starts to exceed the probability of receiving the one below it (Linacre, 2009: 167; Bond & Fox, 2007: 282). The logit values for these transitional points[25] are indicated in the FACETS output and, by using this information, the range of candidate ability that is likely to be assigned each level can be identified. Table 5.4 displays the transitional points for each level.

---

[25] FACETS displays this information under the heading of *expected measure at -0.5.*

Table 5.4

Transitional Points for the Levels

Level	Transitional points
Below A1	
A1	-7.17
A1+	-4.15
A2	-2.85
A2+	-1.60
B1	-.13
B1+	1.49
B2	2.97
B2+	4.66
C1	6.80

As the purpose of this section is to determine the effect of difference in task difficulty, the rater severity for every candidate was set to zero. Let us take as example Candidate 14, who had an ability of 1.12 logits. The logit value he or she would obtain from the calculations for each task is:

On Task A: 1.12 - (-0.14) - 0 = 1.26;

On Task B: 1.12 - (-0.54) - 0 = 1.66.

These values were then compared with the ability ranges for the levels indicated in Table 5.4. As the value of 1.26 would fall in the B1 range ($-0.13 \leq x < 1.49$), Candidate 14 would receive B1 for Task A. However, with 1.66 logits on Task B, which was in the B1+ range ($1.49 \leq x < 2.97$), the assigned level would be B1+.

This calculation was performed for every candidate. It was found that out of 65 candidates who took part in this study, 24 would receive different ratings for the two tasks with a higher level for Task B. Table 5.5 summarises the results with candidate ID numbers.

Table 5.5

Candidates who would be Assigned Different Levels for Tasks A and B

Task A	Task B	Candidates
B1+	B2	2, 5, 6
B1	B1+	1, 11, 14, 32, 33, 40, 50, 70
A2+	B1	19, 26, 55, 59, 62
A2	A2+	4, 15, 24, 43
A1+	A2	30, 51, 63
A1	A1+	36

As can be seen, all 24 candidates would receive a level higher for Task B, and obtaining an adjacent level might not be regarded as so problematic, especially when the difference is between a base level and its plus level (e.g. B1 and B1+). Nevertheless, if these tasks were administered as part of a speaking proficiency test, the effects of the differences would be more critical when base levels were different (e.g. A1+ and A2), and even more so in cases where the CEFR global levels were different (e.g. A2+ and B1). In this regard, the candidates who would be placed in the shaded rows in Table 5.5, where the base levels are different, might be more seriously affected by whichever task was given to them.

5.1.3. Ratings for Range, Accuracy, Fluency, Coherence and Sustained Monologue

5.1.3.1. Examining the Rating Scale

The same procedures used for the Considered Judgement rating scale were taken to examine the category statistics for the five rating scales of Range, Accuracy, Fluency, Coherence and Sustained Monologue. The initial rating scales, with ten levels (Below A1 to C1), did not display monotonic increases in average measures and thresholds due to underuse of the plus levels (i.e. A1+, A2+, B1+, B2+). Therefore, the plus levels were collapsed into the base levels, leaving 5 levels to be assigned: Below

A2, A2/A2+, B1/B1+, B2/B2+, and C1. All the category statistics for the revised rating scales were acceptable, and are shown below in Table 5.6.

Table 5.6
Summary of Category Statistics for the Revised Rating Scales (five CEFR Levels)

Category	Level	Average Measure	Threshold	Outfit MnSq
Range	Below A2	-5.05	None	.90
	A2/A2+	-2.95	-5.54	1.00
	B1/B1+	-.55	-2.03	.90
	B2/B2+	1.93	1.91	1.10
	C1	3.42	5.66	1.10
Accuracy	Below A2	-5.18	None	.90
	A2/A2+	-2.94	-5.32	1.10
	B1/B1+	-.86	-1.86	1.10
	B2/B2+	1.58	1.50	1.10
	C1	3.77	5.67	.80
Fluency	Below A2	-4.68	None	.80
	A2/A2+	-2.73	-4.94	.80
	B1/B1+	-.42	-1.88	.90
	B2/B2+	1.89	1.60	1.00
	C1	3.85	5.22	.90
Coherence	Below A2	-4.98	None	.90
	A2/A2+	-2.96	-5.21	1.20
	B1/B1+	-.96	-2.26	1.20
	B2/B2+	1.45	1.22	1.10
	C1	3.36	6.25	.90
Sustained	Below A2	-4.84	None	.90
Monologue	A2/A2+	-2.74	-5.28	1.00
	B1/B1+	-.44	-1.93	1.00
	B2/B2+	2.00	2.05	1.10
	C1	3.50	5.17	1.20

5.1.3.2. Estimates of Candidate Ability, Task Difficulty, Rater Severity and Rating Category Difficulty

The ratings data were modelled using the Partial Credit Model (Masters, 1982), which constructs a separate rating-scale structure for each category, to estimate candidate ability, task difficulty and rater severity. The resultant FACETS ruler is shown below in Figure 5.2.

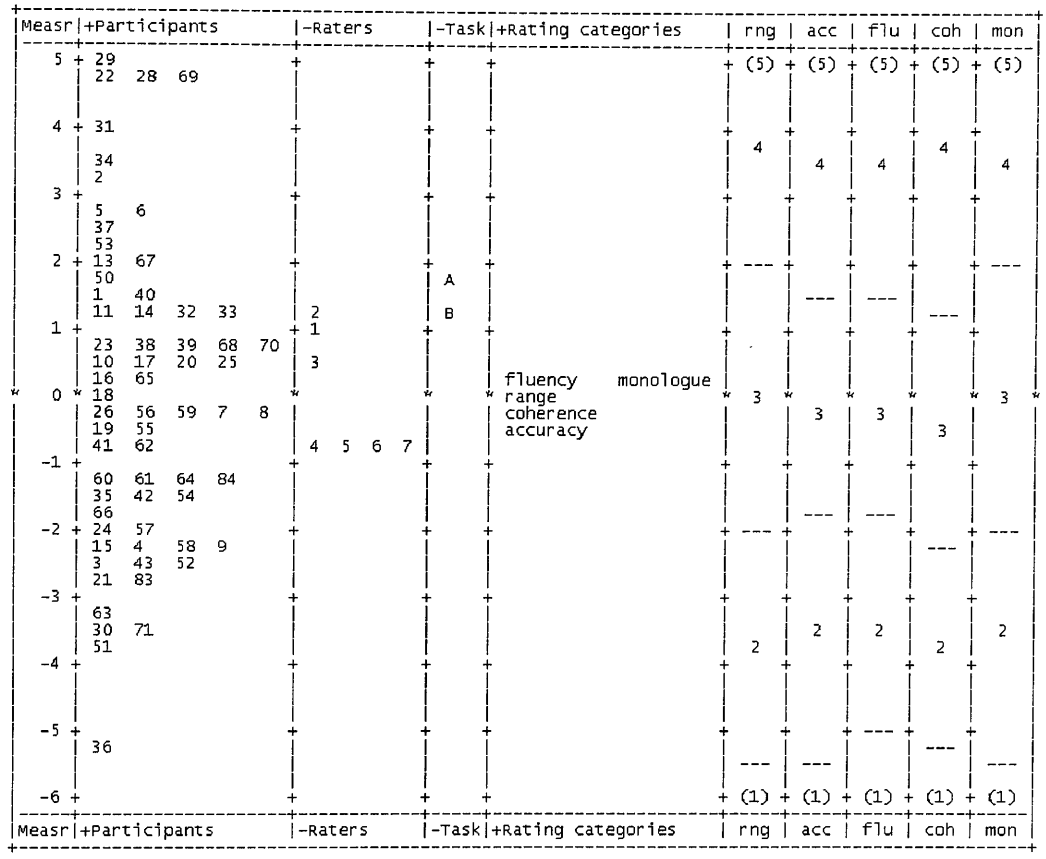


Figure 5.2
FACETS Ruler with Rating Scales for Range, Accuracy, Fluency, Coherence and Sustained Monologue

Again, raters were not positioned horizontally in the third column. Their levels of severity differed (ranging from -.79 (Rater 6) to 1.32 (Rater 2)), which is indicated by the separation reliability[26] of .99. Their rating behaviours fitted well with the model, according to the fit statistics as summarised in Table 5.7, below.

[26] The traditional inter-rater reliability was calculated by using the Spearman-Brown Prophesy Formula, as suggested by Henning (1987: 82-83), and was .912 (Task A) and .916 (Task B) for Range; .918 (A) and .904 (B) for Accuracy; .950 (A) and .936 (B) for Fluency; .910 (A) and .881 (B) for Coherence; .905 (A) and .924 (B) for Sustained Monologue.

Table 5.7
Fit Statistics for Rater Measurement

Rater	Measure	S.E.	Infit MnSq
2	1.32	.07	1.00
1	.94	.07	1.07
3	.58	.07	1.16
5	-.64	.07	.98
7	-.68	.07	.82
4	-.73	.07	.95
6	-.79	.07	.99

The fourth column indicates that, similar to the analysis of CJ ratings, Tasks A and B were not placed horizontally. Their difficulty (Task A: 1.66; Task B: 1.14) were significantly different (χ^2 (1, 2275) = 86.2, $p < .01$). This would, again, cause some candidates to receive different ratings for Tasks A and B, which will be discussed in the next section.

In the fifth column, among the five rating categories, the more difficult categories are placed towards the top of the ruler. It appears that fluency was the most difficult category, and accuracy was the easiest. As in Table 5.8, the fitness values of all categories were appropriate, ranging from .89 to 1.10.

Table 5.8
Fit Statistics for Rating Category Measurement

Rating Category	Measure	S.E.	Infit MnSq
Fluency	.35	.06	.89
Sustained Monologue	.23	.06	1.00
Range	.06	.06	.96
Coherence	-.20	.06	1.10
Accuracy	-.44	.06	1.03

5.1.3.3. Effect of Task Difficulty Difference between Tasks A and B

Using the candidate ability, rater severity, task difficulty and rating category difficulty that were estimated in the previous section, an examination of the effect of the difficulty of the two tasks was conducted with the same procedure as that described in Section 5.1.2.3.

With the rating category difficulty added to the model, the mathematical equation becomes as below (Linacre, 2009: 12):

$\log (P_{nmijk} / P_{nmij(k-1)}) = B_n - A_m - D_i - C_j - F_k$, where:

P_{nmijk} is the probability of category *k* being observed;
$P_{nmij(k-1)}$ is the probability of category *k-1* being observed;
B_n is candidate ability;
A_m is task difficulty;
D_i is rating category difficulty;
C_j is rater severity;
F_k is the difficulty of ratings (i.e. levels) relative to level *k-1*.

Based on this equation, the level that a candidate is likely to be assigned for a task in a rating category can be determined by calculating B_n (candidate ability) - A_m (task difficulty) + D_i (rating category difficulty – shown in Table 5.8 as Measure), and by comparing the resultant value to the transitional points of the levels (i.e. where the probability of being assigned level *k* becomes higher than that for level *k-1*) that are also output by FACETS (summarised in Table 5.9, below).

Table 5.9
Transitional Points for the Levels in Each Rating Category

Level	Range	Accuracy	Fluency	Coherence	Sustained Monologue
Below A2					
A2/A2+	-5.6	-5.38	-5.04	-5.32	-5.35
B1/B1+	-1.99	-1.84	-1.82	-2.19	-1.87
B2/B2+	1.89	1.54	1.6	1.27	1.98
C1	5.71	5.71	5.29	6.26	5.28

Suppose that the rater severity was the same (i.e. zero), then the values for Candidate 5, for example, with an ability of 2.86 in Fluency, would be:

Task A: 2.86 - 1.66 + 0.35 = 1.55;
Task B: 2.86 - 1.14 + 0.35 = 2.07.

These values were then compared to the transitional points (Table 9) in order to find out which level's ability range Candidate 5 fell in for Fluency. As the transitional point between B1/B1+ and B2/B2+ was 1.6, Candidate 5 would receive B1/B1+ for Task A, and B2/B2+ for Task B.

This procedure was applied to all candidates, and those who would receive different ratings for Tasks A and B in each rating category were identified. Table 5.10 shows them with their candidate ID numbers. Altogether, there were 17 candidates out of 65 who would receive different ratings for Tasks A and B in one or more rating category. Table 5.11 classifies these 17 candidates according to the number of rating categories in which they would receive different ratings.

Table 5.10

Candidates who would be Assigned Different Levels for Tasks A and B

Category	Task A	Task B	Candidates
Range	B1/B1+	B2/B2+	2
	A2/A2+	B1/B1+	19, 41, 55, 62
Accuracy	B1/B1+	B2/B2+	2, 34
	A2/A2+	B1/B1+	7, 18, 26
	Below A2	A2/A2+	30, 51, 71
Fluency	B1/B1+	B2/B2+	5, 6, 37
	A2/A2+	B1/B1+	41, 62
	Below A2	A2/A2+	51
Coherence	B1/B1+	B2/B2+	5, 6
	A2/A2+	B1/B1+	8, 19, 41, 55, 59, 62
	Below A2	A2/A2+	30, 51
Sustained Monologue	B1/B1+	B2/B2+	2
	A2/A2+	B1/B1+	41, 55, 62

Table 5.11

Number of Rating Categories with Different Levels for Tasks A and B

Different Rating(s) in	Candidates
1 category	7, 8, 18, 26, 34, 37, 59, 71
2 categories	5, 6, 19, 30
3 categories	2, 55, 51
4 categories	41, 62

It is unlikely that discrepant ratings for a single rating category on a task type would affect the overall eventual level to be assigned in a testing situation; however, different ratings in two or more categories could lead to the assignment of different overall levels, depending on how they perform in other tasks in the test. In the next chapter, the cause of this difference in task difficulty, which emerged despite every effort being made to ensure the parallelness of the two tasks, is discussed, together with the results of Research Questions 2-1 to 2-4.

5.2. Candidate Perceptions of the Two Spoken Narrative Tasks

5.2.1. Data

The responses by the 65 Japanese candidates to the Task Difficulty Questionnaire by Robinson (2001), about each task, were analysed in this section in order to answer Research Question 2-1, "Are the candidates' perceptions of the two tasks the same?", which aimed to investigate candidate perceptions. The questionnaire included five questions, with a 9-point Likert scale, regarding candidates' perceived difficulty of the task, their nervousness in completing it, how well they thought they did on it, how interesting it was, and whether they would have liked to do more tasks like it.

As a first step, the order effect was examined by using two-way mixed design analyses of variance (ANOVA) on the responses to 5 questions by 65 Japanese candidates. If the interaction between the order and tasks was significant, then there was an order effect. This procedure aimed to investigate whether or not the counter-balanced design of this study yielded a practice effect on the candidates' perceived difficulty and, if it did, to exclude problematic questions from further analyses.

Table 5.12 shows the results of the ANOVA by order and task. As indicated by the asterisks, Questions 2 and 3 turned out to have been affected by the task order. For Question 2, Task A-first group showed that they were more nervous about Task A (M = 5.63), whereas the Task B-first group was more nervous about Task B (M = 5.09). Thus, the candidates were more anxious about whichever task was given to them first. Question 3 indicated a similar tendency; the candidates thought that they did better on the second task that they did. Task A-first group thought that they did better on Task B (M = 3.84), and Task B-first group felt they had performed better on Task A (M = 3.76). The practice effects for the two questions are quite understandable, as the candidates may have felt less nervous and thought that they did better on the second task because

they had got more used to narrating by completing the first task. As it is not meaningful to include responses affected by task order in the statistical tests of difference, these two questions were excluded from further analysis.

Table 5.12
ANOVA Results for the Task Difficulty Questionnaire by Order and Task

Questions	Order	Task A		Task B		*N*	Order*Task		
		M	*SD*	*M*	*SD*		*df*	*F*	*p*
1. How difficult was the task?	A→B	5.56	1.80	5.47	1.69	32	1	1.721	.194
	B→A	5.79	1.39	5.30	1.53	33			
2. How nervous were you?	A→B	5.63	1.91	5.06	2.00	32	1	19.585	.000 **
	B→A	4.64	2.54	5.09	2.49	33			
3. How well do you think you did?	A→B	3.19	1.55	3.84	1.82	32	1	8.898	.004 **
	B→A	3.76	1.44	3.55	1.42	33			
4. How interesting did you think the task was?	A→B	6.34	1.18	6.13	1.13	32	1	.783	.380
	B→A	6.39	1.43	6.36	1.34	33			
5. Would you like to do more tasks like this?	A→B	6.31	1.45	6.19	1.60	32	1	.182	.671
	B→A	6.30	1.70	6.27	1.65	33			

Note. $^{**}p < .01$.

5.2.2. Results of *t*-tests

Related sample *t*-tests were conducted on the answers to the three remaining questions on the Task Difficulty Questionnaire: Q1: How difficult was the task?; Q4: How interesting did you think the task was?; and Q5: Would you like to do more tasks like this? The results are summarised in Table 5.13, below.

Table 5.13
Results of t-tests (N = 65)

Questions	Task A		Task B					
	M	*SD*	*M*	*SD*	*t*	*df*	*p*	Cohen's *d*
1. How difficult was the task?	5.68	1.59	5.38	1.60	1.950	64	.056	.188
4. How interesting did you think the task was?	6.37	1.31	6.25	1.24	1.158	64	.251	.094
5. Would you like to do more tasks like this?	6.31	1.57	6.23	1.61	.697	64	.488	.050

No significant differences were found for any of the three questions, although Questionnaire 1 was approaching significance ($t(64) = 1.950$, $p = .056$) with Task A perceived as more difficult ($M = 5.68$, $SD = 1.59$) than Task B ($M = 5.38$, $SD = 1.60$). This might suggest that the candidates' perceptions were in line with the difficulty of the two tasks as calculated by FACETS; however, the difference of .30 seems ignorable considering the 9-point scale in this questionnaire. Thus, the two tasks were parallel in the candidates' perceptions.

5.3. Candidate Perceptions of the Two Spoken Narrative Tasks at Different Levels of Proficiency

To answer Research Question 2-2, "Are the candidates' perceptions of the two spoken narrative tasks the same at different levels of proficiency?", the dataset was divided into four levels according to the 'fair average' values of ability as estimated by FACETS and based on the Considered Judgment (CJ) ratings (i.e. overall CEFR levels).

When FACETS calculates a candidate's ability measure, taking rater severity into account, it also outputs an adjusted measure called 'fair average'. This measure is calculated assuming that the rater severity measures zero (i.e. there is no difference in severity between the raters), thus adjusting raw ratings for severe and lenient raters

(Linacre, 2009: 223). Utilising the participants' fair average ability measure, the various raw ratings could be summarised reliably into one single rating for each rating category.

Once the fair average measures of the CJ ratings for every candidate were calculated, they were rounded off and labelled with the corresponding CEFR levels.[27] The levels in the rating scales used in this study included Below A1 (coded as 1) to C1 (coded as 10); so, for instance, if a candidate had the fair average of 4.69, it was rounded up to 5 (which represents A2+ level) and he or she was assigned an A2+. In this way, a single rating, based on the CJ ratings for each candidate's performance on the two tasks, was obtained. The adjusted levels were then used as a grouping variable for different proficiency levels, and then the plus levels were incorporated into the base levels in order to increase the sample size of each level. The resultant levels were as follows: A2/A2+ (n = 25), B1/B1+ (n = 31), B2/B2+ (n = 8); there was only one candidate at A1/A1+ level, so she was excluded from further analyses.

Then, related sample t-tests were used at A2/A2+ and B1/B1+ levels, and paired-sample Wilcoxon signed-rank tests were conducted at B2/B2+ level because of the small sample size at this level. The results are shown in Tables 5.14 to 5.16, below.

Table 5.14
t-test Results at A2/A2+ Level

Questions	Task A		Task B				
	M	*SD*	*M*	*SD*	*t*	*df*	*p*
1. How difficult was the task?	6.36	1.41	6.40	1.41	-0.166	24	.870
4. How interesting did you think the task was?	6.00	1.22	5.88	1.24	0.592	24	.559
5. Would you like to do more tasks like this?	5.88	1.39	5.56	1.50	1.554	24	.133

[27] This procedure is illustrated in Eckes (2009).

Table 5.15

t-test Results at B1/B1+ Level

Questions	Task A		Task B				
	M	*SD*	*M*	*SD*	*t*	*df*	*p*
1. How difficult was the task?	5.06	1.24	4.81	1.01	1.545	30	.133
4. How interesting did you think the task was?	6.65	1.08	6.45	1.03	1.438	30	.161
5. Would you like to do more tasks like this?	6.58	1.57	6.68	1.45	-0.722	30	.476

Table 5.16

Results of Wilcoxon tests at B2/B2+ Level

Questions	Task A		Task B			
	Median	*Range*	*Median*	*Range*	*Z*	*p*
1. How difficult was the task?	5.50	7.00	5.00	7.00	-1.802	.072
4. How interesting did you think the task was?	5.00	5.00	5.50	4.00	-.447	.655
5. Would you like to do more tasks like this?	5.50	5.00	6.00	5.00	.000	1.000

At all levels, no significant differences were found, which indicates that the two tasks were parallel in the candidates' perceptions. However, an interesting pattern was displayed for Question 1. At A2/A2+ level, there was hardly any difference in the means between Task A (M = 6.36) and Task B (M = 6.40), with a very high p value of .870. As the levels went up, however, Task B was regarded as easier, with smaller p values. At B1/B1+ level (Task A: M = 5.06; Task B: M = 4.81, p = .133), and at B2/B2+ level, the mean difference grew larger (Task A: *Mdn* = 5.50; Task B: *Mdn* = 5.00), with the difference approaching significance (p = .072), despite the small sample size. This might suggest changes in candidate perceptions according to proficiency levels, which will be discussed further in the next chapter.

5.4. Expert Judgements of the Two Spoken Narrative Tasks by Japanese Teachers Regarding Task Complexity Factors

RQ2-3, "Do Japanese teachers judge the two spoken narrative tasks to be

parallel for the candidates in terms of the relevant task complexity factors?", was established in order to examine parallelness from the Japanese teachers' point of view; they were expected to provide expert judgements based on their knowledge of the candidates. The two Japanese teachers who were teaching the majority of the candidates at TUFS answered whether they agreed or disagreed with the 9 statements taken from the Checklist of Difficulty by Weir and Wu (2006). Table 5.17 summarises their responses. Where they answered "No", they provided explanations as to why they thought that way.

Table 5.17
Expert Judgements of Task Complexity Factors by Japanese Teachers

Statement	Teacher 1	Teacher 2
1. The roles of people in the pictures are equally familiar to the candidates.	No	Yes
2. The locations in the pictures are equally familiar to the candidates.	Yes	Yes
3. The events in the pictures are equally familiar to the candidates.	Yes	Yes
4. The objects in the pictures are equally familiar to candidates.	No	Yes
5. The objects and events to be described by the candidates are equally visible in the pictures.	No	No
6. There are enough details in the pictures for the candidates to complete the task.	No	Yes
7. The lexical items required to describe the pictures are equally familiar to the candidates.	No	No
8. The grammatical structures required to describe the pictures are equally familiar to the candidates.	Yes	Yes
9. The language functions required to describe the pictures are equally familiar to the candidates.	Yes	Yes

Both teachers answered "Yes" to Statements 2, 3, 8 and 9, which indicates that the candidates should be equally familiar with the locations, events, required grammatical structures and language functions (i.e. description and narration).

However, at least one of them answered "No" to all other statements, which raised doubts about the parallelness of the two tasks in terms of some of the task complexity factors for the candidates.

The reasons that the teachers offered for disagreeing with the statements clearly illustrate the intricate relationship between task complexity factors and candidate factors. For disagreeing with Statement 1, Teacher 1 answered that the role of the balloon-seller in Task A might not be familiar to the candidates, as such children-targeted sellers, walking door to door, are not seen in Japan. Such knowledge is cultural, which should be avoided in language tests. Moreover, on disagreeing with Statement 4, Teacher 1 also pointed out the candidates' unfamiliarity with washing-related objects, such as the basin (that the mother is using to wash clothes in) and the washing-line in Task A. Washing clothes in a basin may be unfamiliar to many nowadays, considering the widespread use of washing machines, in contrast to the time when the two tasks were first published by Hill (1960). However, a washing line is a different matter, because firm metal bars (placed overhead in a yard or balcony) are used to hang clothes in Japan instead of a rope. This again is cultural knowledge, which also relates to the linguistic knowledge that the Japanese candidates might lack for such objects. In fact, Teacher 1 pointed out this lack of corresponding lexical knowledge when explaining why he disagreed with Statement 7, which will be explained later.

With regard to Statement 5, both teachers disagreed with the concerns about the visibility of the object that replaced the baby in Task B, i.e. whether it was a ball or a mask. For Task A, Teacher 1 mentioned that the gender of one of the children was not instantly recognisable. Additionally, he noted that the piece of clothing used as the body of the ghost-like figure was not clear (whether it was a dress, robe or something else) in Task A.

Teacher 1 answered "No" to Statement 6 about the sufficiency of details drawn in the pictures. He found the changes of setting in Task A perplexing, stating that the locations of the laundry room, window and backyard would not be easily understood by candidates, and therefore it would be difficult for them to describe the story smoothly. He also expressed concern about Task B in which appeared to lack a clear cause as to what finally woke the mother up.

Finally, Statement 7 was disagreed with by both teachers, who doubted that the candidates would have the lexical knowledge to describe the ghost-like figure in Task A. Relating to the notes he made in Statement 4, Teacher 1 pointed to the lack of washing-related vocabulary in the candidates' linguistic knowledge, such as *basin* and *washing line*. According to him, the candidates were unlikely to have been taught such English vocabulary in junior or senior high school in Japan, partly because washing is not a common subject in authorised English textbooks, and partly because the use of basin and the very notion of a washing line may be culturally unfamiliar.

In sum, there was more counter-evidence against the parallelness of the two tasks in terms of the judgements by the Japanese teachers, which were not raised in Pilot Study 5 by the experts with various L1 backgrounds. The clarity of the pictures, as well as the sufficiency of details in the pictures, were questioned. What the responses by Teacher 1 for Statements 4 and 7 suggest is also striking; the linguistic complexity (i.e. *code complexity* in Skehan's (1998) terminology), which is included in the task complexity factors, is not free from the influence of cultural knowledge that candidates hold. These findings are further discussed in the next chapter.

5.5. Perceived Difficulty of the Two Spoken Narrative Tasks by English Native Speakers

Research Question 2-4, "Do English native speakers perceive the two spoken narrative tasks as equally difficult?", explored whether English native speakers (NS), who are free of lack of linguistic knowledge and automaticity, unlike Japanese candidates, would find the two tasks parallel. The 11 NS were asked a question after completing both tasks, "Did you feel either of the two tasks was more difficult?" If they answered yes, the reason behind it was explored further. As a result, 7 out of 11 NS perceived the difficulty of the two tasks as being dissimilar, with 6 of them finding Task A to be more difficult. Table 5.18 summarises the results of their brief interviews.

Table 5.18
Summary of the Perceived Difficulty of the Two Spoken Narrative Tasks by English Native Speakers

NS	Answer: More difficult task	If yes, why?
1	No	-
2	Yes: A	Was able to relate to B more because of babysitting experience.
3	Yes: A	Took some time to figure out what the children were using when creating a man.
4	No	-
5	Yes: B	Could not tell if the mother was more upset or frightened when she woke up.
6	No	-
7	Yes: A	It was difficult to work out what happened to the balloon in the end.
8	Yes: A	Needed to work out where the woman went to and who the children were.
9	No	-
10	Yes: A	Needed to work out who the children were.
11	Yes: A	A was more difficult because of explaining about the balloon man. To figure out the ball in B was slightly difficult.

NS 1, 4, 6, and 9 did not perceive the two tasks to be different in difficulty. The reasons behind the difference in perceived difficulty provided by the other NS included three factors: past experience, lack of clarity of the pictures, and insufficiency of details in the pictures. Firstly, NS2 perceived Task B to be easier to perform because she could relate to the story, as she had previously looked after a baby. This suggests that perceived difficulty can be influenced by past experience, which may relate to topic familiarity. Secondly, two NS reported problems with clarity of the pictures in Task B. NS 5 stated that she was not sure "whether she [the mother] was more upset that the baby wasn't there or she just got afraid because of what was in the basket". Likewise, NS 11 reported, "it was unclear what this [pointing at the ball in Task B] was supposed to be – a ball with a face on it, presumably".

Thirdly, insufficiency of details was raised for Task A by 5 NS, regarding the time gap between the pictures, the relationship between the characters, and how the ghost-like figure was made. For example, NS 8 said, "It didn't seem to link, like this one [pointing at Picture 2 in Task A] and this [pointing at Picture 3 in Task A] didn't seem to link together, like it's kind of she's taking out the washing but now where's she gone and who were they, so you have to sort of make up who these people were and where she'd gone, so it was hard". This remark clearly indicates that NS 8 had to do more thinking when performing Task A, because of the time gap between Pictures 2 and 3, as well as the unclear relationship between the woman who was doing the washing and the two children who suddenly appeared in the backyard. Similarly, NS 10 found Task A more difficult since she "wasn't sure if they were grandchildren or strangers" to the woman. The third element, that lacked sufficient details (i.e. how the ghost-like figure was made), caused 3 NS to perceive Task A to be more difficult. Direct quotations from

their reports are listed below:

- Trying to explain some of the bits like, trying to explain that they were creating a man, dressed in a robe that was from the washing line, that was more difficult. (NS 3)
- At first when I looked at the A picture I didn't realise that it was gonna be one of these balloons, then as I worked it out along the sequence it was fine. (NS 7)
- The second one [Task A], how to describe this, how to describe this strange figure [pointing at the ghost-like figure], that was really difficult to explain. (NS 11)

These results indicate that even NS, who, unlike non-native speakers, are not constrained by a lack of linguistic resources, can find the two tasks differ in their difficulty due to topic familiarity and lack of clarity and detail in the pictures. It is interesting to notice that some of the reasons mentioned overlap with the expert judgements by the Japanese teachers (reported in Section 5.3), which will be discussed further in the next chapter.

5.6. Linguistic Performances in the Two Narrative Tasks

5.6.1. Data

The values of the linguistic variables for the narrative performances by the 65 Japanese candidates in the two tasks were tested using related sample *t*-tests for significant mean differences for RQ3-1, "Are the performances on the two spoken narrative tasks the same in terms of the linguistic variables?". After the data were divided into different proficiency levels, 64 Japanese candidates remained in the dataset. Because of the uneven sample sizes and distributions across the levels, related sample *t*-tests were used for the A2/A2+ and B1/B1+ levels, and Wilcoxon signed-rank tests were applied for the B2/B2+ level. The results are discussed with reference to the baseline data for the eleven native speakers of English, which were also examined by Wilcoxon signed-rank tests due to the small sample size and skewness of the data.

5.6.1.1. Order Effect

As a first step, order effect was examined by using two-way mixed design ANOVA across all the linguistic variables in the dataset of 65 Japanese candidates. If the interaction between the order and tasks was significant, then there was an order effect. This procedure was to investigate whether or not the counter-balanced design of this study yielded a practice effect on any variables and, if it did, then problematic variables could be excluded from further analyses.

Table 5.19 shows the results of ANOVA by order and task. On scrutinising the descriptive statistics, the candidates who were given Task B first and Task A second appeared to have higher values, for most of the variables for both tasks, than those who were given Task A first and Task B second. This may be attributed to the slightly higher

ability of the Task B-first group, although the task order was randomly assigned; the mean logit value of overall ability (as calculated for RQ1) for the Task B-first group was 0.30 (SD = 2.39), whereas the Task A-first group had a mean logit ability value of -0.31 (SD = 1.46). However, the difference was not significant ($t(53.226) = -1.248$, $p = .217$).

The ANOVA results showed no significant interactions between task order and tasks, hence there was no order effect, except for the number of errors per AS-unit ($F = 4.916$, $p < .05$). This was a surprising finding as the three variables for accuracy were supposed to be measuring the same construct, but the order effect was observed only in one of them. It was decided to include the number of errors per AS-unit in a further analysis, since the actual order effect was considered small, as indicated by the magnitude of power,[28] which was .588, as calculated by PASW 17.0. Still, this discrepancy between the three accuracy variables clearly deserve more investigation, thus further examination of this matter will be discussed in Chapter 6.

[28] It was shown under the heading of "observed power" in the PASW output.

Table 5.19

ANOVA Results for Linguistic Variables by Order and Task

Variables	Order	Task A		Task B		*N*	Order*Task		
		M	*SD*	*M*	*SD*		*df*	*F*	*p*
D value	A→B	33.34	8.65	31.55	8.06	32	1	.340	.562
	B→A	31.22	8.35	30.55	9.77	33			
AS-unit length	A→B	8.52	.90	8.41	1.35	32	1	.500	.482
	B→A	8.80	1.79	8.94	1.64	33			
Subordinate clauses per AS-unit	A→B	.15	.11	.22	.14	32	1	.011	.918
	B→A	.19	.17	.26	.16	33			
Speech rate	A→B	80.53	20.08	82.26	21.05	32	1	1.712	.195
	B→A	87.15	32.02	91.92	30.44	33			
% of error-free clauses	A→B	45.91	16.41	55.74	20.05	32	1	2.115	.151
	B→A	58.63	23.34	61.48	23.04	33			
Errors per 100 words	A→B	9.75	3.26	8.47	4.41	32	1	1.951	.167
	B→A	7.41	5.59	7.33	5.36	33			
Errors per AS-unit	A→B	.82	.26	.69	.34	32	1	4.916	.030*
	B→A	.61	.41	.64	.45	33			
Positive additive connectives	A→B	94.29	21.63	92.79	24.60	32	1	.047	.829
	B→A	88.88	25.50	88.67	26.08	33			
No. of main ideas	A→B	8.16	.95	7.38	1.01	32	1	.093	.762
	B→A	8.24	.75	7.36	.93	33			
No. of minor ideas	A→B	4.59	1.56	4.50	1.63	32	1	1.293	.260
	B→A	4.39	1.92	4.88	1.93	33			

Note. *$p < .05$.

5.6.1.2. Descriptive Statistics for the 65 Japanese Candidates' Data (RQ3-1)

The means, standard deviations, standard errors, and values for skewness and kurtosis were calculated for all the remaining variables to gather information on their central tendencies and dispersion. The descriptive statistics are summarised in Table 5.20, below.

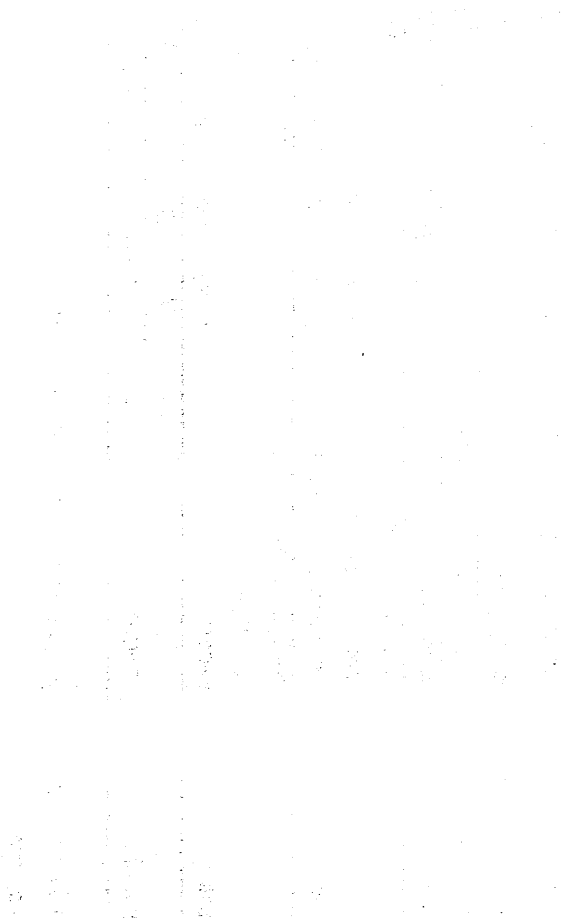


Table 5.20

Descriptive Statistics for Linguistic Variables (N = 65)

Category	Variables	Mean		SD		Skewness		Kurtosis	
		A	B	A	B	A	B	A	B
Range	D value	32.26	31.04	8.50	8.91	.372	.648	.992	.006
	AS-unit length	8.66	8.68	1.42	1.51	1.056	.637	2.901	.247
	Subordinate clauses per AS-unit	.17	.24	.14	.15	.961	.253	.416	-.851
Accuracy	% of error-free clauses	52.37	58.66	21.07	21.64	-.133	-.020	-.671	-.736
	Errors per 100 words	8.56	7.89	4.71	4.91	.524	.551	.028	-.229
	Errors per AS-units	.71	.66	.35	.40	.473	.577	.660	.110
Fluency	Speech rate	83.89	87.17	26.82	26.49	.627	.477	.174	.347
Coherence	Positive additive connectives	91.54	90.70	23.65	25.25	.385	-.096	.278	-.486
Sustained Monologue	No. of main ideas	8.20	7.37	.85	.96	-1.028	-.813	1.466	.001
	No. of minor ideas	4.49	4.69	1.74	1.78	.169	.584	-.115	.527

The normality of each variable's distribution was checked by examining the values for skewness and kurtosis, in order to decide if using parametric tests was appropriate. If the values for skewness and kurtosis are in the range of -2 to +2, then the variable can be assumed to have "a reasonably normal distribution: (Bachman, 2004: 74).

5.6.1.3. Descriptive Statistics for A2/A2+ at Native Speaker Level (RQ3-2)

For the latter half of RQ3-2, "Are the performances in the two spoken narrative tasks the same in terms of the linguistic variables at different levels of proficiency?", the same CEFR levels were assigned to the candidates as for RQ2 (i.e. based on the fair average of Considered Judgement ratings). Accordingly, the A1+ level was removed from the dataset because there was only one candidate. Table 5.21 summarises the descriptive statistics for the 64 remaining candidates at A2/A2+, B1/B1+ and B2/B2+ levels, as well as for the 11 native speakers of English (NS) as baseline data.

Table 5.21

Descriptive Statistics for A2/A2+, B1/B1+, B2/B2+ and Native Speaker Levels

Category	Variables	A2/A2+ (n = 25)								B1/B1+ (n = 31)							
		Mean		SD		Skewness		Kurtosis		Mean		SD		Skewness		Kurtosis	
		A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
Range	D value	28.01	25.98	8.03	6.74	.15	.89	-.12	1.14	33.76	32.46	6.88	7.49	.35	.36	.40	.22
	AS-unit length	8.03	8.19	1.13	1.21	.99	.63	2.96	1.14	8.96	8.84	1.06	1.39	.23	.16	.12	-.35
	Subordinate clauses per AS-unit	.12	.22	.10	.16	.74	.55	.72	-.73	.20	.24	.16	.14	.53	.08	-.85	-.78
Accuracy	% of error-free clauses	41.91	43.98	18.62	15.94	-.02	-.01	-.37	-1.20	56.10	66.47	18.12	19.35	-.47	-.06	-.14	-.35
	Erros per AS-unit	.92	.93	.33	.34	.92	.71	1.98	.76	.63	.51	.29	.32	.10	.62	-.81	.51
	Errors per 100 words	11.64	11.46	4.33	4.12	.49	.55	.24	-.25	7.09	5.93	3.29	3.69	.01	.39	-1.06	-.26
Fluency	Speech rate	65.40	67.93	14.24	15.47	.21	.25	-.67	-.31	90.53	96.01	21.02	21.12	.80	1.18	.47	2.92
Coherence	Positive additive connectives	94.18	97.88	26.21	23.43	.88	.21	.52	.23	94.96	89.48	20.56	21.92	-.34	-.24	.28	-.46
Sustained	No. of main ideas	7.72	7.04	.94	1.06	-.71	-.31	1.50	-.70	8.45	7.71	.62	.74	-.69	-1.06	-.40	1.12
Monologue	No. of minor ideas	3.88	4.48	1.72	1.73	.20	.43	-.72	-.28	4.77	4.52	1.71	1.67	.25	.57	.52	1.04

Variables	B2/B2+ (n = 8)								NS (N = 11)							
	Mean		SD		Skewness		Kurtosis		Mean		SD		Skewness		Kurtosis	
	A	B	A	B	A	B	A	B	A	B	A	B	A	B	A	B
D value	39.44	40.93	10.11	10.78	1.13	-.18	1.43	-2.41	42.71	43.32	7.25	8.13	.03	-.69	-1.16	1.92
AS-unit length	9.83	9.64	2.20	2.34	1.19	.14	1.61	-1.48	9.24	10.50	1.51	1.64	.83	1.35	-.45	2.25
Subordinate clauses per AS-unit	.22	.29	.16	.15	1.84	.18	4.54	-.24	.32	.50	.20	.24	.90	.10	.25	-.82
% of error-free clauses	73.53	78.66	20.45	11.34	-1.48	-.54	1.80	-.24	98.24	98.95	2.59	2.34	-1.10	-1.96	-.09	2.32
Errors per AS-unit	.34	.32	.25	.20	1.15	.45	.83	-1.11	.02	.02	.03	.03	.89	2.13	-1.16	3.49
Errors per 100 words	3.50	3.14	2.70	1.63	1.96	.40	4.29	-.51	.26	.06	.38	.20	1.04	3.32	-.45	10.99
Speech rate	122.61	119.91	21.14	18.93	.15	-.31	.69	-1.52	186.89	178.64	49.89	36.65	.56	-.26	-.43	-1.26
Positive additive connectives	73.60	79.26	19.66	34.49	.10	.46	-1.77	-2.13	66.77	62.19	22.63	17.96	-.32	.11	-.92	.38
No. of main ideas	8.88	7.00	.35	1.07	-2.83	-.94	8.00	.35	9.00	9.00	.00	.00	.	.	.	.
No. of minor ideas	5.25	6.25	1.67	1.83	.46	1.18	-.60	2.06	5.09	5.09	2.51	2.55	.76	.81	-1.22	.26

At A2/A2+ and B1/B1+ levels, all the skewness values and most of the kurtosis values were within an acceptable range of -2 to +2. Exceptions were some kurtosis values for AS-unit length in Task A (2.96), and speech rate in Task B (2.90). However, as deviation of kurtosis is less crucial than that of skewness (Coolican, 2004: 292), it was decided to apply parametric tests (i.e. related sample *t*-tests) to these data. This would allow direct comparison of the effect of the tasks using Cohen's *d* values, which was considered to be very useful for comparing two adjacent levels.

B2/B2+ level and the NS data showed higher skewness and kurtosis values in general, and there were some very high values too (e.g. skewness of -2.83 and kurtosis of 8.00 on a number of the main ideas in Task A; skewness of 3.32 and kurtosis of 10.99 on the number of errors per 100 words in Task B). Therefore, for these two groups, non-parametric Wilcoxon signed-rank tests were used.

This RQ was also intended to examine whether there was any interaction between proficiency levels and tasks, which would require two-way mixed design ANOVA; however, due to the non-normality of the data for the B2/B2+ and NS levels, it was not possible to carry out this part of the analysis.

5.6.1.4. Bonferroni Correction

An issue that needs mentioning here is that of adjusting the significance level when conducting multiple tests on the same dataset, known as the Bonferroni correction. This adjustment aims to terminate Type I errors, rejecting a null hypothesis with significance obtained by chance, and divides the significance level by the number of tests to be conducted in order to make the significance tests much more rigorous. Rice (1989) demonstrated that Type I errors increase as more tests are conducted on the same dataset, and therefore strongly recommended the use of a procedure based on

Bonferroni's correction (also known as the Holm correction (Holm, 1979)). However, other researchers have objected to Rice, arguing that Bonferroni and Holm corrections are so severe that they inevitably risk causing Type II errors instead (i.e. rejecting the alternative hypothesis where it should not be rejected) (Cabin & Mitchell, 2000; Nakagawa, 2004).

As this study used two very similar tasks, differing in only one major factor (i.e. whether or not there are changes of scenes in the sequence), any resulting differences in the linguistic variables were expected to be relatively small and so it was not desirable to risk increasing the chance of Type II errors. Therefore, it was decided that this study would not apply a Bonferroni correction to the significance levels, but would report the effect sizes (Cohen's *d*) of any significant differences found, which Cabin and Mitchell (2000: 248) and Nakagawa (2004: 1045) recommend as an alternative way of reporting significant results.

5.6.2. Results for RQ3-1

Table 5.22 shows the results of related sample *t*-tests on the linguistic variables.

Table 5.22
Results of t-tests (N = 65)

Category	Variables	*M*		*SD*		*t*	*df*	*p*	Cohen's *d*
		A	B	A	B				
Range	D value	32.26	31.04	8.50	8.91	1.27	64	.210	
	AS-unit length	8.66	8.68	1.42	1.51	-0.10	64	.924	
	Subordinate clauses per AS-unit	.17	.24	.14	.15	-3.24	64	.002 **	-.48
Accuracy	% of error-free clauses	52.37	58.66	21.07	21.64	-2.81	64	.006 **	-.76
	Errors per AS-unit	.71	.66	.35	.40	1.26	64	.214	
	Errors per 100 words	8.56	7.89	4.71	4.91	-2.59	64	.010 *	.14
Fluency	Speech rate	83.89	87.17	26.82	26.49	1.54	64	.129	
Coherence	Positive additive connectives	91.54	90.70	23.65	25.25	0.29	64	.775	
Sustained Monologue	No. of main ideas	8.20	7.37	.85	.96	5.23	64	.000 ***	.18
	No. of minor ideas	4.49	4.69	1.74	1.78	-0.78	64	.435	

Note. *p<.05. **p<.01. ***p<.001.

Significant differences were found for subordinate clauses per AS-unit ($t(64) = -3.24, p < .002$), the percentage of error-free clauses ($t(64) = -2.81$, $p = .006$), the number of errors per 100 words ($t(64) = -2.59, p = .01$), and the number of main-idea units ($t(64) = 5.23$, $p = .000$). Effect sizes had a large effect on the percentage of error-free clauses (-.76), a medium effect on subordinate clauses per AS-unit (-.48), and a very small effect on the number of errors per 100 words (.14) and the number of main-idea units (.18). It is striking that the three accuracy variables produced completely different results: a significant difference with a large effect on the percentage of error-free clauses, no significant difference on the number of errors per AS-unit, and a significant difference with a small effect on the number of errors per 100 words. This discrepancy is further reflected upon in the next chapter.

5.6.3. Results for RQ3-2

The data for the 65 Japanese candidates were divided into different proficiency levels for RQ3-2, eventually excluding A1+ level which had only one candidate. The remaining 64 candidates were at A2/A2+ (n = 25), B1/B1+ (n = 31) or B2/B2+ level (n = 8). Related sample t-tests were applied to A2/A2+ and B1/B1+ levels, and paired-sample Wilcoxon signed-rank tests were used for B2/B2+ level and the NS data. As the number of errors per AS-unit was found to be affected by task order, it was excluded from this analysis. The results for the t-tests and Wilxocon tests are shown in Tables 5.23 and 5.24, respectively.

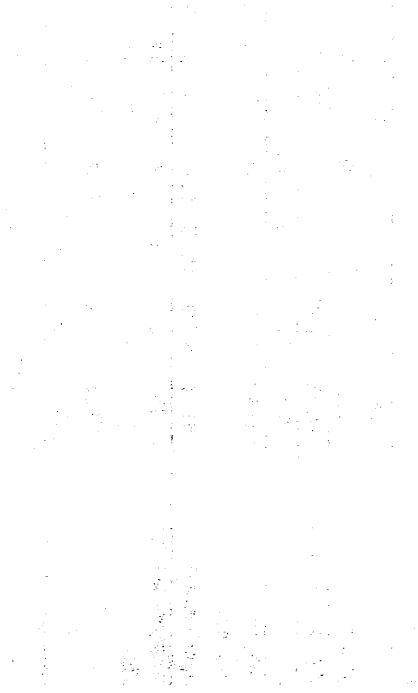


Table 5.23

Results of t-tests for A2/A2+ and B1/B1+ Levels

Category	Variables	A2/A2+ (n = 25) Mean		SD		t	df	p	Cohen's d	B1/B1+ (n = 31) Mean		SD		t	df	p	Cohen's d
		A	B	A	B					A	B	A	B				
Range	D value	28.01	25.98	8.03	6.74	1.28	24	.214		33.76	32.46	6.88	7.49	.98	30	.334	
	AS-unit length	8.03	8.19	1.13	1.21	-.68	24	.504		8.96	8.84	1.06	1.39	.39	30	.697	
	Subordinate clauses per AS-unit	.12	.22	.10	.16	-3.39	24	.002 **	-.75	.20	.24	.16	.14	-.93	30	.362	
Accuracy	% of error-free clauses	41.91	43.98	18.62	15.94	-.53	24	.604		56.10	66.47	18.12	19.35	-2.84	30	.008 **	-1.59
	Errors per AS-unit	.92	.93	.33	.34	-.13	24	.895		.63	.51	.29	.32	2.17	30	.038 *	.39
	Errors per 100 words	11.64	11.46	4.33	4.12	.21	24	.832		7.09	5.93	3.29	3.69	1.95	30	.060	.33
Fluency	Speech rate	65.40	67.93	14.24	15.47	-1.44	24	.163		90.53	96.01	21.02	21.12	-3.20	30	.003 **	-.26
Coherence	Positive additive connectives	94.18	97.88	26.21	23.43	-.89	24	.383		94.96	89.48	20.56	21.92	1.27	30	.212	
Sustained	No. of main ideas	7.72	7.04	.94	1.06	2.32	24	.029 *	.68	8.45	7.71	.62	.74	4.00	30	.000 ***	.16
Monologue	No. of minor ideas	3.88	4.48	1.72	1.73	-1.66	24	.109		4.77	4.52	1.71	1.67	.63	30	.536	

Note. *p < .05. **p < .01. ***p < .001.

Table 5.24

Results of Wilcoxon tests for B2/B2+ Level and English Native Speakers (NS)

Category	Variable	B2/B2+ (n = 8) Median		Range		Z	p	NS (N = 11) Median		Range		Z	p
		A	B	A	B			A	B	A	B		
Range	D value	36.77	43.16	32.74	23.57	-.42	.674	43.55	44.80	20.93	31.73	-.27	.790
	AS-unit length	9.35	9.28	6.89	6.46	-.42	.674	8.85	10.22	4.57	5.83	-2.13	.033 *
	Subordinate clauses per AS-unit	.18	.26	.53	.49	-1.68	.093	.25	.44	.65	.78	-2.22	.026 *
Accuracy	% of error-free clauses	79.47	81.80	60.73	33.89	-.42	.674	100.00	100.00	7.14	6.25	-.31	.753
	Errors per 100 words	2.68	2.72	8.51	4.98	-.42	.674	.00	.00	1.01	.65	-1.36	.173
Fluency	Speech rate	125.34	123.63	66.45	51.99	-.70	.484	196.80	174.78	153.93	108.03	-.71	.477
Coherence	Positive additive connectives	71.03	65.29	48.52	78.65	-.42	.674	67.80	64.29	72.86	62.88	-1.07	.286
Sustained	No. of main ideas	9	7	1	3	-2.56	.011 *	9	9	0	0	.00	1.000
Monologue	No. of minor ideas	5	6	5	6	-1.84	.066	4	5	6	8	-.47	.642

Note. *p < .05.

At A2/A2+ level, two variables yielded a statistically significant difference between Tasks A and B: subordinate clauses per AS-unit ($t(24) = -3.39$, $p = .002$), and the number of main ideas ($t(24) = -1.66$, $p = .029$), with large effects of -.75 and .68, respectively. At B1/B1+ level, more significant differences were found than for any other level (including NS), which suggests that the two tasks may have affected performance differently at different levels. The effect sizes were mostly small at this level: speech rate (-.26), the number of errors per AS-unit (.39) and the number of main idea units (.16). Only the percentage of error-free clauses presented a very large effect of -1.58, while other two accuracy variables yieled only a medium size effect (the number of errors per 100 words produced a near-significant result ($p = .06$) with a medium effect of .33). Thus, again, the results for the three accuracy variables were discrepant, which emphasises the need to investigate them further in Chapter 6, as was mentioned in Sections 5.6.1.1 and 5.6.2.

Considering the very small effect of size on the number of main idea units at B1/B1+ level, the differences between Task A and B for these variables are ignorable. As for the other variables, the results indicate that Task B produced more fluent and accurate performance. The average speech rate counted a larger number of syllables per minute in Task B (96.01), compared to 90.53 syllables in Task A. The percentage of error-free clauses (66.47%), the number of errors per AS-unit (.63 errors) and the number of errors per 100 words (5.93 errors) all suggest higher accuracy in Task B than in Task A (56.10%, .51 errors and 7.09 errors, respectively).

At B2/B2+ level, the number of main idea units was the only variable with a significant difference with fewer units in Task B. In fact, this result was common to the other two levels discussed above, with a large effect of .68 at A2/A2+ and a very small effect of .16 at B1/B1+. It is somewhat perplexing that Task B elicited less detailed

narratives regardless of candidates' proficiency levels, and that a significant difference was found at B2/B2+ with such a small sample size (n = 8). This issue deserves close examination of the transcripts so as to verify that these results were caused by the way in which the idea units were established, which will be conducted in the next chapter.

The NS baseline data suggested significantly higher syntactic complexity in Task B, as measured by AS-unit length (Z = -2.13, p = .033) and subordinate clauses per AS-unit (Z = -2.22, p = .026). The higher value for subordinate clauses per AS-unit was observed at all other levels, although not necessarily with significant differences (only A2/A2+ level had a significant difference with a large effect of -.75).

5.7. Validity of Linguistic Variables

5.7.1. Data

5.7.1.1. Obtaining Fair Averages for the Ratings

Research Question 4 was established as: "How do the linguistic variables correlate with the ratings of spoken narrative performance in the corresponding rating category?", aiming to conduct a validation study of the linguistic variables used in RQs 3-1 and 3-2. For both tasks by each of the 65 Japanese candidates, fair averages of the ratings from the seven raters were obtained for each rating category using FACETS: Range, Accuracy, Fluency, Coherence and Sustained Monologue. After checking the distributions of all the linguistic variables and ratings, the linguistic variables were correlated with the ratings of the corresponding category using Pearson's correlation coefficients.

In order to generate Rasch-scale fair averages, five separate models were created using FACETS, once for each rating category. This was because it was possible that raters behaved differently when rating different categories, i.e. a rater could be

lenient when rating Fluency, severe when rating Accuracy, etc. Therefore, by modelling the ratings for each rating category separately, it was believed that the fair averages obtained would be more accurate than if they were modelled together.

5.7.1.2. Descriptive Statistics

The central tendency and dispersion of the ratings data were calculated and checked for normality. The values for skewness and kurtosis of the ratings data, as indicated in Table 5.25 below, were all within the acceptable range of -2 to 2 (Bachman, 2004: 74). Descriptive statistics for the linguistic variables had already been calculated (they are presented in Table 5.20 in Section 5.6.1.2 (RQ3-1)), and considered suitable for parametric testing.

Table 5.25
Descriptive Statistics for Averaged Ratings

Category	Mean		SD		Skewness		Kurtosis	
	A	B	A	B	A	B	A	B
Range	5.64	5.85	1.17	1.12	.22	.26	-.20	-.40
Accuracy	5.23	5.44	1.20	1.16	.43	.36	-.34	-.46
Fluency	5.66	5.93	1.43	1.31	.13	.06	-.54	-.17
Coherence	5.35	5.87	1.23	1.07	.11	.40	-.31	-.29
Sustained Monologue	5.60	5.91	1.18	1.20	.07	.31	-.10	-.19

5.7.2. Results

5.7.2.1. Range

There were significant correlations between the fair average Range ratings and the corresponding linguistics variables, as shown in Table 5.26 below.

Table 5.26

Correlations between Linguistic Variables and Range Ratings

	Task A	Task B
D value	.470 **	.469 **
AS-unit length	.509 **	.282 *
Subordinate clauses per AS-unit	.265 *	.120

Note. *$p < .05$. **$p < .01$.

All the linguistic variables demonstrated significant correlation with the averaged ratings for Task A. Among them, two variables correlated moderately highly: D value (r = .470) and AS-unit length (r = .509). This indicates that the candidates who were rated high in Range used more varied vocabulary with more words in the AS-units. A weak correlation was found with subordinate clauses per AS-unit (r = .265), which suggests that the candidates may have produced a larger amount of subordination in the narration. Therefore, in Task A, it is assumed that the linguistic variables in Range that were used in the main study captured the characteristics of the higher-level performance which are represented by the averaged Range ratings.

Interestingly, for Task B, only two of the linguistic variables correlated significantly with the averaged ratings of Range: a moderately high correlation with D value (r = .469) and a weak correlation with AS-unit length (r = .282). Similar to Task A, these two variables suggest a tendency for highly-rated candidates to use more varied

vocabulary with more words in the AS-units. However, subordinate clauses per AS-unit showed almost no correlation at .120. Low correlation can be caused by variation among the candidates; however, this was not the case with this variable because the standard deviations were very similar for Tasks A and B.[29] This suggests that the highly-rated candidates did not necessarily produce more subordination in Task B than lower-rated candidates. In summary, considering the significance as well as the strength of correlation, D value appears to be the most reflective of Range ratings for both tasks.

5.7.2.2. Accuracy

Significant correlations were found between the fair average Accuracy ratings and the corresponding linguistic variables for percentages of error-free clauses and errors per 100 words, as summarised below in Table 5.27.

Table 5.27

Correlations between Linguistic Variables and Accuracy Ratings

	Task A	Task B
% of error-free clauses	.644 **	.683 **
Errors per AS-unit	-.652 **	-.687 **
Errors per 100 words	-.723 **	.-731 **

Note. $^{**}p < .01$.

While both variables correlated highly with the ratings, the number of errors per 100 words correlated even more highly for both tasks (-.723 and -.731, respectively). Thus, the number of errors per 100 words is a more sensitive variable which better reflects the human raters' judgement of accuracy.

[29] The standard deviations of Task A and B were .14 and .15, respectively, for the number of subordinate clauses per AS-unit.

5.7.2.3. Fluency

There was a very strong and significant correlation between the fair average Fluency ratings and the corresponding linguistic variable of speech rate in both tasks, as shown below in Table 5.28.

Table 5.28

Correlations between Linguistic Variables Fluency Ratings

	Task A	Task B
Speech rate	806 **	795 **

Note. **$p < .01$.

5.7.2.4. Coherence

None of the variables of coherence correlated significantly or highly with the fair average Coherence ratings, as indicated by Table 5.29, below.

Table 5.29

Correlations between Linguistic Variables Coherence Ratings

	Task A	Task B
Positive additive connectives	-.193	-.122

There was almost no correlation between the ratings and the variable of positive additive connectives; however, these results do not necessarily mean that it is an invalid measure. This will be discussed further in the next chapter.

5.7.2.5. Sustained Monologue

There were significant weak correlations between the fair average ratings in Sustained Monologue and the number of idea units, except for the number of main ideas in Task B, as Table 5.30 below shows.

Table 5.30

Correlations between Idea Units and Sustained Monologue Ratings

	Task A	Task B
No. of main ideas	.501 **	.172
No. of minor ideas	.345 **	.375 **

Note. **$p < .01$.

The number of main idea units in Task B correlated very weakly. This is because higher-level candidates did not necessarily include more main idea units in their narratives, as indicated in RQ3-2, which will be explored further in the next chapter.

5.8. Summary

As there is much to be explored and discussed in the next chapter, in order to reach a full understanding of the data, this section only briefly summarises the findings thus far. Firstly, in response to RQ1, the difficulty of the two narrative tasks was investigated using MFRM software, FACETS, on the ratings data. It was revealed that Task A was significantly more difficult than Task B, both in terms of Considered Judgement (CJ) ratings and the ratings for Range, Accuracy, Fluency, Coherence and Sustained Monologue. Although the differences in logits were small, .40 (CJ) and .52 (other rating categories) between the two tasks, it was demonstrated that such small differences could lead to different CEFR base levels for part of the candidate population. These results should be discussed together with the findings of other research questions so as to consider why significant differences might have been found despite every effort to select and make the two tasks as parallel as possible.

The analysis of RQ2 explored the perceptions of different groups of

participants: candidates (RQ2-1 and 2-2), Japanese teachers who had taught the majority of the candidates (RQ2-3), and English native speakers (RQ2-4). No significant differences were found in the perceptions of the tasks by the candidates. However, the difficulty perceived by B2/B2+ candidates was approaching significance, which might imply some interaction between proficiency levels and perceived difficulty. Expert judgements by the Japanese teachers revealed counter-evidence against the parallelness of the two tasks in terms of relevant task complexity factors. Particularly surprising were the suggestions about the possible interference of lack of cultural knowledge about washing, which might relate to a lack of washing-related vocabulary. In addition, questions were raised about the clarity of pictures in Task B and the sufficiency of detail in the pictures in Task A. In RQ2-4, 7 out of 11 English native speakers perceived different levels of difficulty in the two tasks. The issues of clarity and sufficiency of detail in the pictures were again mentioned.

RQ3 aimed to investigate the linguistic performances of the candidates in the two tasks. With 65 candidates (RQ3-1), the two tasks appeared to be parallel, except for the variables of subordinate clauses per AS-unit, the percentage of error-free clauses, errors per 100 words, and the number of main ideas. All of these variables need to be further discussed in the next chapter, in view of the results for RQ3-2 and RQ4. For RQ3-2 it was shown that significant differences were found at different levels of proficiency, with B1/B1+ candidates yielding the highest number of variables with significant differences. The results for RQ3-2 also need to be explored in light of the validation study for linguistic variables in RQ4.

For RQ4, a validation study of the linguistic variables was conducted to find out which variables were in line with the human ratings. Speech rate (fluency), the number of errors per 100 words (accuracy), D value and AS-unit length (complexity) were shown to be in accordance with the human ratings, and thus validated.

Nevertheless, this does not necessarily suggest that the other variables were invalid. There might have been influences from the rating scales, as mentioned during rater training as a potential issue (reported in Chapter 4). The next chapter will develop the discussion of these findings and aim to integrate them in order to achieve a full understanding of the main data as well as the issues found in this thesis.

Chapter 6: Discussion

This chapter aims to integrate the findings presented in Chapter 5 and discuss them in the light of relevant previous studies in order to explore the contributions of this thesis. Sections are organised according to the research questions, but discussion of RQ4 (validation of the variables) occurs before RQ3 (analysis of linguistic performances) because interpretation of the differences in candidates' linguistic performances on the two tasks relies on an accurate understanding of the variables. Accordingly, this chapter is structured as follows. First, Section 6.1 interprets the differences in difficulty of the two tasks calculated by MFRM analysis in the main study (RQ1), followed by discussion of the perceptions of the two tasks by the candidates, Japanese teachers, and NS (RQ2), in which some important issues regarding task complexity are raised. Then, Section 6.3 examines, both quantitatively and qualitatively, the validity of the variables (RQ4) used for analysing the linguistic performances on the two tasks. Finally, the chapter reverts to discussing task parallelness in terms of linguistic performance.

6.1. Task Difficulty according to MFRM Analysis

The results for RQ1 indicate that the difficulty of Tasks A and B was significantly different according to FACETS, both in the overall Considered Judgment (CJ) ratings and in the ratings of the five rating categories. Although the difference in task difficulty is small (.40 and .52, respectively), its effect was demonstrated as possibly being crucial for part of the candidate population. This section further explores how these differences might be interpreted.

In the previous study to have examined the equivalence of monologic tasks, Weir and Wu (2006) found a .74 logit difference between the two description tasks on Forms 2 and 3 of the test under investigation. Despite this large significant difference in

logit values, they concluded that these two tasks were actually equivalent, arguing that the fair average values were identical; there was only a .03 difference on a 5-point rating scale. In their investigation, Weir and Wu cite a comment by Linacre, from their personal correspondence, whereby the large logit difference found for a very small difference in raw score (i.e. fair average) probably resulted from the very little rater disagreement, and that these differences are actually very trivial (Weir & Wu, 2006: 186).

In Research Question 1 of my main study, the fair average values of CJ ratings for Tasks A and B were 5.67 and 5.91, respectively, on a 10-point rating scale (i.e. Below A1 to C1). Similarly, those for the five rating criteria were 2.59 and 2.75 on a 5-point rating scale (i.e. Below A2, A2/A2+, B1/B1+, B2/B2+, C1). Thus, there was no large discrepancy between the raw scores and logit values as Weir and Wu (2006) found in their study. Regarding the difference of .24 on a 10-point scale for the CJ ratings, Linacre (2011, personal communication) comments:

> On a golf-course, a hole with 0.24 average strokes more than another hole is considered noticeably more difficult. In your situation, I don't know. 0.1 score-points difference would definitely be "the same". 0.5 score-points difference would definitely be "different". 0.24 is in the gray-area where detailed knowledge of the situation is needed. (Linacre, 2011, personal communication)

According to this comment, the difference of .24 is not 'small enough' to be ignored, but not 'large enough' to be absolutely different. This opens up two possibilities for the interpretation to be labelled either as 'ignorable' or 'different'. I would, however, argue that it is rather a large difference, considering that it emerged despite every effort to select and make the two tasks as parallel as possible, as demonstrated in Pilot Study 5.

The two tasks were confirmed to be very similar in terms of their storylines, settings, numbers of events, relationships between characters, and amount of subordination in *a priori* analysis. This reveals that even such careful task design or selection may be insufficient to ensure tasks are parallel.

It is not the intention of this thesis to assert further that significant differences between (or among) tasks, however small, should be rectified, since such decisions and interpretations are, ultimately, to be made by test developers according to their particular testing situations. As mentioned in the previous chapter, one might argue that, in a language testing situation, there will also be other tasks to collect evidence about the different aspects of candidates' speaking ability, thus different (adjacent) ratings for a particular task type may be trivial. Therefore, the rest of this chapter does not pursue this line of discussion, but focuses rather on reflecting why such differences might have been found. In other words, the following sections explore why Task A might have been more difficult despite the careful selection of two tasks in the hope of exploring theoretical and methodological implications for the design of parallel tasks of this type.

6.2. Perceived Difficulty by Candidates and Cognitive Complexity of the Tasks

For RQ2-1, no questions in Robinson's (2001) questionnaire yielded significant differences. Thus, the two tasks were deemed to be parallel in terms of the substantive aspect of validity, which was measured according to perceived difficulty (Q1), nervousness (Q2), self-rating of performance (Q3), interest (Q4), and enjoyment (Q5). However, interestingly, Question 1, which asked about the perceived difficulty of the tasks, showed Task A being perceived as slightly more difficult with near-significance ($p = .056$), although the difference was small and ignorable (.30 on a 9-point Likert scale, with an effect size of .188). RQ2-2 also revealed no significant results, but, with Q1, presented an interesting tendency for higher level candidates to

regard the difficulty of the two tasks as greater in difference than the lower level ones (i.e. A2/A2+). The difference in perceived difficulty of the two tasks was approaching significance at B2/B2+ level (p = .072) despite the small sample size. For B2/B2+ candidates, i.e. the eight candidates placed at the top of the FACETS ruler in Figure 5.1 in Chapter 5, both tasks may have been quite easy. Because the tasks were easy for them, they might have sensed slight differences in the cognitive complexity of the pictures in Tasks A and B, whereas other candidates, at lower levels, might not have managed to do this. In light of Levelt's (1993) model of speaking processes, candidates at higher proficiency levels can speedily execute processes in the Formulator and Articulator, consequently they may be subject to less interference in the processes of conceptualization, which might have enabled them to sense that Task A was slightly more difficult.

What the higher level candidates might have perceived as being different between Tasks A and B can be answered by the findings of RQs 2-3 and 2-4. In RQ2-3, the Japanese teachers raised doubts, concerning the task complexity factors based on the checklist used by Weir and Wu (2006), about the parallelness of the two tasks. The checklist design was based on Skehan's (1998) task-related factors of complexity (i.e. *cognitive complexity* and *code complexity*). This aimed to elicit expert judgements on the assumed familiarity of candidates with the characters, setting, events and required lexical and grammatical items and functions, as well as on the sufficiency of details drawn in the pictures for candidates. From the teachers' responses, it appears that Task A was more difficult for the candidates than Task B. Similarly, for RQ2-4, 6 out of 11 native speakers of English (NS) answered that Task A was more difficult in a brief subsequent interview and gave reasons why. What was commonly mentioned by the Japanese teachers and NS was the insufficiency of details in the pictures in Task A. Specifically, comments from both groups concerned the changes of setting and the

ghost-like figure with the balloon and a piece of clothing from the washing line. First, because the setting in Task A changed from a room to the garden, and then back to the room, there were some time gaps between adjacent pictures. As NS 8 precisely expressed it:

> It didn't seem to link, like this one [pointing at Picture 2] and this [pointing at Picture 3] didn't seem to link together, like it's kind of she's taking out the washing but now where's she gone and who were they, so you have to sort of make up who these people were and where she'd gone, so it was hard.

Similarly, as there was a change in setting between Picture 5 (where the children were drawing a face on the balloon and taking the piece of clothing down in the garden) and Picture 6 (the shocked mother in the room near the ghost-like figure at the window), there was a time gap as well as a lack of detail in the process of how the ghost-like figure was created. In addition, one of the Japanese teachers pointed out that how the window, room and garden were located would be hard to understand and to explain, which also relates to the change in setting. Therefore, such details had to be 'made up' or figured out by the candidates in Task A, whereas no change in setting occurs in Task B.

What was striking was the inseparability of cognitive complexity and code complexity, as Japanese Teacher 1 suggested in his responses. He described the role of the balloon seller and washing-related objects as being unfamiliar to candidates because of the lack of equivalents in Japanese culture. Moreover, he pointed to the lack of candidates' lexical knowledge of washing-related objects, as a result of such unfamiliarity. This indicates that lexical knowledge is not separate from cultural knowledge, and therefore code complexity cannot be examined accurately, independent of cognitive complexity factors. This is a very important implication for the discussion

of task complexity factors. As Figure 2.5 in Section 2.7.1.3 (Chapter 2) illustrates, Bachman (2002) regards Skehan's (1998) code complexity as task-inherent, and cognitive complexity as the interaction between task and candidate factors. However, the comments of Teacher 1 imply that even code complexity is not purely task-inherent, but interacts with candidate factors. Moreover, it is not separate from cognitive complexity either. Therefore, the concepts of Robinson's (2001) task complexity, which encompass Skehan's code complexity and cognitive complexity, may not be as "fixed and invariant" (Robinson, 2001: 29) as expected, but need rather to be re-conceptualised as variant and complex interactions between task and candidate factors.

6.3. Linguistic Performances on the Two Tasks

Having discussed the analysis results for RQ1 and RQ2, this chapter now turns to exploring the possible effects of differences in cognitive complexity factors on candidates' linguistic performances. First, Section 6.3.1 discusses the effects of relevant task complexity factors, as predicted by Skehan (2009) and Robinson (1995), and then briefly reviews the results of analysis for RQ3-1 and RQ3-2 in order to make comparisons between predictions of the two theories and actual linguistic performances by the Japanese candidates and native speakers of English. As interpreting the findings of RQ3-1 and RQ3-2 largely depends on the construct of linguistic variables used in the investigation, Section 6.3.2 discusses the answers to RQ4, "How do the linguistic variables correlate with the ratings of spoken narrative performance in the corresponding rating categories?". Moreover, in Section 6.3.3, some transcripts are analysed qualitatively in order to investigate further the construct of some linguistic variables. Furthermore, Section 6.3.4 discusses the notion of 'task induction', which arises from the discussion in Section 6.3.3. Finally, Section 6.3.5 re-examines the

results for RQ3-1 and RQ3-2 in the light of all the discussions, leading to the summary in Section 6.4.

6.3.1. Discussing Linguistic Performances in the Light of Theories of Task Complexity

From the results for RQs 2-3 and 2-4 it is apparent that the change in settings in Task A has resulted in more cognitive demands than Task B. This could have resulted in lower average ratings for Task A, which might have eventually led to the difference in task difficulty calculated by MFRM analysis. It should be noted here that these findings were unexpected before the main study; they appear to relate to the task complexity factors that were initially excluded as being irrelevant to this thesis (see Section 2.7.2). Therefore, the excluded factors of task complexity are reintroduced in this section so as to shed light on how the change in setting in Task A might affect linguistic performances, according to predictions by the hypotheses of Skehan (1998) and Robinson (2001).

The effects of the changes in setting in Task A on linguistic performances can be explained with reference to Skehan's (2009) study. Considering that such changes may have required candidates to 'make up' or fill in some time gaps between the pictures and seek ways to make the story coherent, they may be related to *information organisation*, which is included in Skehan's *cognitive complexity* factors. The change in setting might have made Task A slightly less structured because of the resultant time gaps in the pictures. Skehan (2009: 516), reviewing his previous studies, mentions that the 'unstructured' spoken narrative task used in Foster and Skehan (1996), which has no obvious storyline, and therefore requires candidates to develop one, elicits more complex performance in terms of the number of clauses per C-unit. Skehan claims that

developing a storyline necessitates the manipulation and organisation of information, and explains this phenomenon as follows:

> Information manipulation and integration seems to require more extensive Conceptualizer use, which is reflected in a more complex pre-verbal message, and this, in turn, leads to the need to formulate more complex language. (Skehan, 2009: 517)

If Task A is regarded as less structured, then similar influences on linguistic performance might be observed. In the light of Skehan's (1998) Trade-off Hypothesis, if complexity is prioritised and increased, then fluency and accuracy are expected to decrease, since L2 speakers might not be able to allocate sufficient attention to all areas of performance. Quite similarly, Robinson (2003: 74) argued that more complex tasks along *resource-dispersing* dimensions, for example, increasing demands for attentional resource by making learners tell a story as well as figure out the sequence (this condition is called as +/- single task (or task structure)), would lead to negative effects on all the aspects of fluency, accuracy, and complexity.

However, Robinson (1995) suggested different predictions of effects on performance by more complex tasks along *resource-directing* dimensions. With regard to organising information, Robinson (1995) also discussed the effects of increased demands in conceptualization in his study which investigated the influences of task administration conditions. These conditions were expected to represent a *resource-directing* factor in Robinson's task complexity: *There-and-Then* versus *Here-and-Now* conditions for performing spoken narrative tasks. The *There-and-Then* condition refers to tasks where candidates are asked to look at pictures, plan their speech, and narrate a story in the past tense without having the pictures in front of them. In the absence of pictures to base the story on, this condition is thought to be more

cognitively demanding (compared to the *Here-and-Now* condition where candidates narrate in the present tense with pictures in front of them), because candidates have to recall events, organise them, and establish coherent connections between the events in their narratives. Robinson (1995: 121) attributes the reduced fluency under the There-and-Then condition to the cognitive demands of organising and connecting events in the conceptualization stage. The absence of pictures under the There-and-Then condition is somewhat similar to the lack of detail in the pictures in Task A in the main study, in that it demands establishing appropriate connections between the pictures largely on the narrator's part. In contrast, Task B includes no changes of setting that induce major gaps between the pictures, which might have enabled candidates to narrate a story without consuming as many cognitive resources in the Conceptualizer stage, compared with Task A.

Nevertheless, despite the partial similarity to the There-and-Then condition, it is not yet known whether the lack/presence of details and connections drawn in the pictures (in Task A) belongs to resource-directing factors of task complexity. This is because the conditions of There-and-Then and Here-and-Now are also thought to draw on memory demands (Robinson, 1995: 107) which are irrelevant in this thesis. It appears reasonable to assume that the changes in setting make up some of the resource-directing factors, as such changes in Task A may have required slightly more complex event construal (Robinson & Gilabert, 2007: 166) than did Task B. On this assumption, Task A should elicit linguistic performances with less fluency, as a There-and-Then condition would do, but with increased accuracy and complexity. Robinson and Gilabert (2007) explained this prediction as follows:

> Following arguments by Givon (1985; 1995; cf. Sato 1988, 1990) that structural complexity tends to accompany functional complexity in discourse,

> and that demanding, formal communicative tasks and contexts elicit a syntactic mode of production (characterized by greater use of morphology, greater syntactic subordination, and a higher noun to verb ratio) in contrast to a simpler pragmatic mode, the Cognition Hypothesis predicts greater accuracy and complexity using such general measures of production, complex versus simpler tasks along all resource-directing dimensions of tasks. (Robinson & Gilabert, 2007: 166)

Having discussed the predictions of Skehan (1998) and Robinson (1995, 2001, 2003), this section now explores the answer to RQ3-1, "Are the narrative performances on the two spoken narrative tasks the same in terms of the linguistic performance variables?". The analysis of linguistic performances by 65 Japanese candidates for RQ3-1 revealed no significant differences between the linguistic performances on the two tasks in terms of D value, AS-unit length, and speech rate. Significant differences were found in subordinate clauses per AS-unit (lower for Task A), percentage of error-free clauses (lower for Task A), number of errors per 100 words (larger for Task A), and number of main idea units (larger for Task A). Therefore, the possible effects of Task A being more complex on fluency, accuracy and complexity of linguistic performances can be summarised as follows:

- No change in fluency (speech rate);
- Decreased accuracy (% of error-free clauses; errors per 100 words);
- Decreased complexity (subordinate clauses per AS-unit).

In light of the predictions by Skehan (1998) and Robinson (1995; 2001; 2007) concerning the influence of the change in setting in Task A, the results of the analysis

for RQ3-1 are summarised below, in Table 6.1.

Table 6.1

Predicted and Actual Changes in Linguistic Performance for Task A (RQ3-1)

	Fluency	Accuracy	Complexity
Skehan	↓	↓	↑
Robinson (*resource-directing*)	↓	↑	↑
Robinson (resource-dispersing)	↓	↓	↓
Results (N = 65)	→	↓	↓

The results do not fully support any of the predictions. Fluency (i.e. speech rate) actually decreased in Task A (M = 83.89, SD = 26.82; Task B: M = 87.17, SD = 26.49), though the difference was not significant. It could be argued that the predictions of both Skehan and Robinson presume greater differences in task complexity between the tasks used in the experiments. As mentioned in the previous section, Skehan's argument (2009) about the effect of the organisation of information stems from comparison between an 'unstructured' task which does not have any obvious storyline and a 'structured' task with a clear temporal sequence. Compared to the difference between the 'structured' and 'unstructured' tasks, the difference between Tasks A and B might have been much smaller. Likewise, the difference between Tasks A and B may not have required as much attentional demand or memory demand as the +/- single task or There-and-Then conditions of Robinson (1995).

Unlike fluency, the areas of accuracy and syntactic complexity (in terms of the amount of subordination) revealed significant differences. Decreased accuracy is in line with the predictions of Skehan, but syntactic complexity also decreased in opposition to his predictions. Decreased complexity is also the opposite of the predictions of

Robinson, who predicts both increased complexity and accuracy for more cognitively difficult tasks. However, as the following paragraphs will reveal, it raises a serious question about whether increased task complexity can be attributed to the increased syntactic complexity of linguistic performance. In addition, it is important to remember that the three accuracy variables yielded different results (i.e. a significant difference with a large effect on the percentage of error-free clauses; no significant difference in errors per AS-unit; a significant difference with a very small effect on errors per 100 words), which necessitate further investigation into how such discrepancies were found between the three variables. After reviewing the results of the analysis for RQ3-2 in the next paragraph, Section 6.3.2 presents the validation studies for the linguistic variables, with correlations, and Section 6.3.3 examines some of the transcripts in order to investigate further the linguistic variables of accuracy and syntactic complexity.

Similar to the findings for RQ3-1, it is revealed that neither of the predictions by Skehan and Robinson is supported by the analysis of linguistic performances at different levels of proficiency (RQ3-2). Table 6.2 summarises the results.

Table 6.2

Predicted and Actual Changes in Linguistic Performance on Task A (RQ3-2)

	Fluency	Accuracy	Complexity
Skehan	↓	↓	↑
Robinson (*resource-directing*)	↓	↑	↑
Robinson (*resource-dispersing*)	↓	↓	↓
Results (A2/A2+)	→	→	↓
Results (B1/B1+)	↓	↓	→
Results (B2/B2+)	→	→	→
Results (NS)	→	→	↓

At A2/A2+ level, candidates' linguistic performances on the two tasks did not yield any significant differences in fluency or accuracy. As was the case with the 65 candidates, a significant difference was found in syntactic complexity (i.e. subordinate clauses per AS-unit). At B1/B1+ level, fluency and accuracy (measured by two variables) were significantly different. Complexity did not differ; however, subordinate clauses per AS-unit also displayed a decrease on Task A (M = .20, SD = .16; Task B: M = .24, SD = .14). Again, the accuracy variables revealed discrepancies at this level; the percentage of error-free clauses and errors per AS-unit were significantly different, but not the errors per 100 words, which again emphasises the necessity to investigate these variables further. At B2/B2+ level, none of the three areas of linguistic performance yielded significant differences. Finally, whilst NS performances did not differ significantly in fluency or accuracy, they did for both variables of syntactic complexity (i.e. AS-unit length and subordinate clauses per AS-unit). It is also worth noting that a significant decrease in syntactic complexity (i.e. less subordination) on Task A was observed at NS level as well as at A2/A2+ level. Considering the differences in the available linguistic resources and processing speed between NS and the candidates at A2/A2+ level, it is very interesting that the two groups displayed similar profiles of linguistic performance. This indicates the possibility of 'task induction' for Task B, not task complexity, that induced more subordination regardless of levels of proficiency. This issue is revisited in Section 6.3.3.

In brief, the candidates' linguistic performances at different levels of proficiency were not in line with the two theories of task complexity from Skehan (2009) and Robinson (1995). Especially, no increased accuracy or complexity, as predicted by Robinson (1995), was observed. However, before discussing further the possible effects of differences in task complexity on linguistic performance, it is crucial that we examine the construct and validity of the linguistic variables, as appropriate

interpretation and implications are premised on accurate understanding of what the variables actually measure. For this reason, the next section (Section 6.3.2) discusses the validity of the variables based on the results of the correlation studies for RQ4, "How do the linguistic variables correlate with the ratings of spoken narrative performance in the corresponding rating category?". The section presents subsections that are allocated to each of the rating categories: Fluency, Accuracy, Range, Coherence and Sustained Monologue. Of the five categories, Fluency, Accuracy and Range are of particular relevance to the discussion of task complexity and linguistic performance (in the areas of fluency, accuracy and complexity) that have been noted so far.

6.3.2. Validation of Linguistic Variables

6.3.2.1. Fluency

Speech rate yielded very high correlation with human ratings, around .80, which is in line with previous studies on variables of fluency (e.g. Kormos & Dénes, 2004; Fulcher, 1996). This suggests that the results of analysis for linguistic performance via this variable are credible. Speech rate is a variable concerned with speed (Skehan, 2009: 513); it does not directly represent breakdown fluency (i.e. pauses) or repair fluency (i.e. reformulation, replacements, repetition and false starts). Nevertheless, it successfully captured the characteristics of fluency represented in the Fluency rating scales, which include descriptors for breakdown fluency and repair fluency. These results strengthened the rationale for using this variable to examine spoken narrative performance, as speech rate is easy to calculate and does not involve measuring pauses which can be quite tedious, time-consuming and impractical.

6.3.2.2. Accuracy

Although the three accuracy variables all correlated highly with one another, the number of errors per 100 words correlated slightly higher than the other two, demonstrating its suitability to reflect better the human raters' judgement of accuracy of candidate performances. This supports the claim by Mehnert (1998) that this variable may be appropriate for candidates with relative lower proficiency, though for a slightly different reason than the one she gave. Mehnert prefers this variable because having 100 words as a denominator avoids the use of clausal units whose definitions can be problematic. However, as Section 6.3.3 will demonstrate with the transcripts, having clausal units as a denominator may, depending on whether errors are spread across clauses, produce completely different results. As the spread of errors might also depend on the proficiency levels of candidates, using clausal units as denominators could function well with a narrower range of proficiency levels.

Another researcher who argued for using an accuracy variable over others was Bygate (2001), who suggests that calculating the number of errors per T-unit might be appropriate because it does not obscure the actual occurrences of errors as counting error-free units does. However, as Section 6.3.3 will show, segmenting transcripts into clausal units, such as AS-units, can also change the results. In this thesis, errors per 100 words aligns best with the human ratings, and thus can be considered the most valid measure of accuracy in my dataset that consisted mostly of candidates at A2/A2+ and B1/B1+ levels. Nevertheless, as mentioned above, having clausal units as denominators could be appropriate and reflect human ratings for a narrower range of proficiency levels, as a narrower ability range may have less variation in the length of clausal units, thus not altering the distribution of the resulting values too much.

The issue of error gravity (Hughes & Lascaratou, 1982) is also worth noting. The fact that correlations were only moderately high may be attributed to the issue of

error gravity. Because counting the number of errors does not distinguish between global and local errors, such accuracy variables are not able to differentiate between candidates with local errors (e.g. article omission or subject-verb agreement) and those with global errors (e.g. incomprehensible lexical choice). The difference in the 'seriousness' of the errors may distinguish lower level and higher level candidates, although these general variables of accuracy can rank higher level candidates with a lot of local errors lower than ones with fewer number of global errors. Thus, it is assumed that coefficients in the range |.6| to |.7| might be as high the correlations can go, if general variables of accuracy are to be used, and the remaining variance could be explained by how serious each error might have been.

6.3.2.3. Range

The ratings for Range were correlated with complexity variables: D value (lexical variety), AS-unit length and subordinate clauses per AS-unit (syntactic complexity). Significant correlations ranged from |.259| to .509, which led to the conclusion that D value was reasonably reflective of the ratings for this category. It can be assumed that correlations were only weak or moderate because some of the descriptors in the Range rating scale did not quite correspond to what the variables were measuring. For example, the Range descriptor at B1 level states, "Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events" [my emphasis]. This descriptor entails an aspect of fluency and the use of strategies to which no linguistic variables corresponded.

Still, it is important to note that correlations for syntactic complexity variables were quite low on Task B compared to Task A. The coefficients on Task B were .282 and .120 on AS-unit length and subordinate clauses per AS-unit respectively, whereas

they were .509 and .265 on Task A. This strongly suggests that highly-rated candidates did not necessarily produce more subordination or more words on Task B than lower-rated candidates did. Again, this might be because of the task induction effect of Task B that elicited subordination across all levels (this is examined further in Section 6.3.3).

6.3.2.4. Coherence

The positive additive connectives, aimed at measuring cohesive devices of coordination, showed almost no correlation with the ratings for Coherence. Therefore, it can be argued that the rating scale for Coherence was unable to measure the coherence of spoken narrative performance. As noted in Chapter 3, most of the descriptors for Coherence in the CEFR grid mentioned cohesive devices such as *and, or, then,* and descriptors of coherence start to appear at B2 in the CEFR grid, phrased as "can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse [...]". The raters reported problems regarding similar descriptors across levels in that they could not distinguish better performances based on the Coherence rating scales (as reported in Chapter 4). This addresses an important question of whether it was because of a problem with the rating scale or the construct. If an absence of correlation between ratings and cohesive devices was due to a lack of descriptors for coherence and organisation in the CEFR, then it may be because such tasks, which require coherence and organisation (in addition to the use of cohesive devices), are not thought to be a representative task type that learners are likely to perform at B1 level and below. Alternatively, it might be due to an incorrect assumption for the construct of cohesion and coherence of a narrative. The main study focused on examining cohesion by the amount of coordination based on the argument by Carrell

(1982: 485) that coherent text tends to be cohesive, and on my assumption that the elicited performance would contain a lot of simple cohesive connectives such as *and* and *then*, especially at lower levels of proficiency. Although the results of analysis for RQ3-2 showed that the variable of positive additive connectives by Coh-Metrix generally decreased as the levels went up[30] (see Tables 5.23 and 5.24), the lack of correlation with the ratings may suggest that cohesion of a narrative does not depend on only connectives, and that other dimensions of cohesion need to be investigated such as reference (e.g. lexical repetition, anaphora). Further data with a larger number of candidates as well as more NS should be collected in order to examine this. Conducting such an investigation would lead to one of two decisions: one is to revise the rating scale so as to tailor it for differences in coherence in this task type. The other, if spoken narrative tasks were found not to be capable of eliciting a lot of varied cohesive (and coherence) devices, is to omit the Coherence rating scale from the grid.

6.3.2.5. Sustained Monologue

The number of main idea units in Task A correlated moderately highly with the Sustained Monologue ratings (r = .501), but the correlation coefficient was only .172 for Task B because higher level candidates did not necessarily include more main idea units in their narratives. Interestingly, the number of main idea units has displayed a consistent pattern across all candidate levels; Task A constantly elicited a larger number of main idea units. Moreover, it was one of the variables which yielded a significant difference in the results for RQ3-1 (with 65 candidates). Examining transcripts revealed the cause of this somewhat surprising result. There were nine main idea units in each

[30] The values of positive additive connectives were 94.18 (A) and 97.88 (B) at A2/A2+, 94.96 (A) and 89.48 (B) at B1/B1+, 71.03 (A) and 65.29 (B) at B2/B2+, and 67.80 (A) and 64.29 (B) at NS level.

task, as shown in Table 6.3, and they were defined as the ones used by all 11 native speakers of English, as explained in Chapter 3. The transcripts suggested that two main idea units in Task B were actually redundant and were often omitted by the Japanese candidates, namely Unit Numbers 7 and 8. These idea units express the mother's noticing that the baby has gone and been replaced by a strange ball; the exchange of the baby and the ball is already described by Unit Numbers 4 and 5. A sample transcript for an A2/A2+ level candidate (Candidate 61) on each task is provided below and shows that Unit Number 8 is omitted in Task B. Clauses are indicated by a forward slash, /, and Unit Numbers are expressed in brackets.

Table 6.3

Main Idea Units

Task A	Task B
1 A mother is washing some clothes	1 A mother is reading a book
2 She goes outside to hang the clothes out	2 Her baby is sleeping in a basket
3 Two children appear in the garden	3 She falls asleep
4 They find a man carrying balloons	4 Two children take the baby from the basket
5 They buy a balloon	5 They put the ball with a face in the basket instead
6 The girl draws a face on it	6 The mother wakes up
7 The boy takes down a washed shirt	7 She finds that her baby is not there
8 They take the shirt with a balloon face to the window	8 It is the ball with a face lying in the basket
9 The mother is very surprised	9 She is very surprised

Task A by Candidate 61

(1) one day the elderly woman was washing her clothes / she lived with black cat happily / that was sunny day / so she finished washing clothes / (2) and then she brought her clothes to outside / then she went back to her house / (4) the man who has balloon / appeared / (3) two children ran from the house / and their looked happily / (5) so the man presented one balloon to children / the two children were thinking to play with balloon / (6) one child write the face on the balloon / (7) and another child was preparing / makes a trick to the woman / so

children bring the face to the shirts / and one child get in the clothes / and caught a balloon / (8) and he appeared next to the woman's window / and the elderly woman looked it / (9) and was surprised / and the cat also surprised / and ran away from the room /

Task B by Candidate 61

in the room there are two people the baby and the woman / the woman was sitting on a chair / (1) and reading a book / (3) the woman got sleep / (2) and the baby was sleeping / the woman put the book on her chair / and she got sleep / and then one boy and one girl came into the room / and they were thinking something / were smiling / (4) the girl caught a baby from the baby chair / and the boy was point outside of the room / (5) then the boy got the ball like baby / and the boy exchanged of the baby / but the woman didn't find it / (6) after that the woman got up / (7) and find / that the baby was stolen from one boy and one girl / (9) and she was surprised /

As can be seen, omitting Unit Number 8 (or 7) does not actually interfere with the storyline, although these were used by all 11 native speakers of English. This suggests that the degree of 'necessity' of each idea unit may differ between the two tasks, leading to the discrepancy in the numbers of idea units used across the levels.

Thus, it has been found that the main idea units in Task B contained two idea units that conveyed the same content as the other idea units, and that higher level candidates did not include these 'redundant' idea units either. The number of minor idea units, however, correlated significantly with the ratings in Sustained Monologue. This is not surprising since the category of Sustained Monologue includes descriptors about how 'detailed' a performance is. Nevertheless, the correlation between minor idea units

and ratings was rather weak (.345 for Task A; .375 for Task B), and may therefore not have yielded credible results. In sum, the conclusion here is that the idea units, established based on NS performance, were not in line with the ratings in this category.

6.3.2.6. Summary

Thus far, Section 6.3.2 has discussed the results for RQ4 which present the correlation coefficients between human ratings and the corresponding linguistic variables. Speech rate, errors per 100 words (although the two other variables of accuracy demonstrated moderately high correlation coefficients, this variable correlated slightly higher), and D value are in line with the human ratings for both tasks, and can thus be considered valid measures of linguistic performance. In the next section, recurring questions about the discrepancies between the three accuracy variables and the possible effects of 'task induction' by Task B on syntactic complexity are investigated qualitatively using transcripts.

6.3.3. Construct of the Linguistic Variables

6.3.3.1. Accuracy

The decreased accuracy on Task A seems to be in line with Skehan's (2009) prediction; however, there was a discrepancy between the results shown for the three variables of accuracy although they should be tapping into the same construct. As noted in Table 5.19, the variable of errors per AS-unit was found to be affected by task order, but was included in further analysis because the size of the order effect was thought to be small. This variable did not yield any significant difference, whereas errors per 100 words showed a significant difference with a small size effect (Cohen's d = .14). Moreover, a significant difference was also found in the percentage of error-free clauses,

but with a large effect of -.76. In order to explore how such discrepancies between the three variables might have been caused, this section first discusses the errors per AS-unit and errors per 100 words, as they have the same numerator (i.e. number of errors) but different denominators. Then, the ensuing discussion contrasts the errors per 100 words and the percentage of error-free clauses.

If the two variables of accuracy, errors per AS-unit and errors per 100 words, yield different results, then it must have been due to the difference in the denominators. Transcripts were examined for this trait, and those of the same candidate (Candidate 65) on the two tasks are shown as examples, below, of how it could happen. AS-units are indicated by forward slash, and errors are marked by underscore (omission), underlining (inflection/mischoice) and strikethroughs (unnecessary parts).

Task A by Candidate 65

(0.6 errors per AS-unit; 6.52 errors per 100 words)

there is a housewife washing her clothes and her family__ / and beside her there is a black cat / and then after washing them she starts hanging them outside / then she gets inside / after maybe one hour or two hours the clothes got dry and here comes a man with some balloons in his hand / and the children from the house come out / and get _ balloon from the man / and they get some bad not so bad but some funny idea / and a girl starts painting some face on the balloon / a boy get_ ~~out~~ the cloth from the rope / and they makes a balloon and a clothes like a man / and from the window they show~~ed~~ this figure / and the housewife is very surprised / and even the black cat is afraid of them / and then run_ away /

Task B by Candidate 65

(0.67 errors per AS-unit; 6.06 errors per 100 words)

there is a woman reading some book on the chair in her room / and beside her or in front of her there is a baby sleeping in __ small basket / but during reading the mother fell asleep / and two children a boy and a girl come in the room / and they try to take the baby from the basket / and instead of the baby they bring a ball painted a face on it / and put it on the basket / and after few minutes the woman wakes up and see_ it / and ___ surprised because the baby has changed to a ball /

As can be seen from the length of the transcripts, the performance on Task A was longer and contained more words (138 words) than that on Task B (99 words). The number of errors was also larger on Task A (9 errors) than B (6 errors). This produced a larger number of errors per 100 words for Task A (6.52) than B (6.06). However, because the AS-units were slightly longer in Task B, each AS-unit contained a higher number of errors, which resulted in a larger number of errors per AS-unit for Task B (0.66). Therefore, Task A was more erroneous according to the errors per 100 words, whilst the errors per AS-unit are the opposite. This offers an important implication for the use of these two variables as segmenting the transcripts into clausal units of analysis can produce quite different results.

Now, let us turn to examining the third accuracy variable, the percentage of error-free clauses, in comparison with the number of errors per 100 words. These two accuracy variables showed significant differences for RQ3-1, both indicating the performances on Task B to be more accurate. However, surprisingly, the effect size, as measured by Cohen's d, was very small for errors per 100 words (.14), while the percentage of error-free clauses had a large effect (-.76). With such a small size effect, the difference found for errors per 100 words can be ignored. In contrast, the difference

between the tasks appeared to have had a strong effect on the percentage of error-free clauses. Thus questions arose as to how the two accuracy variables could have revealed such different results, and which variable better reflects the actual difference between the accuracy of the performances.

To answer the first question, the raw data and transcripts were revisited. It was found that the cause of the discrepancy in size effect between the two accuracy variables may have been, again, the difference in denominators (i.e. 100 words or clauses), because the denominators must have affected the extent to which each error mattered when calculating values. Even if a candidate made about the same number of errors per 100 words, the percentage of error-free clauses could be quite different. Let us take two narrative performances, by Candidates 15 and 16 on Task A, below, as examples. Clauses are indicated by forward slash, and errors are marked by underscore (omission), underlining (inflection/mischoice), and strikethroughs (unnecessary parts).

Candidate 15 (5.60 errors per 100 words; 66.67% of error-free clauses)

in a room a mother is washing clothes on the table / and under the table a cat is looking at her / number two she is hanging __ washed clothes to dry in the yard / number three a man is walking along a fence / and he sells balloons / and the two children __ running up / number four the children buy a balloon from him / number five a girl is drawing a face on the balloon / and a boy is standing on the box / and he is taking a clothes to do something / number six the first lady is surprised by children's ___ / a mother and a cat is surprised by __ monster-like man / he was made by the children / his face is the balloon / and his body is __ washed clothes /

Candidate 16 (5.42 errors per 100 words; 72.73% of error-free clauses)

there was a lady // who was washing her clothes in the kitchen / and in the kitchen there was a black cat too / after washing the clothes // she went out of the house / and she put them to the ropes / and she went back to the room / and then the man came ~~to~~ near the house / and in his hand he had balloons to sell / and now the boy and the girl bought a balloon from the man / and then an idea came up to them / the girl painted a face / it looks like a Humpty Dumpty on the face of the balloon / and the boy picked up a shirt / or maybe it's a sheet~~s~~ from the rope / and then they acted like a big man / and they stood in front of the window ~~from the~~ outside / and then because the lady was in the kitchen // and looking __ the big man from the window / she was so astonished / she was surprised / and the black cat run away /

Candidates 15 and 16 had about the same number of errors per 100 words (5.60 and 5.42 respectively), but the percentages of error-free clauses were different (66.67% and 72.73% respectively) depending on how the errors were spread across clauses. Whilst most of the errors that Candidate 15 made were spread out as one error per clause, those of Candidate 16 were clustered together to produce a count of two errors per clause, which resulted in a higher ratio of error-free clauses. Therefore, with the number of errors per 100 words, each error is taken into account when calculating a value, while the percentage of error-free clauses can have clustered errors in certain clauses, leaving others error-free, and hence produce quite a different value.

Thus far, it has been demonstrated how a difference in the spread of errors can result in quite different values from different accuracy variables. Still, the question remains as to which variable reflects the accuracy of performances better in the main study. As shown in the previous chapter, the results for RQ4 showed the highest correlation coefficients between the numbers of errors per 100 words and the averaged

Accuracy ratings by the raters. This fact indicates that errors per 100 words may be the best index of accuracy for linguistic performance in this thesis. In addition, a *t*-test on averaged Accuracy ratings for Task A ($M = 2.47$, $SD = .68$) and Task B ($M = 2.55$, $SD = .63$) revealed a significant difference ($t(64) = -2.614$, $p < .05$) with a very small effect of -.12. This is the same as the result for errors per 100 words produced for RQ3-1. Considering that the human ratings showed the same results as the *t*-test on errors per 100 words (i.e. a significant difference with a very small effect), as well as correlating the highest with errors per 100 words, it can be concluded that errors per 100 words is the best index for the effect of different tasks on accuracy in the main study.

6.3.3.2. Syntactic Complexity

The results of analysis for RQ3-1 show decreased complexity in terms of the number of subordinate clauses per AS-unit on Task A. This indicates that the linguistic performances of the candidates were syntactically more complex on Task B, the cognitively easier task. This is the opposite of the predictions by both Skehan (2009) and Robinson (2007), both of whom used a variable of syntactic complexity relating to the amount of subordination. As was cited in Section 6.3.1, Skehan (2009: 516) pointed to the increased complexity of the 'unstructured' task in terms of the average number of clauses per C-unit. Similarly, Robinson (1995) used clauses per T-unit for syntactic complexity when comparing the effects of There-and-Then and Here-and-Now conditions. Both C-units and T-units are very similar to AS-units.[31] Moreover, the larger the number of such clauses per unit there are, the more subordinate clauses are expected to be included. Thus, these variables are thought to be closely related to the amount of subordination, which the number of subordinate clauses per AS-unit intends

[31] See Section 2.7.3.2 for discussion.

to measure in this thesis. Robinson (1995) did not find significant differences in clauses per T-unit between the two task administration conditions in his study, the values were identical. It is, therefore, surprising that the cognitively more difficult Task A produced significantly less subordination in my study.

The transcripts from both tasks were examined in order to find out what types of subordination were elicited by each task. As a result, it was found that Task B elicited more subordination in terms of relative clauses and subordinate conjunctions such as *while, when* and *because*. The transcripts of two candidates from both tasks are shown below. Forward slashes indicate AS-unit boundaries and squared words indicate where subordination occurs.

Task A by Candidate 29 (Total of 4 subordinate clauses)

there is a woman / they always say [that] she's quite serious person / she likes doing homemaking jobs / and she's serious about everything / and she doesn't like kids / she doesn't like talking to people / she doesn't believe in anything [()] she doesn't see / and today she is washing clothes / and she was trying to dry the clothes / and then some kids came into her garden / and then they saw a man [who] was selling balloons / and then they got the idea to surprise her a little bit / and then they bought a balloon / and then one of the kid painted the face on the balloon / and then the other got one of her clothes / and then they thought [()] they could make it look like a ghost / and then they put the balloon and clothes together / and then they put the balloon monster outside of the window to surprise her / and then the woman was so surprised / and she was so shocked / and then she fell over to floor /

Task B by Candidate 29 (Total of 8 subordinate clauses)

there is a woman in the room with her baby / and she's reading a book / and the baby is fast asleep / and then she was so tired / and then while she was reading she fell asleep / and nobody was there in the room / but two kids came into the room / one was a girl and one is a boy / and the girl got interested in the baby / she seems to want to play with the baby / the boy was not sure about if it was okay / but she says it's okay the mother is sleeping / so she took a baby / and the boy was saying oh be careful not to wake up the mother / and well I've got the idea / maybe I can bring a ball to make it look like a baby / so that the mother won't find out / and then he came up with the ball / and he put the ball into the bucket that the baby was in / and the girl is quite happy / because she could hold the baby / and then they were playing with baby / while the mother still kept sleeping / and then a few minutes later the mother find out that there was no baby in the basket / and she was so surprised because she doesn't see the children holding her baby / children were also surprised that she woke up / that's it /

It can be seen that Task B elicited twice as much subordination in the case of Candidate 29. This candidate was the 'most able' student according to MFRM analysis (see Section 5.1.2.2, Figure 5.1), and it could be argued that being able to produce subordination relates to the level of proficiency. However, it turned out that the least able candidate in my data (Candidate 36) was also producing more subordination on Task B, as shown below.

Task A by Candidate 36 (Total of 1 subordinate clause)

a woman is washing clothes / a cat gaze her / she dries big coat / after she going to home a man comes to her house / he sells balloons / her children go home /

and they buy a balloon / a girl painted a human face on the balloon / and a boy open coat drying / after a while she shocked to something strange looking at window / a big big something strange / she very shocked / and cat also very shock / and run away /

Task B by Candidate 36 (Total of 4 subordinate clauses)

a mother was reading book / nearby basket her baby was sleeping / she read book / but after finish reading she slept too / then her children a boy and girl come to room / they want to treat / first a girl try to hide baby / next a boy put strange ball in the basket which baby slept / after their mother woke up she very shocked because in the basket there is no baby but very strange ball /

As can be seen, Task B elicited more subordination from Candidate 36. The transcripts of both candidates appear to suggest that producing more subordination on Task B is partly due to the mother's constant presence in the pictures, which requires occasional mentioning of the mother's state using *while* and *after.* It is also attributable to the plot, in which the baby is replaced by a ball in the basket, which might have necessitated the use of relative pronouns such as *that.*

This suggests the strong possibility that subordination will be elicited regardless of how 'cognitively difficult' a task is thought to be, which in turn raises serious doubt about measuring syntactic complexity of the resultant performance by means of the amount of subordination. This may threaten the fundamental assumption of the arguments about task complexity, i.e. that the more cognitively complex (in terms of information organisation (Skehan, 2009) or absence of pictures (Robinson, 1995)) that tasks are, the more syntactically complex the elicited performance will be. In this regard, task complexity might not be related to the resultant complexity of the linguistic

performance at all. It seems that, as will be discussed further in the next section, there needs to be a separate notion, other than concepts of task complexity, for the linguistic structures likely to be elicited by the pictures, which I have called 'task induction'.

6.3.4. Task Complexity or Task Induction?

The results of Wilcoxon tests on the linguistic variables from the NS data show no significant differences between the two tasks except for two variables of syntactic complexity, AS-unit length (M = 8.85(A), 10.22(B)) and subordinate clauses per AS-unit (M = .25(A), .44(B)), both of which indicate Task B to be more syntactically complex. It is probable that, because Task B elicited more subordination, it also resulted in greater AS-unit length. Interestingly, as discussed in Section 6.2, six out of 11 NS perceived Task A to be more difficult in terms of cognitive complexity factors such as the amount of thinking, figuring out, and 'making up' details to fill in the time gaps between the pictures originated by the change in setting. Thus, again, the cognitive complexity reported by the NS was not in line with the predictions of increased syntactic complexity by Robinson (1995) or Skehan (2009). Moreover, at A2/A2+ level, significantly more subordinate clauses were observed. It is very interesting that even at such a low proficiency level where "basic sentence patterns" (Council of Europe, 2001: 28) are most expected, subordinate clauses were used in narratives. Taken together with the fact that even the least able candidate (Candidate 36, as shown in Section 6.3.3.2), who was excluded from the analysis of RQ3-2 due to her Below-A2 proficiency, managed to produce more subordination on Task B. This might reinforce the possibility of subordinate clauses being elicited by the 'task induction' effect of Task B. The NS and candidates at A2/A2+ level revealed significant differences, and the other two levels between them displayed the same tendency (although without statistical

significance).

Task induction, as I define it, is different from task complexity. The two theories of task complexity, from Skehan (2009) and Robinson (1995), assume that more cognitively complex tasks will require candidates to use more complex language. Foster and Tavakoli (2009) concluded that this prediction was confirmed with the same tendency for increased syntactic complexity[32] being shown by candidates at different levels of proficiency and NS in their study. However, in this thesis, Task B, the less cognitively complex task, elicited more complex language in terms of the amount of subordination across all levels of proficiency and from the NS. Foster and Tavakoli (2009) found increased syntactic complexity on spoken narrative tasks with more events drawn in the background,[33] which elicited more subordinate clauses such as relative clauses, *if/when* clauses, *unless/although/in case* clauses (Foster & Tavakoli, 2009: 885). They assumed that more events in the background would make the tasks more complex, so their results were in line with Skehan's prediction of increased syntactic complexity. In the case of the two spoken narrative tasks used in this thesis, however, more subordination was elicited because Task B entails the constant presence of the mother (eliciting *while* and *when* clauses) and includes in the plot the replacing of the baby by a ball (eliciting relative clauses), which do not relate to what the Japanese teachers and the NS reported about the cognitive complexity of the tasks at all. In addition, the Japanese teachers answered that the candidates would be equally familiar with the vocabulary and grammatical structures needed to complete the two tasks, so the increased syntactic complexity on Task B did not seem to be related to the code complexity of the task either. Therefore, it can be concluded that more syntactically

---

[32] They used the variables of AS-unit length and clauses per AS-unit.

[33] They stated that such tasks had "dual storylines" in which there were other events drawn in addition to the actions of the main character(s).

complex language can be elicited regardless of task complexity (that entails code complexity and cognitive complexity), and that there needs to be a separate notion for it. This is what I have called 'task induction', referring to task characteristics which induce the use of certain syntactic structures but are not related to task complexity.

It is also worth noting that task induction might have pushed even lower-level candidates, such as A2 level learners, to produce more complex structures than those specified in the CEFR. From this finding, along with separating the notions of task complexity and task induction, it is also essential to reconsider the variables used for measuring syntactic complexity. If learners with lower levels of proficiency can be pushed to produce subordination by task induction, then the amount of subordination may not be a reliable indication of one's language proficiency being higher or more advanced, as has often been assumed in task-based research. A possible alternative variable of syntactic complexity has been suggested by Norris and Ortega (2009: 561), which is not affected by the amount of subordination: phrasal complexity. It refers to the length of clauses, and any increase in this variable can only come from complexified phrases resulting from pre- or post-modifications using adjectives, adverbs, prepositional phrases, or non-finite clauses. The main study in this thesis employed AS-unit length and subordinate clauses per AS-unit, based on the literature review, so as to relate the discussion to previous studies. However, since it has clearly been demonstrated that the amount of subordination can increase due to the task induction effect and that this variable may not differentiate learners at lower and higher proficiency levels, phrasal complexity might be a better variable to use in future research.

6.3.5. Task Parallelness in Terms of Linguistic Performance

Now that the validity and construct of the linguistic variables and the notion of task induction have been discussed, this chapter revisits the results for RQ3-1 and RQ3-2, which aimed to examine whether the two tasks were parallel in terms of the linguistic performances by 65 Japanese candidates (RQ3-1) and at different levels of proficiency (RQ3-2). The linguistic performances by the 65 Japanese candidates revealed significant differences in the variables of subordinate clauses per AS-unit, the percentage of error-free clauses, errors per 100 words, and the number of main idea units. From the discussion above it is apparent that idea units did not prove to be a valid measure, thus this variable is not discussed here. Regarding the other two aspects of performance (i.e. accuracy and syntactic complexity), Tasks A and B were not parallel in terms of the linguistic performances by the 65 Japanese candidates.

For RQ3-2, the linguistic performances at different levels of proficiency have been investigated, and the two tasks were found not to be parallel in NS performances in terms of syntactic complexity. This may have been due to the task induction effect of Task B which elicited more subordination (and therefore, longer AS-units). Likewise, the task induction effect may have elicited significantly more subordinate clauses at A2/A2+ level. Thus, at A2/A2+ level and NS level, the two tasks were not parallel. B2/B2+ level was the only level that appeared to be parallel (with no significant differences) in the areas of fluency, accuracy and complexity. The numbers of main idea units were significantly different; however, this variable was not found to be a valid measure. B1/B1+ level yielded significant differences in a larger number of variables than any other level, which also rejects the parallelness of the two tasks and suggests that the differences in task complexity factors might have influenced candidates at this level the most. This, in turn, suggests that task parallelness may depend on candidates' proficiency levels. The effect sizes were small for speech rate (-.26) and very small for

the number of main idea units (.16). With such small effect sizes, in addition to the discussion earlier, the difference in the number of main idea units can be ignored. Two of the accuracy variables revealed a significant difference: percentage of error-free clauses with a very large effect of -1.58, and the errors per AS-unit with a medium effect of .39. The remaining variable, errors per 100 words, produced a near-significant result (p = .06) with an effect size of .33. As discussed in Sections 6.3.3.1, the discrepancies between these three variables are due to the different denominators and the spread of errors. Moreover, as the number of errors per 100 words reflects the human ratings of accuracy better (see Section 6.3.2.2), it is appropriate to refer to the results for this variable. In sum, at B1/B1+ level, Task A elicited linguistic performances with decreased fluency and accuracy (but with unchanged complexity).

To summarise, Tasks A and B in this thesis were not parallel in terms of the linguistic performances of the candidates and the NS, presumably because of both task complexity and task induction. On the one hand, the task induction effect of Task B elicited more subordination across all levels of proficiency, which resulted in the higher syntactic complexity of the linguistic performances. On the other hand, the higher task complexity (i.e. mainly the lack of detail in the pictures) may have resulted in decreased fluency and accuracy on Task A.

As mentioned in Section 6.3.1, the results for RQ3-1 and RQ3-2 do not fully support the theories of either Skehan (2009) or Robinson (1995). However, decreased fluency and accuracy appear to be, at least, more in favour of Skehan's predictions than those of Robinson who predicts decreased fluency and increased accuracy on cognitively more difficult tasks. Because of the change in setting, Task A is considered to have required more attention at the stage of conceptualization, and is likely to have interfered with attention allocation in the pre-task planning phase (i.e. the two minutes of planning time) as well as online planning, which might have led to decreased fluency

and accuracy in linguistic performance. At higher levels (i.e. B2/B2+ and NS), this pattern was not observed. At NS level of proficiency, the attentional demands on the Conceptualizer during online planning do not appear to have influenced the linguistic performance; as NS 7 (see Section 5.5) stated, "At first when I looked at the A picture I didn't realise that it was gonna be one of these balloons, then as I worked out along the sequence it was fine." This report indicates that she noticed the need to 'fill in' how the ghost-like figure is made during her performance, but she was able to work it out without loss of fluency while online planning as the story proceeded. Lower level candidates probably would not be able to do this because of their slower grammatical and lexical encoding procedures and incomplete L2 knowledge, which could have led to a decrease in both fluency and accuracy. Regarding L2 speakers' decreased fluency on the cognitively more complex tasks, Foster and Tavakoli (2009) summarise the discussion in relation to Levelt's L1 speech model and their NS data as follows:

> We could perhaps safely assume that in the prelinguistic *conceptualization* phase, native and non-native speakers are on a level playing field, sharing whatever conceptual demands the speaking task imposes. However, in the *formulation* stage, during which linguistic coding of the message takes place, the playing field is not even. Native speakers have greater knowledge of and faster access to the linguistic code and can formulate information even as *conceptualization* feeds it in. Non-native speakers with less knowledge [...] and slower retrieval may have to conceptualize and formulate utterances in alteration rather than in tandem due to insufficient attentional capacity [...]. (Foster & Tavakoli, 2009: 887, italics in the original)

It is clear from Foster and Tavakoli's remarks as well as the reports by the NS in this thesis (RQ2-4) that even NS are not free from the cognitive demands that higher task complexity might impose, though they are able to maintain fluency and accuracy

because of their knowledge and control of linguistic resources. Investigating NS performances also benefited this thesis by identifying the existence of task induction, which has very important implications for the future design of spoken narrative tasks in language testing and task-based research.

6.4. Summary

This chapter has discussed the results of the analyses for the research questions. It has explored how a significant difference, small but unexpected, in task difficulty from MFRM analysis (RQ1) can be interpreted. Evidence from the responses of the Japanese teachers and NS suggested Task A was cognitively more difficult (RQ2), which might have resulted in less fluent and less accurate performances on it, leading to the difference in task difficulty. These responses raise a question about the usefulness of the frameworks of task complexity by Robinson (2001) and Skehan (1998), in that code complexity and cognitive complexity may not be as fixed, invariant or separable as these frameworks seem to suggest. Moreover, given that Task A was more cognitively difficult, the linguistic performances on the two tasks do not fully support the predictions of either Robinson or Skehan (RQ3-1 and RQ3-2).

The variables employed in the main study of this thesis were examined for their validity (RQ4) in terms of correlation with the ratings in each rating category. The variables shown to be most valid include D value, errors per 100 words, and speech rate. The results for the other variables were discussed in terms of the possible reasons why they did not correlate highly. It should be noted that, while the rating scales used in this thesis were shown to be quite useful for accurately rating the samples (see Chapter 4), the results could change if different scales were used, or if descriptors were tailored for spoken narrative tasks or to correspond better with the linguistic variables. Moreover,

detailed examination of the transcripts has explained the discrepancies between the three accuracy variables and the construct of syntactic complexity variables. In so doing, important issues have been raised regarding the relevant variables and the notion of task induction, which has been shown to be separate from the concept of task complexity.

Then, the analysis results of the linguistic performances were revisited, indicating that the two tasks used in the main study were not actually parallel because of the differences in both task complexity and task induction. It was found that the higher cognitive complexity of Task A might have influenced candidates at B1/B1+ level the most, leading to decreased fluency and accuracy. The decreases in fluency and accuracy at this level might lend partial support for Skehan's predictions, and were discussed in relation to Levelt's model (1993) of speech production. Based on the results and discussion developed in the last two chapters, the next chapter, Conclusion, considers the implications and contributions of this thesis, as well as its limitations.

Chapter 7: Conclusion

7.1. Introduction

The purpose of this thesis has been to explore how task parallelness might be established; this is of fundamental importance to any discussion in the areas of language testing and task-based research where equivalence of tasks is a prerequisite. Following a review of relevant literature in the fields of language testing and task-based research (Chapter 2), five pilot studies were conducted (Chapter 3) including two feasibility studies using several linguistic variables to analyse candidate performances (PS1 & 3), a study of expert judgements on the two SST spoken narrative tasks (PS2), a study of NS linguistic performance (PS4), and a study to identify an appropriate pair of tasks for the main study (PS5). The main study examined the parallelness of two spoken narrative tasks by Hill (1960) in terms of the ratings, linguistic performances and expert judgements by Japanese teachers of English, and the perceptions of Japanese candidates and native speakers of English. The validity of the linguistic variables was also examined. The methodology was described in Chapter 4. Chapter 5 reported the results of analyses, in which it was found that the two tasks were not actually parallel in terms of ratings, expert judgements and NS perceptions, or the linguistic performances of Japanese candidates and NS, despite the effort to ensure *a priori* parallelness via the pilot studies. The findings were extensively discussed in Chapter 6 in relation to the theories of task complexity from Robinson (2001) and Skehan (1998). The results of analyses of the linguistic variables also raised several questions regarding variables of accuracy and syntactic complexity. Taken together, the findings of this thesis significantly add to the understanding of task parallelness and the results of my work can be applied not only to the design and selection of tasks but also to the investigation of linguistic performance in the fields of language testing and task-based research.

This chapter concludes the thesis by firstly summarising the findings for each of the research questions. Then, implications and contributions are discussed, followed by the limitations of this thesis and suggestions for future research.

7.2. Synthesis and Summary of Findings

7.2.1. RQ1: Task Difficulty according to MFRM Analysis

The analyses for RQ1, "Is the difficulty of the two tasks the same according to MFRM analysis?", were conducted on the ratings given to the spoken narrative performances by 65 Japanese candidates. Seven raters gave ratings from Below A1 to C1, based on the CEFR Oral Assessment Grid, in the categories of Range, Fluency, Accuracy, Coherence and Sustained Monologue, and then decided on a single overall level (i.e. a Considered Judgement) on the performances on both tasks. Using FACETS, task difficulty was calculated based on Considered Judgement as well as five other rating categories. The results of the ratings of Considered Judgement revealed a significant .40 logit difference between the difficulty of the two tasks (Task A: -0.14 logits; Task B: -0.54 logits). The rating scales had to be collapsed (see Section 5.1.3.1), however, the difficulty based on the ratings in the five rating categories also produced a significant .52 logit difference between the two tasks (Task A: 1.66 logits; Task B: 1.14 logits). According to the MFRM analyses, Task A was significantly more difficult, and it was demonstrated that some of the candidates would receive different (neighbouring) CEFR levels on the two tasks (see Tables 5.5, 5.10 and 5.11). Considering that Tasks A and B were selected and modified so as to make them as parallel as possible, the significant differences found in the values of task difficulty were large and the effects were not ignorable.

7.2.2. RQs 2-1 & 2-2: Candidate Perceptions of the Tasks

Perceptions of the Japanese candidates were investigated in order to collect evidence for the substantive aspect of validity of the two tasks for RQs 2-1 and 2-2: "Are the candidates' perceptions of the two tasks the same?" and "Are the candidates' perceptions of the two tasks the same at different levels of proficiency?" Paired sample *t*-tests were conducted on the responses to a 9-point scale questionnaire originally developed by Robinson (2001) to obtain perceptions of task difficulty, anxiety, self-rating of performance, enjoyment and interest. For RQ2-2, the same analysis was conducted at different levels of proficiency. Each candidate was assigned a CEFR level based on their ability value in logits calculated by FACETS (see Section 5.3). At B2/B2+ and NS levels, Wilcoxon signed rank tests were used due to the small sample size.

Candidate perceptions of the anxiety they felt during the task and self-ratings of performance showed a significant order effect, indicating that the students felt more relaxed and that their performance was better on the second task. For the other three questions involving their perceptions of task difficulty, enjoyment and interest, no significant difference was found between Tasks A and B for RQ2-1 (i.e. with the 65 candidates), although the perceived difficulty was slightly higher for Task A. When participants' perceptions were examined at different levels of proficiency, no significant differences were found between Tasks A and B; however, at B2/B2+ level ($n = 8$), the difference in perceived difficulty was approaching significance ($p = .072$), with Task A being perceived as more difficult. At lower levels of proficiency, the same trend of Task A being perceived as more difficult was maintained, although the findings were not significant.

This suggests the possibility that the candidates with higher proficiency might have been able to perceive that Task A was slightly more difficult because the tasks

were relatively easy for their ability level, and they had less interference at the stage of speech production when the message was being conceptualised. Less able candidates might however not have been able to notice differences in difficulty as their attention might have been occupied with linguistic encoding and overcoming processing difficulties. These findings imply that candidate perceptions of task difficulty might depend on their level of proficiency, which is in line with the remarks by Elder et al. (2002: 364) who suggested that candidate perception is a complex phenomenon involving task characteristics and candidate factors.

7.2.3. Native Speaker Perceptions and Expert Judgements of the Tasks

The responses by two Japanese teachers on the Checklist for Difficulty (Weir & Wu, 2006) were summarised for RQ2-3, "Do Japanese teachers judge the two spoken narrative tasks to be parallel for the candidates in terms of the relevant task complexity factors?" The results indicated that the teachers disagreed with statements that claimed that the students were equally familiar with the roles of people, objects and corresponding lexical items for them in the two tasks. They also doubted that the two tasks shared an equal degree of visibility of events and sufficiency of details. Likewise, six out of 11 NS did not perceive the two tasks to be equally difficult, judging from the results for RQ2-4. Background knowledge (i.e. having experienced babysitting), visibility of pictures and lack of detail were attributed as the sources of difference. In summary, the differences were thought to be due to the changes of setting in Task A, which created a time gap between the pictures and therefore required the narrators to figure out or 'fill in the gap' to describe what happened. Moreover, the findings from the NS perceptions for RQ2-4 support Foster and Tavakoli's (2009) views, who suggested that even NS are not free from the influence of task complexity but can maintain their

performance without loss of fluency due to their automatised knowledge of the language, which enables them to process information quickly as they progress in narration.

Together with the synthesis of the findings from Pilot Study 2, in which expert judgements of the two SST spoken narrative tasks were examined, this thesis suggests important implications for the design of parallel tasks. It is clear that, to establish task parallelness, there are elements other than the factors suggested by the theories of task complexity to be considered. In Pilot Study 2, the two 'supposedly-parallel' SST tasks with a conflict in a public place were not actually parallel in terms of the expert judgements because of the differences in the relationships among the characters, the prominence of characters and the degree of damage, which resulted in the elicitation of different functions that the narrators had to weave into their stories (e.g. 'justifying' was needed to explain the actions of the car driver in the car accident task, whereas it was not elicited in the train station task). Thus, when designing parallel tasks, it is not sufficient to ensure that tasks share the same numbers of elements (i.e. characters, objects, events), a similar range of vocabulary which is assumed to be familiar to the candidates, and the same topic (i.e. a children's trick at a family home) which was assumed to require a very similar amount of reasoning as Robinson (2007) suggests; it is also necessary to pay attention to changes in the setting (Brown & Yule, 1983) as well as the relationships between characters and the functions that the characters perform in the pictures. In addition, collecting judgements from experts who know the candidate population well is essential in order to ensure that the tasks do not require lexical items which are of different degrees of familiarity to the students. It was also pointed out that code complexity, i.e. the language required to complete the task, is not "task-inherent" (Bachman, 2004) or "fixed and invariant" (Robinson, 2001), and is not separable from cognitive complexity, i.e. the amount of computation (or thinking) required to complete

a task because candidate factors, such as background and cultural knowledge, can affect how complex the required language for a candidate is. Since it is assumed to be inevitable that language tests will include lexical items that differ with regard to familiarity, more research is needed to find out to what extent such variability might affect linguistic performance, and therefore the ratings that will be given.

7.2.4. RQ3 & RQ4: Linguistic Performances and Linguistic Variables

Examining the linguistic performances of the candidates on the two tasks aimed to collect evidence of generalisability as an aspect of validity, as per Messick (1996). Paired sample *t*-tests (or Wilcoxon signed rank tests at B2/B2+ and NS levels) were used to compare performance with regard to the frequency of various linguistic variables for RQ3-1, "Are the performances of the two spoken narrative tasks the same in terms of the linguistic variables?", and RQ3-2, "Are the performances of the two spoken narrative tasks the same in terms of the linguistic variables at different levels of proficiency?". The linguistic variables included speech rate (fluency), the percentage of error-free clauses, errors per 100 words, errors per AS-unit (accuracy), D value (lexical complexity), AS-unit length, subordinate clauses per AS-unit (syntactic complexity), positive additive connectives (i.e. incidence of coordination) (syntactic complexity), and the number of main idea units and minor idea units (idea units).

Discussion of the results for RQ3 follows this section on the validity of the linguistic variables which was examined for RQ4: "How do the linguistic variables correlate with the ratings of spoken narrative performance in the corresponding rating categories?" The averaged ratings in the categories of Range, Fluency, Accuracy, Coherence and Sustained Monologue were correlated with the linguistic variables using Pearson's coefficients as follows:

- Range ratings with D value (lexical complexity), AS-unit length and No. of subordinate clauses per AS-unit (syntactic complexity)
- Fluency ratings with speech rate
- Accuracy ratings with % of error-free clauses, No. of errors per 100 words and No. of errors per AS-unit
- Coherence ratings with incidence of coordination
- Sustained Monologue ratings with the number of idea units

The (moderately) high correlation coefficients between ratings and linguistic variables (indicated for Task A, Task B, respectively, in parentheses) were found for the following variables: D value (.470, .469), speech rate (.806, 795), and accuracy variables (% of error-free clauses (.644, .683), errors per 100 words (-.723, -.731), errors per AS-unit (-.652, -.687)). The correlation between these variables and the ratings for Range, Fluency and Accuracy were significant at $p < .01$ level. In accordance with the results of the qualitative analysis of the transcripts, it was found that the number of errors per 100 words was most in line with the Accuracy ratings, which suggests that the overall frequency of errors is a more reliable indicator of overall accuracy than the frequency of error-free clauses.

The transcripts also revealed that even lower rated candidates produced subordination in Task B, which may have resulted in the low and non-significant correlations between the variables of syntactic complexity and the Range ratings. This was attributed to the constant presence of the mother and the plot of exchanging the baby with a ball in Task B, which elicited subordinate clauses with *when, while, where* and *that*. In addition, the transcripts showed that some of the main idea units in Task B contained redundant content, and this may have resulted in the smaller number of main idea units in Task B than in Task A. Thus, the variables of syntactic complexity and idea

units were not shown to be valid with the current data.

The results of the analyses for RQ3-1 showed a significant difference with regard to subordinate clauses per AS-unit, the percentage of error-free clauses, errors per 100 words, and the number of main idea units, indicating that there was less complex (in terms of subordination) and less accurate performance with more main idea units on Task A than on Task B. For RQ3-2, significant differences were found with the following variables at different levels of proficiency. The "A" and arrows in parentheses indicate how the value changed for Task A (i.e. the cognitively more complex task):

- **A2/A2+**: subordinate clauses per AS-unit (A↓); No. of main idea units (A↑);
- **B1/B1+**: % of error-free clauses (A↓); errors per AS-unit (A↑); speech rate (A↓); and No. of main idea units (A↑);
- **B2/B2+**: No. of main idea units (A↑);
- **NS**: AS-unit length (A↓); subordinate clauses per AS-unit (A↓).

The change in the number of main idea units was not considered to be affected by task complexity, as it was shown by the results of the analyses for RQ4 that this variable was not valid.

Among all the levels, the B1/B1+ candidates appeared to have been affected the most by the difference in task complexity between the two tasks; their performance was syntactically less complex (in terms of subordination), less accurate, and less fluent on Task A than on Task B. This suggests that the effects of task complexity on linguistic performance may differ at different levels of proficiency.

In contrast, the number of subordinate clauses per AS-unit displayed a uniform trend across the levels; the amount of subordination was higher on Task B at all levels.

This was not in line with the predictions from the theories of task complexity from Skehan (2009) or Robinson (1995), which suggested that more subordination would be found in cognitively more complex tasks. From this uniform trend, it became clear that subordination can be elicited regardless of the complexity of tasks and candidate proficiency, which is explained as the effect of task induction. This raises questions about the generalisability of predictions from theories of task complexity. Moreover, this finding challenges the conventional use of subordination to measure the syntactic complexity of performances in task-based research. Instead, measuring phrasal complexity (i.e. words per clause) might be more appropriate, as argued by Norris and Ortega (2009).

7.3. Implications of the Findings and Contributions of the Thesis

7.3.1. For Language Testing Research

This thesis has a number of implications for research and practice in the areas of language testing and task-based teaching and learning. For the field of language testing, this thesis has added another multi-method study in the areas of equivalence and parallelness, following the examples of Shohamy (1996), O'Loughlin (2001), and Weir and Wu (2006). More specifically, the parallelness of two narrative tasks was examined in terms of task difficulty calculated by MFRM analysis based on ratings (i.e. score parallelness), responses by Japanese teachers of English and NS (i.e. *a priori* analysis about task complexity), and linguistic performances. Candidate perceptions of the tasks did not show any significant differences; however, the trend of Task A being perceived as more difficult was uniform across the levels, and it was approaching significance at B2/B2+ despite a very small sample size.

This thesis has demonstrated the issues that need to be considered when

designing spoken narrative tasks, incorporating knowledge of SLA research. Unfortunately, O'Loughlin (2001: 126-127) reported a minimum level of concerns about equivalence by a test development team and the item editorial committee of the ***access*** test, and it is hoped that the implications of this thesis would contribute to the understanding of the issue by test developers and task designers who are involved with this task type. Together with the findings of Pilot Study 2, it has been demonstrated that there are different ways in which tasks can be 'seemingly-parallel' but not actually parallel at all. In the case of the SST spoken narrative tasks, both the *train station* and *car accident* tasks depicted 'a conflict in a public place' with the same numbers of characters. However, the differences in the prominence of the characters and relationships between them gave rise to different functions that each character had to perform, which led to the participants perceiving the difficulty of the two SST tasks as different. Interestingly, the two SST tasks are supposed to be given to candidates who are estimated (by the interviewer) to be at the same SST level or above; however, the average TOEIC scores of the SST candidates in Pilot Study 1 indicate that the car accident task might have been given to slightly more proficient candidates (see Table 3.4).[34] Although Mr. Hirano, the Head of the Educational System Department at ALC Press (the developer and administrator of the SST), mentioned that the two tasks have been treated as being of the 'same level' (2008, personal communication), there might have been some bias (or different impressions of the two tasks) on the SST interviewers' part when deciding which tasks to give to the candidates. If this has been the case, differences between tasks within the same task-type do not only pose a problem in task design, but might also influence interviewer decisions which can then affect the final

[34] The average TOEIC scores of the SST candidates who were given the train station task or the car accident task were 586.40 and 666.25, respectively (SST lv.4). They were very similar at SST lv.7, but the SD was much larger for the car accident task.

scores awarded to candidates.

Likewise, the findings of the main study suggest that the changes of setting (Brown & Yule, 1983) have created time gaps between the pictures and insufficiency of details which the candidates have to fill in by themselves. This is assumed to have led to the differences in perceived difficulty by the candidates and NS, expert judgements, and logit values of task difficulty (calculated by MFRM). Although the actual difference in task difficulty was relatively small and could easily be levelled out with other tasks in real-life testing situations, the findings provide valuable insights and should be interpreted as a need to take precautions against non-parallelness or non-equivalence because such small difference in pictures might lead to significant differences in results. Consulting candidate perceptions, together with expert judgements and NS perceptions, may be beneficial at the early stages of task development so that any issues would be detected and rectified. Although they do not always accurately reflect the actual task difficulty, Elder et al. (2002: 363) recommends the use of candidate perceptions as below:

> Test-takers [...] may have some insight into whether a particular task feature or performance condition makes it easier to perform the task, and should perhaps – as Alderson (1988), Stansfield (1991) and Brown (1993) suggested – be consulted at the early stages of test development, along with other parties, to give their feedback on task selection and task design. (Elder et al., 2002: 363)

The second unique contribution of this thesis is the detailed analysis of the candidates' linguistic performances together with a validity study of the linguistic variables used to examine them. Regarding the validity and construct of the linguistic variables, it was demonstrated that speech rate, errors per 100 words, and D value were in accordance with the ratings. In real-life testing practice, examining a number of

transcripts and recordings using such a detailed list of variables may not be realistic. However, it has been indicated that the use of speech rate, errors per 100 words, and D value were valid, all of which could be relatively easily identified using computer technologies for speech recognition (and syllable counting), word counting, and programs to calculate D values.

7.3.2. For Task-Based Research

The first implication of this thesis for the field of task-based research is the importance of ensuring the parallelness (or equivalence, depending on the purpose of research) of tasks before manipulating any variables of task complexity. Specifically, as demonstrated in the previous chapter, it is crucial to conduct an initial pilot study on the tasks in question in order to justify attribution of changes in linguistic performance to manipulation of the factors of task complexity. It is also recommended that judgements from the NS be collected in the piloting phase, as they help identify issues stemming from sources other than L2 processing. Without such piloting, the credibility and reliability of research results are seriously threatened.

Additionally, the conventional use of the amount of subordination as an index of syntactic complexity should be questioned. Through collecting and analysing NS performance, as strongly recommended by Skehan (2009) on the scarcity of such data in the field, it became clear that more subordination was elicited by Task B because of the constant presence of the mother and the plot of replacing the baby with a ball, not because of higher task complexity. Moreover, at other levels including A2/A2+, more subordination was also observed on Task B. If learners with lower levels of proficiency can be pushed to produce subordination by task induction, then the amount of subordination may not be a reliable indication of one's language proficiency being

higher or more advanced. Thus, it is suggested that alternative variables should be used for measuring syntactic complexity, and that one of the possible alternatives might be phrasal complexity, as argued by Norris and Ortega (2009: 561).

The third implication is the challenge that this thesis has made to the theories of task complexity. Robinson (1995) predicted that the absence of pictures (together with memory demands in his conditions) might lead to decreased fluency but increased accuracy and complexity, while Skehan (2009) suggested that tasks which require organising more information may elicit less fluent and accurate, but more complex, performance. The main study in this thesis did not manipulate exactly the same factors as Robinson and Skehan; however, the time gaps and lack of details in Task A required 'filling in the blanks' which relates to the absence of pictures (by Robinson), and required more information to be organised (as Skehan and Foster (1997) manipulated in their study). In the main study, all the investigated areas of performance decreased on the more complex task (Task A), whereas the less complex task (Task B) elicited more complex language in terms of subordination, which led to the development of the concept of task induction. Task induction refers to the effect of pictures inducing the use of subordinate clauses. The difference from the effects of task complexity is that task induction does not assume that more complex language is to be used to express more complex ideas. If we are to generalise and incorporate such predictions into an accurate understanding of task complexity, it is essential to re-examine the results of previous studies to see whether any increase in syntactic complexity can be attributed to relevant factors of task complexity rather than task induction.

Regarding the theories of task complexity, another essential issue to consider is how to ensure that one task is 'more cognitively complex' than others, which is indispensable in confidently attributing any changes in linguistic performance to differences in cognitive complexity (i.e. task complexity). Among the few previous

studies that disclose actual pictures, Foster and Tavakoli (2009) decided that one task was more cognitively complex than another without showing any evidence. It is, therefore, assumed that such decisions have been made intuitively by researchers. However, unless evidence from experts and/or candidates is collected, such assumptions are arguably open to question. Foster and Tavakoli (2009) assumed one of their tasks would be more complex because it contained two storylines (which seem to refer to children's actions and a puppy's actions) which would elicit a larger amount of subordination, and their results confirmed their prediction. However, looking at the pictures, it is questionable whether the storyline of the first task should be regarded as more complex than that of the other task.[35] I would argue that there is still a possibility that more subordination was elicited due to task induction, rather than the complexity of the storylines, unless evidence can be provided to ensure higher cognitive complexity of the first task.

In this thesis, relevant factors of task complexity have been investigated via expert judgements and NS perceptions. However, these may be different from what the Japanese candidates would have judged or perceived, and collecting interview data and the ratio of agreement to a statement (as collected from the Japanese teachers) may not be very useful where a number of tasks need to be evaluated for their complexity. Thus, it may be ideal if independent and quantifiable measures of the cognitive complexity of tasks can also be provided. One such measure with strong potential is subjective time estimation for task completion, which is reported after completing a task by the candidate (Nesbit & Hadwin, 2006: 830). This measure was shown to be an accurate index of cognitive complexity by a study on the speed of information processing and cognitive complexity of tasks using paper-and-pencil-based coding (e.g. copying the

[35] The tasks that were used to compare the effects of storyline complexity by Foster and Tavakoli (2009) were called the *Picnic* task (i.e. assumed to be more complex) and the *Football* task.

given alphabet; replacing a given alphabet with the one before or after it) and letter-matching tasks (Fink & Neubauer, 2001). Although it has rarely been used in educational research (Nesbit & Hadwin, 2006: 830), it could be applied to task-based research, as adding such an independent measure would facilitate comparisons between study results.

7.4. Limitations of This Thesis and Future Research

7.4.1. Limitations of This Thesis

The limitation of this thesis lies primarily in the relatively small sample size, especially at higher levels of proficiency, which raises questions about the generalisability of the findings. Due to the time constraints for data collection for the main study, there were only four weeks in Japan when the author was able to recruit candidates, administer the Oxford QPT and questionnaires, and conduct one-to-one interviews. This was largely affected by the flu pandemic in 2009, which deprived me of the first ten days in Japan due to a university policy about returnees from the UK at that time. Ten more days might have provided thirty or more candidates for the main study, and it might have been more suitable to have a more balanced spread across different levels of proficiency. More data may contribute more to understanding the complex relationship between candidate perceptions, proficiency and task complexity (as Elder et al. (2002) suggested). Especially, more data at higher proficiency levels, including NS levels, are needed to examine to what extent such differences in cognitive complexity might affect candidates at different levels of proficiency. This would lead to deeper understanding of task difficulty relative to candidates' levels of proficiency.

The number of spoken narrative tasks that were examined was also very small, although the main study could not include any more tasks due to time and financial

constraints. It would have been better to include one or two spoken narrative tasks which were not as seemingly-parallel as Task A and B, but quite similar in terms of the relevant factors of task complexity, so that this thesis would have been able to address 'how different' spoken narrative tasks can be in order to assert equivalence, which is the fundamental question in designing equivalent tasks in language testing. Moreover, this thesis has investigated only one task type: spoken narrative tasks. Thus, the generalisability of the findings is also limited in this respect. Additionally, the results of correlation studies (RQ4) would be different if conducted with a different set of tasks. Therefore, again, more data are needed to generalise the validity of the linguistic variables.

Thirdly, use of the CEFR Assessment Grid might not have been the best choice due to the lack of correspondence between the descriptors and the construct of the linguistic variables used in the main study. Although the grid was chosen for the perceived benefits of rater training materials and benchmarked performances, there is still much to be researched about the CEFR (Fulcher, 2003: 110), and its descriptors may not have been the most appropriate for rating spoken narrative performances. Another issue regarding rating scales has been addressed in the interpretations of the results of a validation study of the linguistic variables. It is not clear at this point if the low correlations between the ratings and linguistic variables were due to the insufficient range of descriptors in the rating scales, which resulted in failure to capture the characteristics of linguistic performances that were present, or if this was due to the nature (i.e. construct) of performances on spoken narrative tasks that do not match the descriptors. Clearly, collecting and analysing the performances on various tasks of this type is indispensable to answering this question.

Lastly, there needs to be a methodological improvement regarding collecting candidate perceptions of tasks. It is regrettable that there are no interview data from the

Japanese candidates, because questionnaires can be much more limited in the scope of responses that they collect. Interviews or stimulated recall protocols would have allowed for obtaining considerably more detailed information about the processes during task completion and candidate perceptions. The responses to Robinson's task difficulty questionnaire (2001) did not show any significant differences between the two tasks. However, interviews or stimulated recall might have revealed that the source of the perceived difficulty might have differed. Time and budget permitting, it would have been ideal to have had triangulation of the perceptions and judgements by the candidates, NS and Japanese teachers.

7.4.2. Areas of Future Research

Following the limitations listed above, this section will summarise the directions for future research that could be conducted to extend the findings of this thesis. The research design which involves triangulation of evidence from expert judgements (i.e. *a priori* evidence), analysis of linguistic performance and perceptions by the candidates, teachers and NS (i.e. *a posteriori* evidence) should be applied to researching other tasks of this type to allow for suggestions for practical use in language testing. In addition to the qualitative data collected in this thesis, interview or stimulated recall protocol data from the candidates should be included.

Investigating a set of tasks which includes 'less seemingly-parallel' tasks with different characters, objects, and storylines (which would elicit different functions that the characters perform) will benefit language testers, who need to develop equivalent tasks, and not parallel tasks, for test security purposes. Analysing such tasks, which may be 'more different' than Tasks A and B were in this thesis, may help further support (or question) the argument of this thesis that the theories of task complexity

by Robinson (2001) and Skehan (1998) may be inadequate. In order for this to be achieved, further validity studies of the linguistic variables using different sets of tasks is indispensable, as well as development of more suitable rating scales. Above all, for the results of such research to be reliably interpreted, a larger sample size will be necessary.

Conducting all of such studies will be a time-consuming and iterative process, but as long as language tests and second language research are conducted to make decisions about a person's life, whether it be by test results, research results to feed back to teaching methods, materials and syllabi, the issue of task equivalence is deserving of scrutiny. Although having several limitations, I believe this thesis has succeeded in conducting solid research and in providing valuable implications and suggestions for future research in the fields of language testing and task-based teaching and learning. In contrast to a number of previous studies, this thesis has examined a fundamental issue of comparability, using tasks that are as seemingly-parallel as possible. It is my hope that the results of this thesis will be useful to test developers and task-based researchers and practitioners not only in Japan but also in overseas countries, so that the credibility and reliability of tests and studies are not threatened due to low awareness or neglect in this area.

Bibliography

Aarssen, J. (2001). Development of temporal relations in narratives by Turkish-Dutch bilingual children. In L. Verhoeven & S. Strömqvist (Eds.), *Narrative development in a multilingual context* (pp. 209-232). Amsterdam: John Benjamins.

Albert, Á., & Kormos, J. (2004). Creativity and narrative task performance: An exploratory study. *Language Learning, 54*(2), 277-310.

Alderson, J. C. (1988). New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing, 5*, 220-232.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation.* Cambridge: Cambridge University Press.

Appel, G. (1984). Improving second language production. In H. W. Dechert, D. Möhle & M. Raupach (Eds.), *Second language productions* (pp. 186-210). Tübingen: Gunter Narr.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing, 19*, 453-476.

Bachman, L. F. (2004). *Statistical analyses for language assessment.* Cambridge: Cambridge University Press.

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. (1995). *An investigation into the comparability of two tests of English as a foreign language.* Cambridge: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1982). The construct validation of some components of communicative proficiency. *TESOL Quarterly, 16*(4), 449-465.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26*(2), 390-395.

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind.* Cambridge, MA: MIT Press.

Bauman, J. (n.d.). About the General Service List. Retrieved April 10, 2010, from http://jbauman.com/aboutgsl.html

Berry, V. (2004). *A study of the interaction between individual personality differences and oral performance test facets.* Unpublished PhD thesis, King's College, University of London.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

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.

Brown, A. (1993). The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing, 10*, 277-303.

Brown, A. (2003). Interviewer variation and the co-construction of speaking proficiency *Language Testing, 20*(1), 1-25.

Brown, G., & Yule, G. (1983). *Teaching the spoken language: An approach based on the analysis of conversational English.* Cambridge: Cambridge University Press.

Brown, J. D., Hudson, T., Norris, J., & William, J. B. (2002). *An investigation of second language task-based performance assessments.* Honolulu: University of Hawai'i at Manoa.

Bygate, M. (1999). Quality of language and purpose of task: patterns of learners' language on two oral communication tasks. *Language Teaching Research 3*(3), 185-214.

Bygate, M. (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 23-48). Essex, UK: Pearson Education.

Cabin, R. J., & Mitchell, R. J. (2000). To Bonferroni or not to Bonferroni: When and how are the questions. *Bulletin of Ecological Society of America, 81*(3), 246-248.

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

Candlin, C. N. (1987). Toward task-based language learning. In C. N. Candlin & D. F. Murphy (Eds.), *Language learning tasks* (pp. 5-22). Englewood Cliffs, NJ: Prentice Hall International.

Carrel, P. L. (1982). Cohesion is not coherence. *TESOL Quarterly, 16*, 479-488.

Chalhoub-Deville, M. (1997). Theoretical models, assessment frameworks and test construction. *Language Testing, 14*(1), 3-22.

Chalhoub-Deville, M., & Turner, C. E. (2000). What to look for in ESL admission tests: Cambridge certificate exams, IELTS, and TOEFL. *System, 28*(4), 523-539.

Chapelle, C. A. (1999). Validity in language assessment. *Annual Review of Applied Linguistics 19*, 254-272.

Clark, H. H., & Clark, E. V. (1977). *Psychology and language.* New York: Harcourt Brace Jovanovitch.

Clark, J. L. D., & Li, Y. (1986). *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated assessment model for other less commonly taught languages.* Washington, DC: Center for Applied Linguistics.

Coolican, H. (2004). *Research methods and statistics in psychology* (4th ed.). London:

Hodder & Stoughton.

Corder, S. P. (1981). *Error analysis and interlanguage*. Oxford: Oxford University Press.

Council of Europe. (2001). *Common European framework of reference for languages: Leaning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2009). Manual for relating Language Examinations to the Common European Framework of Reference for Languages (CEFR). Retrieved November 24, 2009, from http://www.coe.int/t/dg4/linguistic/Manuel1_EN.asp

Crookes, G. (1989). Planning and interlanguage variability. *Studies in Second Language Acquisition, 11*, 367-383.

Daigaku Hensachi Juku. (n.d.). Tokyo gaigo daigaku hensachi [Deviation value of Tokyo University of Foreign Studies] Retrieved March 30, 2010, from http://tabutijyuku.livedoor.biz/archives/50788515.html

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

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

de Groot, A. M. D. (1992). Bilingual lexical representation: A closer look at conceptual representations. In R. Frost & L. Katz (Eds.), *Orthography, phonology, morphology, and meaning* (pp. 389-412). Amsterdam: Elsevier.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*(4), 655-679.

Ebel, R. L., & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs, NJ: Prentice Hall.

Eckes, T. (2009). Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment; Section H: Many-Facet Rasch measurement. Retrieved February 23, 2010, from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf

Educational Testing Service. (2008). Reliability and comparability of TOEFL iBT scores. Retrieved January 27, 2009, from http://www.ets.org/Media/Tests/TOEFL/pdf/TOEFL_iBT_Reliability.pdf

Ejzenberg, R. (2000). The juggling act of oral fluency: A psycho-sociolinguistic metaphor. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 287-313). Michigan: The University of Michigan Press.

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing, 19*, 347-368.

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

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

Press.

Ffrench, A. (n.d.). DVD "Spoken performances illustrating the 6 levels of the Common European Framework of Reference for Languages": Comments on the assigned levels in English. Retrieved November 3, 2009, from http://www.ciep.fr/en/publi_evalcert/dvd-productions-orales-cecrl/docs/comments_en.pdf

Fink, A., & Neubauer, A. C. (2001). Speed of information processing, psychometric intelligence and time estimation as an index of cognitive load. *Personality and Individual Differences, 30*(6), 1009-1021.

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

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59*(4), 866-896.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*, 354-375.

Fromkin, V. A. (1971). The non-anomalous nature of anomalous utterances. *Language, 4*, 27-52.

Fulcher, G. (1996a). Does thick description leada to smart tests? A data-based approach to rating scale construction. *Language Testing, 13*, 208-238.

Fulcher, G. (1996b). Testing tasks: Issues in task design and the group oral. *Language Testing, 13*(1), 23-51.

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

Gathercole, S. E., & Baddeley, A. D. (1993). *Working memory and language.* Hillsdale, NJ: Laurence Erlbaum.

Gilabert, R. (2005). *Task complexity and oral narrative production.* Unpublished PhD Thesis, Universitat de Barcelona, Barcelona.

Givon, T. (1985). Functionn, structure and language acquisition. In D. Slobin (Ed.), *The crosslinguistic study of language acquisition* (Vol. 1, pp. 1008-1025). Hillsdale, NJ: Lawrence Erlbaum.

Givon, T. (1995). *Functionalism and grammar.* Amsterdam: John Benjamins.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English.* Harlow: Longman.

Henning, G. (1987). *A guide to language testing: Development, evaluation, research.* Cambridge: MA: Newbury House Publishers.

Hill, L. A. (1960). *Picture composition book.* London: Longman.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics 6*(2), 65-70.

Holmes, J. (2003). Narrative structure: Some contrasts. In C. B. Paulson & G. R. Tucker (Eds.), *Sociolinguistics: The essential readings.* Oxford: Blackwell.

Horai, T. (2009). Intra-task comparison in a monologic oral performance test: The impact of

task manipulation on performance. In A. Brown & K. Hill (Eds.), *Tasks and criteria in performance assessment.* Frankfurt: Peter Lang.

Hughes, A., & Lascaratou, C. (1982). Competing criteria for error gravity. *ELT Journal, 36*(3), 175-182.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). Harmondsworth: Penguin.

Inoue, C. (2010). Investigating the sensitivity of the measures of fluency, accuracy, complexity and idea units with a narrative task. *Papers from the Lancaster University Postgraduate Conference in Linguistics and Language Teaching, 4,* 129-153.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401-436.

JACET. (2003). *JACET8000.* Tokyo: JACET English Vocabulary SIG.

James, C. (1998). *Errors in language learning and use.* Essex: Addison Wesley Longman.

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*(1), 57-84.

Kempen, G., & Hoenkamp, E. (1987). An incremental procedural grammar for sentence formulation. *Cognitive Science, 11,* 202-258.

Kenyon, D. M., & Malabonga, V. (2001). Comparing examinee attitudes toward computer-assisted and other oral proficiency assessments. *Language Learning & Technology, 5,* 60-83.

Kenyon, D. M., & Tschirner, E. (2000). The rating of direct and semi-direct oral proficiency interviews: Comparing performance at lower proficiency levels. *The Modern Language Journal, 84*(1), 85-101.

Koizumi, R. (2005). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET Journal, 9,* 1-33.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.

Koponen, M., & Riggenbach, H. (2000). Overview: Varying perspectives on fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 5-23). Michigan: The University of Michigan Press.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning, 49*(2), 303-342.

Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, NJ: Lawrence Erlbaum.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32,* 145-164.

Kroll, B. (1990). What does time buy? ESL student performance on home versus class compositions. In B. Kroll (Ed.), *Second language writing* (pp. 140-154). Cambridge:

Cambridge University Press.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics, 45*(3), 261-284.

Labov, W. (1972). *Language in the inner city: Studies in the Black English vernacular.* Oxford: Basil Blackwell.

Lado, R. (1961). *Language testing.* London: Longman.

Laufer, B., & Nation, P. (1995). Vocabulary size and use - Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-322.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Michigan: The University of Michigan Press.

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

Linacre, J. M. (1994). *Many-facet Rasch measurement* (2nd ed.). Chicago: MESA Press.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith Jr. & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 258-278). Maple Grove, MN: JAM Press.

Linacre, J. M. (2009). A user's guide to FACETS: Rasch-model computer programs. Retrieved February 15, 2010, from http://www.winsteps.com/facets.htm

Liskin-Gasparro, J. E. (1996). Narrative strategies: A case study of developing storytelling skills by a learner of Spanish. *The Modern Language Journal, 80*(3), 271-286.

Lumley, T., & McNamara, T. (1995). Rater characteristics and rater bias: Implications for training *Language Testing, 12*(1), 54-71.

Luoma, S. (2004). *Assessing speaking.* Cambridge: Cambridge University Press.

Masters, G. N. (1982). A Rasch model for partial scoring. *Psychometrika, 47*, 149-174.

McNamara, T. F. (1996). *Measuring second language performance.* London: Longman.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*(3), 5-19.

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

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13-103). New York: Macmillan/American Council on Education.

Messick, S. (1996). Validity and washback in language testing. *Language Testing, 13*(2), 241-256.

Montanari, S. (2004). The development of narrative competence in the L1 and L2 of Spanish-English bilingual children. *Journal of Bilingualism, 8*(4), 449-497.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*(6), 1044-1045.

Nesbit, J. C., & Hadwin, A. F. (2006). Methodological issues in educational psychology. In E.

Anderman, P. H. Winne, P. A. Alexander & C. Lyn (Eds.), *Handbook of educational psychology* (2nd ed., pp. 825-848). New York: Routledge.

Norris, J. M., Brown, J. D., Hudson, T. D., & Bonk, W. (2002). Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing, 19*(4), 395-418.

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.

North, B. (2000). *The development of a common framework scale of language proficiency.* New York: Peter Lang.

O'Loughlin, K. (2001). *The equivalence of direct and semi-direct speaking tests.* Cambridge: Cambridge University Press

O'Sullivan, B. (2000). Exploring gender and Oral Proficiency Interview performance *System, 28*(3), 373-386.

O'Sullivan, B., Weir, C., & Saville, N. (2002). Using observation checklists to validate speaking-test tasks. *Language Testing, 19(1)*, 33-56.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*(1), 109-148.

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language.* Amsterdam: John Benjamins.

Paradis, M. (1987). *The assessment of bilingual aphasia.* Hilldsdale, NJ: Lawrence Erlbaum.

Polio, C. G. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*(1), 101-143.

Poulisse, N. (1997). Language production in bilinguals In A. M. D. de Groot & J. F. Kroll (Eds.), *Tutorials in bilingualism: Psycholinguistic perspectives.* Mahwah, NJ: Lawrence Erlbaum.

Poulisse, N., & Bongaert, T. (1994). First language use in second language production. *Applied Linguistics, 15*(1), 36-57.

Qian, D. D. (2009). Comparing direct and semi-direct modes for speaking assessment: Effects on test takers. *Language Assessment Quarterly, 6*(2), 113-125.

Révész, A. (2007). *Focus on form in task-based language teaching: Recasts, task complexity, and L2 learning.* Unpublished PhD Thesis, Columbia University, New York.

Rice, W. R. (1989). Analyzing tables of statistical tests. *Evolution, 43*(1), 223-225.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(1), 99-140.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 23*, 27-57.

Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language

learning. *Second Language Studies, 21*(2), 45-105.

Robinson, P. (2005). Cognitive complexity and task sequencing:Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*, 1-32.

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. d. P. García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7-26). Clevedon: Multilingual Matters.

Robinson, P., Ting, S., & Urwin, J. J. (1995). Investigating second language task complexity. *RELC Journal, 26*(2), 62-79.

Sato, C. (1988). Origin of complex syntax in interlanguage development. *Studies in Second Language Acquisition, 10*, 371-395.

Sato, C. (1990). *The syntax of conversation in interlanguage development.* Tubingen: Gunter Narr.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Segalowitz, N. (2010). *Cognitive bases of second language fluency.* New York: Routledge.

Shohamy, E. (1994). The validity of direct versus semi-direct oral tests. *Language Testing, 11(2)*, 99-124.

Skehan, P. (1996). A framework for the implementation of task based instruction. *Applied Linguistics, 17*, 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. *Applied Linguistics, 30*(4), 510-532.

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*, 93-120.

Spolsky, B. (1995). *Measured words.* Oxford: Oxford University Press.

Stansfield, C. W. (1990). An evaluation of Simulated Oral Proficiency Interviews as measure of oral proficiency. In J. E. Alatis (Ed.), *Linguistics, language teaching and language acquisition: The interdependence of theory, practice and research* (pp. 228-234). Washington, D.C.: Georgetown University Press.

Stansfield, C. W. (1991). A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (Ed.), *Current developments in language testing* (pp. 199-209). Singapore: SEAMEO RELC.

The Society for Testing English Proficiency. (2010). Overview of the EIKEN tests Retrieved September 25, 2010, from http://stepeiken.org/overview-eiken-tests

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French *Applied Linguistics, 17*(1), 84-119.

UCLES. (2001a). *CD ROM User Manual: Quick Placement Test.* Oxford: Oxford University

Press.

UCLES. (2001b). FCE Handbook Retrieved April 10, 2010, from http://www.metodocallan.net/fce_hb_intro.pdf

van Lier, L. (1989). Reeling, writhing, drawling, stretching, and fainting in coils: Oral Proficiency Interviews as conversation. *TESOL Quarterly, 23*(3), 489-508.

Verhoeven, L., & Strömqvist, S. (2001). *Narrative development in a multilingual context.* Amsterdam: John Benjamins.

Viberg, Å. (2001). Age-related and L2-related features in bilingual narrative development in Sweden. In L. Verhoeven & S. Strömqvist (Eds.), *Narrative development in a multilingual context* (pp. 87-128). Amsterdam: John Benjamins.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weir, C. J. (2005). *Language testing and validation.* Houndmills: Palgrave Macmillan.

Weir, C. J., O'Sullivan, B., & Horai, T. (2009). Exploring difficulty in speaking tasks: An intra-task perspective. Retrieved May 12, 2011, from http://www.ielts.org/pdf/vol6_report5.pdf

Weir, C. J., & Wu, J. R. W. (2006). Establishing test form and individual task comparability: a case study of a semi-direct speaking test. *Language Testing, 23*(3), 167-197.

Wickens, C. (1984). Processing resources in attention. In R. Parasuraman & D. Davies (Eds.), *Varieties of attention* (pp. 63-102). New York: Academic Press.

Wickens, C. (2007). Attention to the second language. *International Review of Applied Linguistics, 45*(3), 177-191.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*, 167-197.

Wolfe-Quintero, K., Inagaki, S., & Hae-Young, K. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity.* Honolulu: University of Hawaii, Manoa.

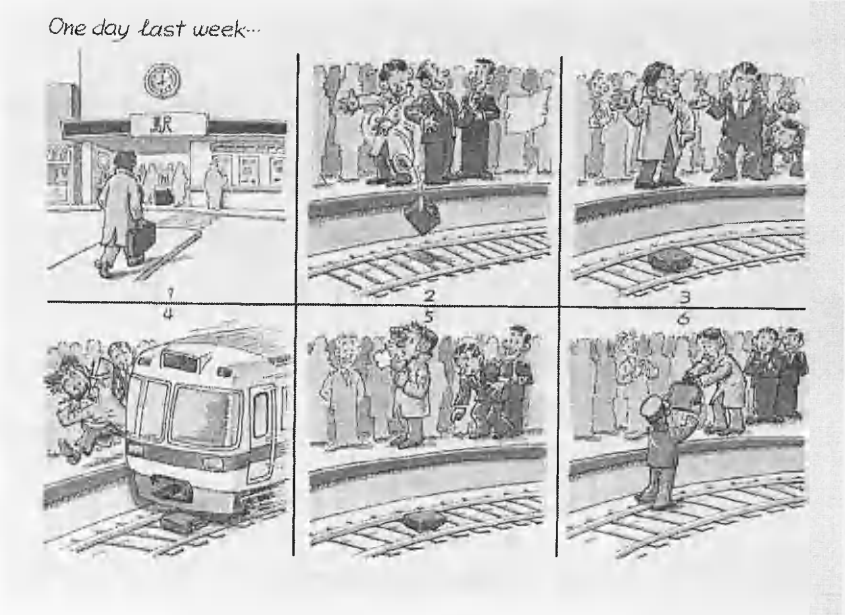
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

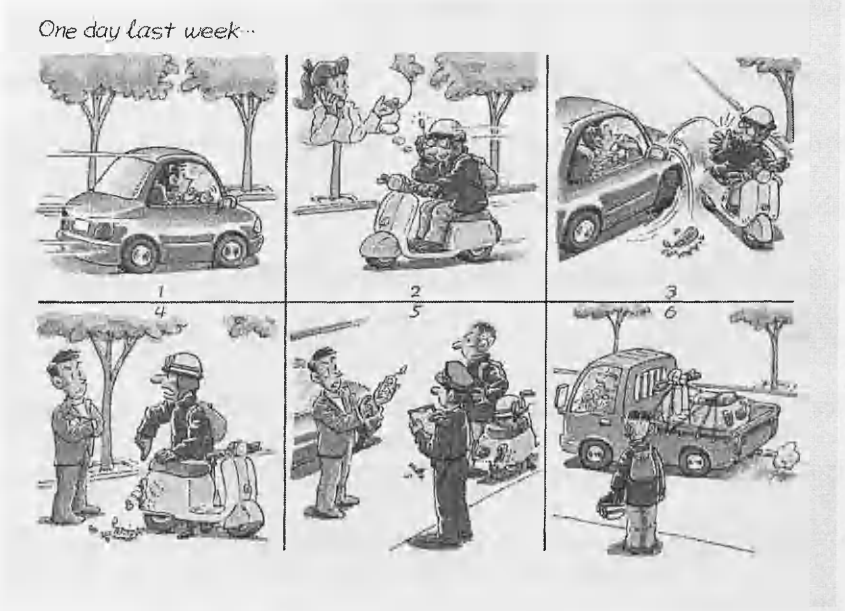
Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity, and accuracy in L2 monologic oral production. *Applied Linguistics, 24*, 1-27.

Zhou, Y. (2009). *Effects of computer delivery mode on testing second language speaking: The case of monologic tasks.* Unpublished PhD thesis, Tokyo University of Foreign Studies Tokyo.

Appendices

Appendix 1: Two Spoken Narrative Tasks from the SST





Note. The *train station* task (above) and the *car accident* task (below) are reprinted in monochrome in this thesis with special permission from the ALC Press. As these tasks are part of the test materials of the ongoing SST, it is strictly prohibited to photocopy and/or distribute them.

Appendix 2: Sample Transcripts by the SST candidates

[Level 4: *train station* task]

on my way to the office I had the happening last week / at station it was eight o'clock / and it was a usual day / but I was in the form / and I waited for the train / when I waited for the train I falled the bag / and there is the man who was next to me / and his arm hit my arm / and I fall the bag / and the bag was the under the form / so I had a trouble / and the train came in the station / and I was surprised / and I thought my bag was broken by the train / but the train left the station / I found the bag was safe / and I called the station clerk / and I got it /

[Level 7*: train station* task]

last week I went to the station as usual on my way to my office / I usually take a train around eight o'clock because my work starts at nine / and it takes around an hour or so / and when I went to the platform to take a train the man who stand in front of me hit my elbow / and I dropped my briefcase / and suddenly my briefcase fell down my elbows and my arms / and it fell down to the lane / and what made me angry that a man said "What happened because I didn't nothing." / but the man hit me so that my briefcase fell down the lane / and he looks that something happened / but he's not serious / so I was very angry with him / and after that the express train passed / and I thought my briefcase will be crashed / but fortunately my briefcase was safe / so inside my briefcase it didn't happen anything / and the train officer kindly take up my briefcase / and gave it to me / so you know my documents and files and everything was safe / so I was able to bring my briefcase to my office / so that's a good story /

Appendix 3: Sample Transcripts by a Native Speaker of English

[*Train station* task]

one day last week Mr. Brown was heading to the station to go to work as usual / he was just on time for his train at eight o'clock / but when he was standing on the very crowded platform a terrible thing happened / and his bag fell on to the track / he didn't know how to respond to this because he thought it was the fault of a very gregarious man standing beside him waving his arms about / when he questioned this man the man just shrugged / and said it's nothing to do with me while his friend kind of looked on in a puzzled fashion / this didn't really help with the situation / the bag was still on the track / and Mr. Brown wasn't feeling too pleased about such things / he had some important papers in his bag that he was going to need in a meeting later / so they stood there discussing it / and then came a long rushed train / Mr. Brown was [laughter] he was shocked beyond all belief / he couldn't just stand to see his bag with his papers and his laptop going underneath the train / he nearly threw himself out onto the tracks / he was appalled / all his hard work he put for the last number of weeks was all in that bag / and it was all going underneath the train / however once the train rushed past and the wind died down he was mopping his sweating blouse / he thought about his boss and what was going to happen in the office when the smug man in the blue suit that had been stood beside him started laughing and pointing / and he could see that the train had actually run straight over the bag / and hadn't damaged anything at all / Mr. Brown was slightly embarrassed / everybody else was looking round smiling / and eventually a very helpful guard jumped out on the tracks / and pitched the bag up for Mr. Brown / and he had to go and hide himself in the crowd / go to work /

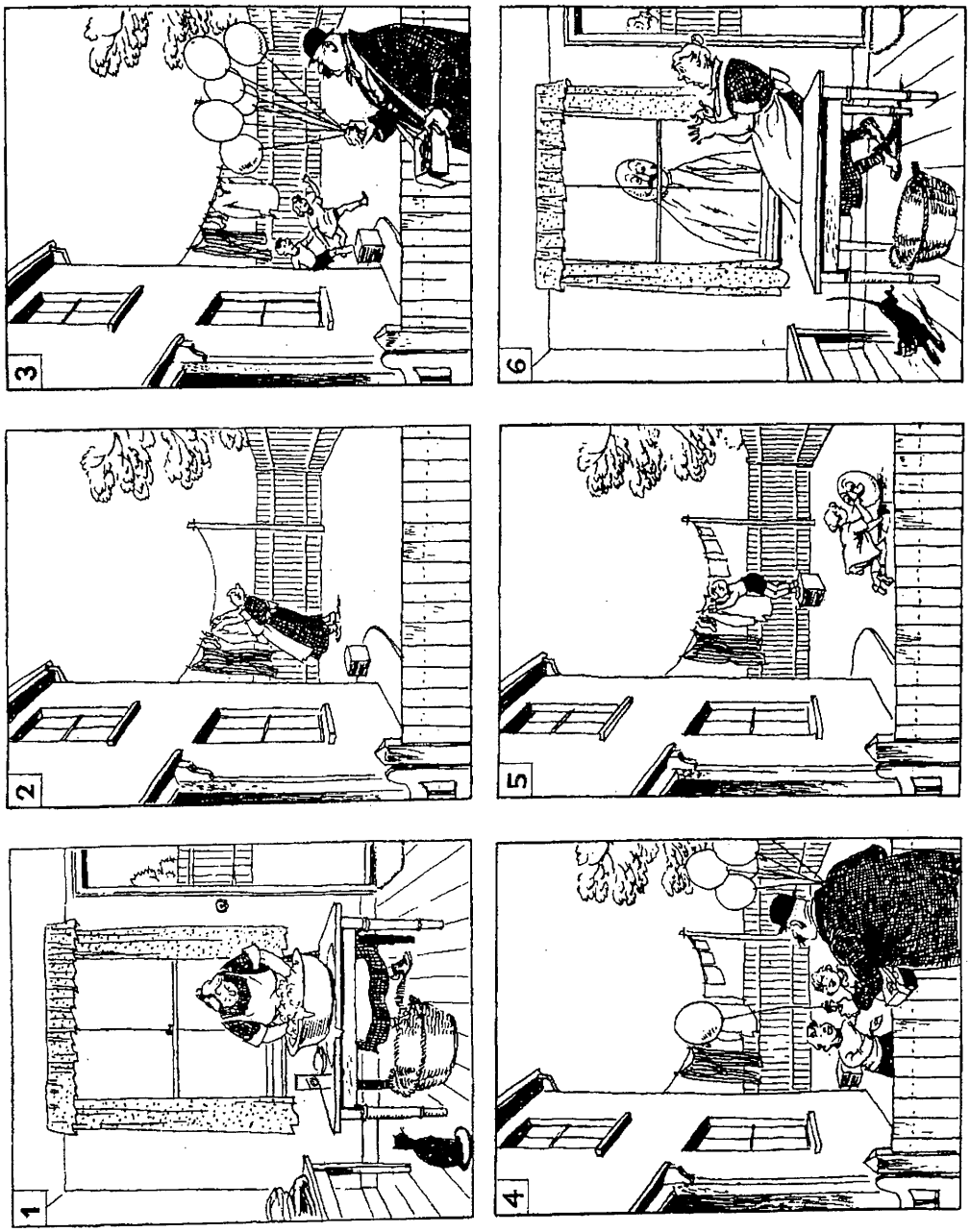
[*Car accident* task]

one day last week Digby Jones was driving along in his small orange car which he was

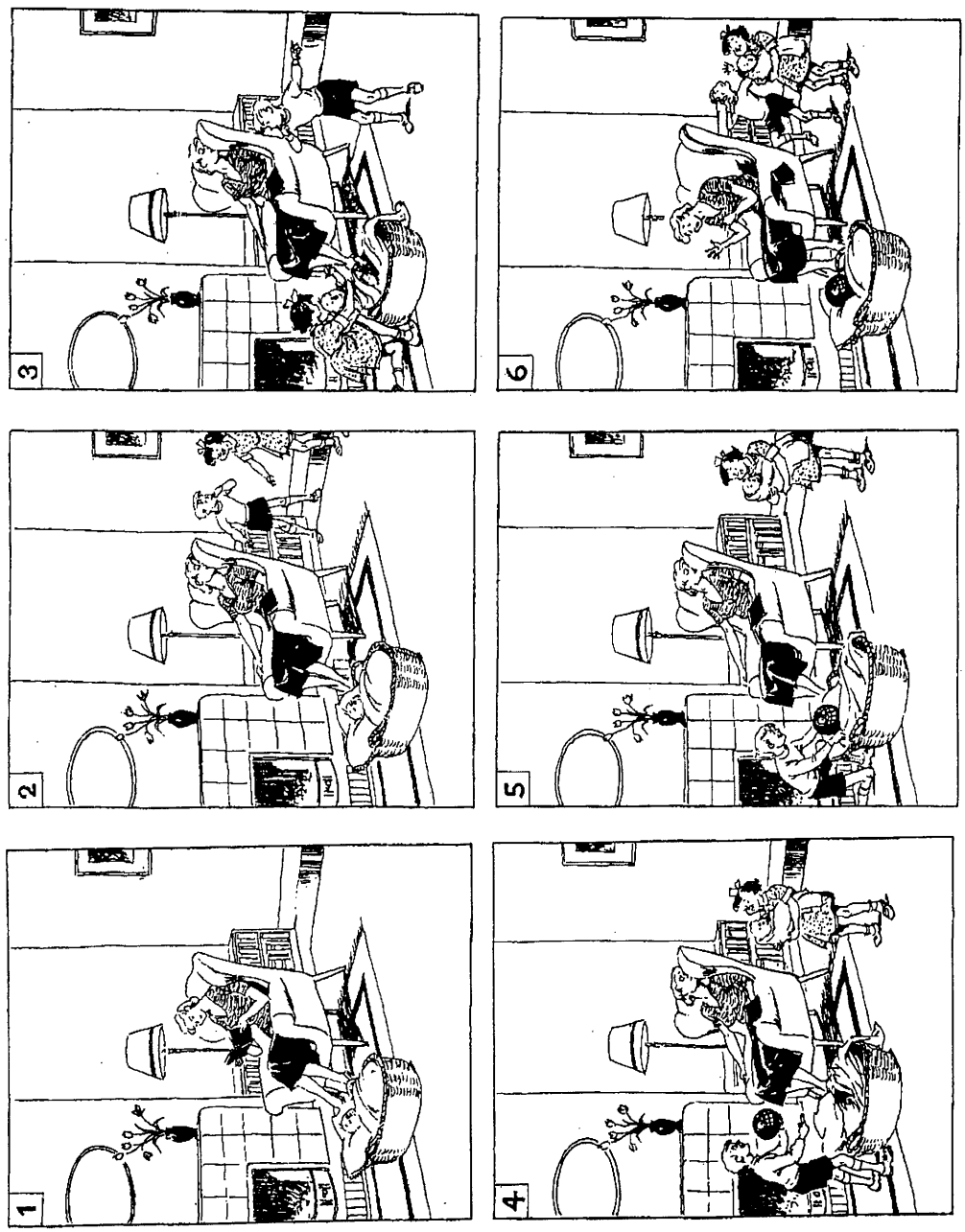
very proud of / he was rushing off to go and see his mum whose birthday it was / however Steven Smith was coming the opposite direction on his little green scooter / and now Steven Smith had only just started a relationship with a with a girl called Mandy / and he was chatting to her and thinking about her / and he had his goggles on / but his cell phone was out / and he wasn't really thinking about what he was doing / and when suddenly Digby Jones in his little orange car came rushing up and they somehow managed to clash and the cell phone hit the hit the wing mirror of the car / and the car screeched to a halt / and Digby was taken away from the thoughts of his birthday cake and mum / and Steven Smith was stopped thinking about Mandy / and it all went wrong / and basically Digby wasn't pleased / he'd been following the highway code / he'd been driving along in a straight line / and this bloke who had no reason to be speaking on his cell phone had like caused all sorts of problems / but Steven Smith on the other hand really wasn't pleased because he thought he was driving along the right side of the road / and suddenly his scooter was damaged which he was particularly proud of / the phone was on the floor / and he hadn't been able to say goodbye to Mandy / eventually some bystander called the police who came along to take a record of the incident / and Digby spent a lot of time telling the policeman all about the cell phone / and how it certainly hadn't been his fault / and basically it was all to do with the careless scooter driver / the police were listening to everything and made no comment and then eventually / I don't know / the issues were resolved [rising intonation] / and the scooter was taken away to be mended by the insurance man / and Digby Jones went off to see his mum celebrating happy birthday /

Appendix 4: Spoken Narrative Tasks for the Main Study by Hill (1960)

A



B



Appendix 5: Perceived Task Difficulty Questionnaire

Please circle one of the numbers as appropriate about the task that you have just narrated.

1. How easy was this task?

0	1	2	3	4	5	6	7	8	9
Very easy		Easy		Moderately easy	Moderately difficult		Difficult		Very difficult

2. How nervous were you to do this task?

0	1	2	3	4	5	6	7	8	9
Very relaxed		Relaxed		Moderately relaxed	Moderately nervous		Nervous		Very nervous

3. How well do you think you did this task?

0	1	2	3	4	5	6	7	8	9
Didn't do well at all		Didn't do well		Moderately poor	Moderately well		Well		Very well

4. How interesting did you think this task was?

0	1	2	3	4	5	6	7	8	9
Not at all interesting		Not interesting		Moderately not interesting	Moderately interesting		Interesting		Very interesting

5. Would you like to do more tasks like this?

0	1	2	3	4	5	6	7	8	9
Not at all		No		Moderately no	Moderately yes		Yes		Very much yes

Note. This questionnaire was given in Japanese in the main study.

Appendix 6: CEFR Assessment Grids

Table 1: GLOBAL ORAL ASSESSMENT SCALE

C2	***Conveys finer shades of meaning precisely and naturally.*** Can express him/herself spontaneously and very fluently, interacting with ease and skill, and differentiating finer shades of meaning precisely. Can produce clear, smoothly-flowing, well-structured descriptions.
C1	***Shows fluent, spontaneous expression in clear, well-structured speech.*** Can express him/herself fluently and spontaneously, almost effortlessly, with a smooth flow of language. Can give clear, detailed descriptions of complex subjects. High degree of accuracy; errors are rare.
B2+	
B2	***Expresses points of view without noticeable strain.*** Can interact on a wide range of topics and produce stretches of language with a fairly even tempo. Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest. Does not make errors which cause misunderstanding.
B1 +	
B1	***Relates comprehensibly the main points he/she wants to make.*** Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair may be very evident. Can link discrete, simple elements into a connected, sequence to give straightforward descriptions on a variety of familiar subjects within his/her field of interest. Reasonably accurate use of main repertoire associated with more predictable situations.
A2+	
A2	***Relates basic information on, e.g. work, family, free time etc.*** Can communicate in a simple and direct exchange of information on familiar matters. Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident. Can describe in simple terms family, living conditions, educational background, present or most recent job. Uses some simple structures correctly, but may systematically make basic mistakes.
A1	***Makes simple statements on personal details and very familiar topics.*** Can make him/herself understood in a simple way, asking and answering questions about personal details, provided the other person talks slowly and clearly and is prepared to help. Can manage very short, isolated, mainly pre-packaged utterances. Much pausing to search for expressions, to articulate less familiar words.
Below A1	Does not reach the standard for A1.

Table 2: ORAL ASSESSMENT CRITERIA GRID

	RANGE	ACCURACY	FLUENCY	COHERENCE	SUSTAINED MONOLOGUE
C1	Has a good command of a broad range of language allowing him/her to select a formulation to express him/ herself clearly in an appropriate style on a wide range of general, academic, professional or leisure topics without having to restrict what he/she wants to say.	Consistently maintains a high degree of grammatical accuracy; errors are rare, difficult to spot and generally corrected when they do occur.	Can express him/herself fluently and spontaneously, almost effortlessly. Only a conceptually difficult subject can hinder a natural, smooth flow of language.	Can produce clear, smoothly flowing, well-structured speech, showing controlled use of organisational patterns, connectors and cohesive devices.	Can give elaborate descriptions and narratives, integrating sub-themes, developing particular points and rounding off with an appropriate conclusion
B2+					
B2	Has a sufficient range of language to be able to give clear descriptions, express viewpoints on most general topics, without much conspicuous searching for words, using some complex sentence forms to do so.	Shows a relatively high degree of grammatical control. Does not make errors which cause misunderstanding, and can correct most of his/her mistakes.	Can produce stretches of language with a fairly even tempo; although he/she can be hesitant as he or she searches for patterns and expressions, there are few noticeably long pauses.	Can use a limited number of cohesive devices to link his/her utterances into clear, coherent discourse, though there may be some "jumpiness" in a long contribution.	Can give clear, detailed descriptions on a wide range of subjects related to his/her field of interest.
B1+					
B1	Has enough language to get by, with sufficient vocabulary to express him/herself with some hesitation and circumlocutions on topics such as family, hobbies and interests, work, travel, and current events.	Uses reasonably accurately a repertoire of frequently used "routines" and patterns associated with more predictable situations.	Can keep going comprehensibly, even though pausing for grammatical and lexical planning and repair is very evident, especially in longer stretches of free production.	Can link a series of shorter, discrete simple elements into a connected, linear sequence of points.	Can reasonably fluently relate a straightforward narrative or description as a linear sequence of points. Can give detailed accounts of experiences, describing feelings and reactions. Can describe events, real or imagined. Can narrate a story.
A2+					
A2	Uses basic sentence patterns with memorised phrases, groups of a few words and formulae in order to communicate limited information in simple everyday situations.	Uses some simple structures correctly, but still systematically makes basic mistakes.	Can make him/herself understood in very short utterances, even though pauses, false starts and reformulation are very evident.	Can link groups of words with simple connectors like "and, "but" and "because".	Can describe people, places, and possessions in simple terms. Can give a simple description or presentation of people, living or working conditions, daily routines, likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list.
A1	Has a very basic repertoire of words and simple phrases related to personal details and particular concrete situations.	Shows only limited control of a few simple grammatical structures and sentence patterns in a memorised repertoire.	Can manage very short, isolated, mainly pre-packaged utterances, with much pausing to search for expressions, to articulate less familiar words, and to repair communication.	Can link words or groups of words with very basic linear connectors like "and" or "then".	Can produce simple mainly isolated phrases about people and places.

Table 3: SUPPLEMENTARY CRITERIA GRID: "Plus Levels"

	RANGE	ACCURACY	FLUENCY	COHERENCE	SUSTAINED MONOLOGUE
C1					
B2+	Can express him/herself clearly and without much sign of having to restrict what he/she wants to say.	Shows good grammatical control; occasional "slips" or non-systematic errors and minor flaws in sentence structure may still occur, but they are rare and can often be corrected in retrospect.	Can communicate spontaneously, often showing remarkable fluency and ease of expression in even longer complex stretches of speech. Can use circumlocution and paraphrase to cover gaps in vocabulary and structure.	Can use a variety of linking words efficiently to mark clearly the relationships between ideas.	*No descriptor available*
B2					
B1+	Has a sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films.	Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influences.	Can express him/herself with relative ease. Despite some problems with formulation resulting in pauses and "cul-de-sacs", he/she is able to keep going effectively without help.	*No descriptor available*	*No descriptor available*
B1					
A2+	Has sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics, though he/she will generally have to compromise the message and search for words.	*No descriptor available*	Can adapt rehearsed memorised simple phrases to particular situations with sufficient ease to handle short routine exchanges without undue effort, despite very noticeable hesitation and false starts.	Can use the most frequently occurring connectors to link simple sentences in order to tell a story or describe something as a simple list of points.	Can tell a story or describe something in a simple list of points. Can give short, basic descriptions of events and activities. Can use simple descriptive language to make brief statements about and compare objects and possessions.
A2					
A1					

Appendix 7: Rating Sheet

SAMPLE NO. _____

Levels : Below-A1, A1, A2, A2+, B1, B1+, B2, B2+, C1, C2

1. Initial Impression

2. Detailed Analysis with Grid

RANGE	ACCURACY	FLUENCY	COHERENCE	SUSTAINED MONOLOGUE

3. Considered Judgment

Notes: