Title: The impact of pre-task planning on speaking test performance for English-medium university study

Name: Stefan O'Grady

THE IMPACT OF PRE-TASK PLANNING ON SPEAKING TEST PERFORMANCE FOR ENGLISH-MEDIUM UNIVERSITY STUDY

By

Stefan O'Grady

A thesis submitted to the University of Bedfordshire, in partial fulfilment of the requirements for the degree of Doctor of Philosophy

January 2018

Academic Thesis: Declaration of Authorship

I, Stefan O’Grady declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

The impact of pre-task planning on speaking test performance for English-medium university study

I confirm that:

1. This work was done wholly or mainly while in candidature for a research degree at this University;
2. Where any part of this thesis has been submitted for a degree of any other qualification at this University or any other institution, this has been clearly stated;
3. Where I have cited the published work of others, this is always clearly attributed;
4. Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
5. I have acknowledged all main sources of help;
6. Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
7. None of this work has been published before submission.

Name of Candidate: Stefan O’Grady Signature: Stefan O'Grady

Date: 2nd January 2018

THE IMPACT OF PRE-TASK PLANNING ON SPEAKING TEST PERFORMANCE FOR ENGLISH-MEDIUM UNIVERSITY STUDY

STEFAN O’GRADY

ABSTRACT

This study investigated the impact of different lengths of pre-task planning time on performance in a test of second language speaking ability for university admission. The research was conducted in a university in Turkey where the increasing popularity of English-medium instruction has heightened the need for valid assessment of prospective students’ English language ability.

In the study, 47 Turkish speaking learners of English aged between 18 and 22, sat a test of English language speaking ability. The participants were divided into two groups according to their language proficiency estimated through a paper-based English placement test (an A1+/A2 level and B1 level group, Council of Europe, 2001). They each completed four monologue tasks: two picture-based narrative tasks and two description tasks. In a balanced design, each test taker was allowed a different length of planning time before responding to each of the four tasks. The four planning conditions were 30 seconds, one minute, five minutes, and ten minutes.

The effect of variation in pre-task planning time was analysed using a set of measures of complexity, accuracy and fluency identified through the literature review and refined through piloting. In addition, 16 trained raters awarded scores to the test takers using an analytic rating scale and a context specific, binary choice rating scale designed specifically for the study. The results of the rater scores were analysed using multi-faceted Rasch measurement.

The impact of pre-task planning on test scores was found to be influenced by four variables: the method of assessment, the task type that test takers completed, the length of planning time provided, and the test takers’ levels of proficiency in the second language.

Firstly, contrary to common accounts in the literature, pre-task planning did not have an impact on the complexity, accuracy, and fluency of the spoken output. Rather, planning for longer periods of time increased the number of idea units test takers produced (an indication of the propositional completeness and complexity of the task content), and led to marginal increases in test scores. The increases in scores were larger on the picture-based narrative tasks than the two description tasks.

The results also revealed a relationship between proficiency and pre-task planning whereby statistical significance was only reached for the increases in the scores of the lowest (CEFR ‘A’) level test takers. Regarding the amount of planning time, the five-minute planning condition led to the largest overall increases in scores. The research findings offer contributions to the study of pre-task planning and will be of particular interest to institutions seeking to assess the speaking ability of prospective students in English-medium educational environments.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Acknowledgements

I wish to express my sincerest thanks to Professor Anthony Green for being my supervisor. I am indebted to him for his generous support, guidance and endless patience throughout my PhD studies. Dr Chihiro Inoue for taking an interest in my project, acting as my second supervisor and for many insightful discussions about language testing and task-based language teaching. Dr John Field for his support during the early stages of the project and for sharing his expertise on the important role of cognition in language production. Professor Barry O'Sullivan and Dr Sathena Chan for acting as my examiners and for their detailed comments on the thesis.

Dr John O'Dwyer for suggesting that I begin the PhD project in 2013 and for outlining the potential importance of research in to English language assessment in Turkey. Dr Esin Caglayan and Dr Evrim Ustunluoglu for their encouragement and support. I would also like to thank Tom Rogers for many discussions about our research projects, and for being a reliable companion over the past four years.

The staff and students in the school of Foreign Languages at Izmir University of Economics, whose participation made the research possible. I am grateful to you all for your time and effort.

Last but not least, my family, Mum, Dad, Rhiannon, Nathaniel and Lucie. Above all, my wife Simge and my son Harry Kartal, who was born during the writing of this dissertation and is lucky enough to be raised in a bilingual, Turkish/English environment.

1 Research Context

1.1 English-medium higher education in Turkey

English-medium instruction is increasingly common in Turkish higher education (Selvi, 2014). The prestige and future economic reward an English education is able to bestow compels many prospective students to apply to English-medium universities for undergraduate studies. However, despite widespread national recognition of the importance for English, a high percentage of Turkish students leave secondary education without sufficient proficiency to follow English-medium tuition at the undergraduate level. Recent research into the teaching of the language in secondary education has shown that:

- An environment conducive to language learning has been difficult to establish in classes that often consist of 35-40 students (Kirkgoz, 2009).
- Foreign language teachers may tend to focus on the development of grammar knowledge at the expense of communicative competence in the language (Cepik and Polat, 2014).
- The absence of an English language element or version of the centralized university entrance examination undermines the importance of the language for many secondary school students (Selvi, 2014).

In addition to the centralized examination discussed in Selvi (2014), which acts as an overall screening exam for both Turkish-medium and English-medium universities in Turkey, English-medium universities need to collect information on prospective students' English proficiency. This is typically achieved with an institutional admission English language test that assesses test taker ability to use English at a level that is adequate to follow undergraduate instruction within the university. University admission tests are high stakes as failure to achieve a sufficient score can limit prospective students' educational and professional lives.

1.2 Assessing speaking ability for an English-medium educational context

Developers of tests need to provide evidence that their testing methodology is valid. This is especially the case when the test stakes are high. Our contemporary concept of validity is that it resides in the inferences we are able to make about a test taker's underlying language ability from their performance on a language test (Messick, 1989, Fulcher and Davidson, 2007, Weir 2005). In order for university admission language testing to be valid, it must provide evidence about prospective students' ability to perform in an English-medium, higher educational environment: the 'target language use domain' (Purpura, 2016, p. 193). The ability to speak in the language of instruction is an important skill for undergraduate students. During undergraduate study, students utilise second language (L2) English to participate in seminar discussions, meet with instructors and deliver presentations. Speech production is thus an essential component of language assessment for purposes of English-medium university admission.

The test takers that sit the university admission test come from a secondary school environment in which academic language is not explicitly taught. As a result, the content of the speaking test involves accessible topics that are likely to be familiar to test takers (e.g. test takers are typically required to discuss personal interests, family, experiences, travel, and current affairs) and does not presume any background knowledge with academic conventions. Test takers that pass the admission test receive in-sessional language support through the university's academic English courses. The admission test therefore measures the ability to express relatively simple ideas in the L2 to determine the extent to which test takers will be able to discuss the more complicated concepts they encounter during the early stages of undergraduate study (see Taylor, 2011 on similar uses of the Cambridge ESOL FCE and CAE exams).

1.3 Making a case for planning in speech assessment

Many speaking tasks in an academic context involve some form of planning (Wigglesworth and Elder, 2010). This is especially the case for many of the monologue tasks that are required at the undergraduate level (e.g. academic presentations). This has implications for the design of speaking test tasks. To represent the domain of use, language tests for university admission should include speaking tasks that involve a period of pre-task planning (Skehan, 1998).

The inclusion of planning for a speaking test task is compatible with socio-cognitive validity theory (Weir, 2005): the context element, the extent that the test is representative of the real world tasks that the test taker will encounter beyond the test, and the theory based element (Weir, 2005), now termed the cognitive element (Field, 2011, O'Sullivan and Weir, 2011), the extent to which a language test task elicits the cognitive processes that would occur if the task were performed in a naturalistic setting.

In making a case for the inclusion of pre-task planning time, an important concept is Swain's (1985, p. 42) argument that language tests should 'bias for best'. Developers of language tests should create conditions that allow test takers to produce their best possible performance: bias for the best performance. The literature indicates that planning before a language task is beneficial to the process of second language speech production (Ellis, 2005, O'Sullivan, 2012, Robinson, 2005, Skehan, 2016). There is thus a compelling argument for planning to be included as part of speaking tests for test takers to demonstrate their full capabilities: 'if we add it, performance improves; remove it or reduce it, and performance worsens' (O'Sullivan, 2012, p. 235). Biasing for best is particularly important in the Turkish higher educational context, where the educational conditions are not conducive to the development of English spoken ability and testing is particularly stressful.

In sum, pre-task planning should be included as part of a speaking test task for reasons of test validity and to bias for best. However, research is yet to provide a clear account of the impact on test scores of including a period of pre-task planning. Empirical research findings in language testing are inconsistent but generally indicate that planning may not have a large impact on test scores (Wigglesworth, 1997, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006). This is in stark contrast to findings reported in task-based language teaching (TBLT) where pedagogically oriented research has consistently demonstrated impacts of pre-task planning on task performance.

1.4 Thesis outline

This thesis examines the impact of including pre-task planning time on the results of a test of second language speaking for the purposes of admission to an English-medium university. This first chapter has described the research context and provided a rationale for the research. The following section outlines the structure of the thesis. A brief synopsis of each chapter is presented and the key themes are set out.

Chapter two is the literature review. It summarises research in both task-based language teaching (TBLT) and language testing that has focussed on the impact of pre-task planning on second language speech performance. Following this, the chapter presents Weir's (2005) socio-cognitive framework as the organising principle of the study and describes the kinds of evidence that need to be provided by developers of language tests to demonstrate validity. This evidence is categorised as evidence of context, cognitive, and scoring elements of validity. These sources of validity evidence are used in the chapter to interpret the research in both TBLT and language testing that seeks to establish the effect of pre-task planning on second language speech production. The literature review identifies four sources of variation between the research studies that may account for differences in the reported impact of pre-task planning on task results. These sources are task type, planning time, test taker proficiency, and measurement. Following the literature review, chapter three summarises the key themes of the research and states the research questions.

Chapter four describes the methodology and results of two pilot studies. The chapter presents information about research participants and sets out the procedures for the collection of data and the approaches to data analysis. Information relating to the development of two EBB rating scales (*empirically* derived, *binary* choice rating scales that describe *boundaries* between performance levels; Turner and Upshur, 1996), the application of an analytic rating scale (Iwashita, McNamara, and Elder, 2001) and the multi-faceted Rasch measurement (MFRM) of the rating scale results is described in detail. The chapter also describes the development and application of measures of complexity, accuracy and fluency (CAF). Following the presentation and discussion of the pilot study results, the implications of the pilot study findings for the main study methodology are stated.

Chapter five describes the main study research methodology. This chapter begins by explaining the rationale of the study and restates the research questions. The following sections describe the quantitative research methods and the analytical procedures adopted in the research, and justify these choices with reference to the literature review and results of piloting. The chapter includes information about the research participants, data collection procedures, rating scale development and rater training, transcription of test taker speech, CAF measures and statistical procedures.

Chapter six describes the results of the main study. The chapter begins by presenting the results of MFRM of the scores awarded to the test takers by raters working with two rating scales: an EBB rating scale (Turner and Upshur, 1996) and an analytic rating scale (Iwashita et al., 2001). This section describes the impact of variation in pre-task planning time on test scores. The results of a series of MFRM analyses that investigate the independent variables, task type, planning time, test taker proficiency, and rating scale are presented in the next section. Following this, the results of an analysis of the test takers' test transcripts using measures of CAF are presented. This section examines the impact of interactions between planning time, task type, and language proficiency on CAF results.

Chapter seven discusses the results of the main study. The chapter is separated into six sections, which correspond to the research questions. The results of the main study are discussed and interpreted with reference to the literature review. The chapter analyses transcript samples to provide examples of the quantitative findings and interprets the impact of pre-task planning on test scores and CAF measures in light of this analysis.

Chapter eight is the conclusion. The chapter summarises the purpose of the study and restates the research findings. Following this, the chapter describes the implications and contributions of the research. This section is divided into implications and contributions to language testing, task-based language teaching, and to the understanding of language learners/test takers. The limitations of the study are then set out and areas for future research are stated. The chapter concludes by offering final comments on the study.

2 Literature Review

2.1 Introduction

This chapter reviews the literature relating to pre-task planning time and its relation to quality of test performance. The review first presents the research findings reported in the language testing and task-based language teaching (TBLT) literature and discusses common research methodology in the study of pre-task planning (Section 2.2). Weir's (2005) socio-cognitive framework is then set out as the organising principle of the study (Section 2.3). Following this, the review is separated into four sections which discuss key considerations that need to be made when introducing a period of pre-task planning time to a test of second language speaking. This review demonstrates that there are considerable shortcomings in the existing research with regards to the way planning impacts have been evaluated in the fields of both language testing and TBLT.

Section 2.4 describes the importance of accounting for test taker characteristics, such as emotional state and exam familiarity, in the development of speaking tests (O'Sullivan, 2000a, Weir, 2005). It focuses on experiential and psychological influences on test taker performance in language tests. The section discusses the educational environment in which the study was conducted and describes research that has investigated the interaction between working memory capacity and pre-task planning.

Section 2.5 discusses the context element of validity (Weir, 2005). This section examines the literature for reports of interactions between task demands and pre-task planning and discusses the concept of task challenge and its relevance in pre-task planning research. It identifies picture-based narrative tasks and non-picture-based description tasks as important task types for pre-task planning research and discusses issues of validity involved in the use of these tasks. Secondly, the section discusses the amount of pre-task planning time that has been investigated in research studies. It identifies four periods of planning time that are common in the literature and have been shown to affect second language speech production in different ways. These periods are 30 seconds, one minute, five minutes and ten minutes.

Section 2.6 discusses the cognitive element validity (Field, 2011, O'Sullivan and Weir, 2011). This section describes relevant theory about speech production and the development of second language proficiency. It examines the role that proficiency plays in pre-task planning and identifies the proficiency of the test taking population as a core variable in the pre-task planning research. The section reviews research that has compared the impact of pre-task planning between different proficiency groups.

Section 2.7 discusses the scoring element of validity (Weir, 2005). This section describes the dominant approaches to the assessment of language performance in the pre-task planning literature. The section describes two methods of assessment a) measures of complexity, accuracy and fluency (CAF) and b) scores awarded using a rating scale. The two methods have generated different results concerning the impact of planning on speech performance and the section therefore posits measurement as a key issue in pre-task planning research. The measurement of complexity, accuracy and fluency is first described. Following this, approaches to rating scale development are discussed and the use of *empirically* derived, *binary* choice rating scales that describe *boundaries* between performance levels (EBB scales) for purposes of language testing is examined.

2.2 Planning time in Language Testing and TBLT

Studies that have investigated pre-task planning time as a variable in second language testing research have produced mixed results (Wigglesworth, 1997, Elder et al., 2002, Elder and Iwashita, 2005, Tavakoli and Skehan, 2005, Elder and Wigglesworth, 2006, Weir et al., 2006, Xi, 2005, 2010, Nitta and Nakatsuhara, 2014, Li et al., 2014). This body of research spans two decades and has been produced in a variety of contexts. A summary of the research is presented in Table 1.

The table reports the country in which the research was conducted. This is important information as learning English as a second language (e.g. in Australia, the UK, or the USA) affords more opportunities to use the language outside the language classroom and develop proficiency as a second language speaker than learning English as a foreign language (e.g. in Turkey, Iran, Japan, and China). The table contains information about the number of participants, their first language backgrounds, and their levels of proficiency in English. Proficiency is reported as a result on a language test (e.g. TOEFL, IELTS) or common reference level such as those described in the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001) where possible. However, when this information is not provided, the terms used by the researchers are reported (e.g. Tavakoli and Skehan, 2005).

The task types used in the studies are indicated (task type is discussed in detail in Section 2.5.1.2 and Section 2.5.1.3). The amount of planning time used in each study is reported in the following column (see Section 2.5.2). Cells in this column that contain more than one amount of time indicate that the research compared the effects of different amounts of planning time. The following column shows whether guidance was offered to test takers during pre-task planning: crosses indicate that no guidance on planning was provided.

The statistical approach to the analysis of test scores is reported in the next column. *Facets* (Linacre, 1989) is software used to perform multi-faceted Rasch measurement (MFRM). This column also indicates whether an adjusted critical significance level (alpha) was used to account for multiple statistical tests on CAF results. Using a critical significance level of .05, there is 5 per cent chance (i.e. a chance of 1 in 20) of committing a type one error: failing to reject the null hypothesis when it is true (Brown, 1990, Panchin and Tuzhikov, 2017, Siegel, 1990). The chance of committing a type one error increases with every test that is simultaneously conducted. Brown (1990, p. 771) demonstrates this using the following formula (c is the number of statistical tests):

$$1 - (1 - \alpha)^c$$

For instance, using a critical significance level of .05 to test for statistically significant differences between two amounts of planning time on the results of six CAF measures, the chance of committing a type one error is 26 per cent. Spurious significant results are more likely to occur when an unadjusted alpha is used. This is an important limitation of the CAF analysis in the pre-task planning literature that means reported results must be interpreted with caution (see Section 2.7.2.4).

The final columns report the effect of the planning variable on the test scores. The effect of planning is reported in terms of measures of CAF (see Section 2.7.2) and on test scores generated through rater assessments. Ticks refer to statistically significant impacts of pre-task planning on the results of the analysis. N/A indicates that either CAF or rater assessment was not investigated in the study.

As the table shows there is variation between the studies in terms of research methods and results. The majority of the studies were conducted in English as a second language contexts with language learners from diverse L1 backgrounds. Sample sizes range from relatively small (e.g. 32 participants in Nitta and Nakatsuhara, 2014) to large (e.g. 236 participants in Xi, 2005, 2010). *Facets* is used to perform MFRM on the rater results, with the exception of Xi (2005, 2010) who uses structural equation modelling. When CAF measures are used, the statistical analysis is generally performed with an analysis of variance (ANOVA) or multivariate analysis of variance (MANOVA) to account for multiple variables in the research design. Many of the studies investigate the impact of pre-task planning on both measures of CAF and test scores. The two approaches typically record similar overall results (i.e. planning impacts the results of both analytical approaches), although Wigglesworth (1997) finds statistically significant effects of planning on CAF but not test scores. Overall, pre-task planning impacted the results of Xi (2005, 2010), Tavakoli and Skehan (2005), Li, Chen and Sun (2014), Weir, O'Sullivan and

Horai (2006), Nitta and Nakatsuhara (2014), whereas Iwashita, McNamara and Elder (2001), Elder, Iwashita and McNamara (2002), Elder and Iwashita (2005), and Elder and Wigglesworth (2006) report no impact of the planning variable.

Table 1 Language testing research into task planning

	Setting	Proficiency	Participants	L1	Task type	Planning	Guidance	Analysis	CAF	Rater
Wigglesworth (1997)	Australia	ACCESS Low: 2 High: 3	107	Various	Picture description/comparison, conversation summary, telephone answering message, graphical description/ discussion	1 min.	✗	*Facets*/Chi-square (unadjusted alpha)	✓	✗
Iwashita et al. (2001)	Australia	TOEFL 427-670	193	Various	Picture based narrative	3 min.	✗	*Facets*/ MANOVA (unadjusted alpha)	✗	✗
Elder et al. (2002)	Australia	TOEFL 427-670	201	Various	Picture based narrative	3 min.	✗	*Facets*	NA	✗
Elder and Iwashita (2005)	Australia	TOEFL 427-670	197	Various	Picture based narrative	3 min.	✗	*Facets*/ ANOVA (unadjusted alpha)	✗	✗
Tavakoli and Skehan (2005)	Iran	Elementary/Intermediate	80	Farsi	Picture based narrative	5 min.	✗	ANOVA/ T-test (unadjusted alpha)	✓	NA
Elder and Wigglesworth (2006)	Australia	Advanced/Intermediate	90	Various	Descriptive monologue	A.1 min. B. 2 min.	✗	*Facets*	A.✗ B. ✗	A.✗ B. ✗
Weir et al. (2006)	UK	IELTS High: 6.5 Borderline: 6.0-6.5 Low: Below 6.0	74	Various	Descriptive monologue	1 min.	✗	Correlation analysis/ ANOVA	NA	✓
Xi (2005/2010)	USA	Not mentioned	236	Various	Graph description	1 min.	✗	Structural Equation Modeling	NA	✓
Nitta and Nakatsuhara (2014)	Japan	B1	32	Japanese	Discussion	3 min.	✗	*Facets*/ T-test (unadjusted alpha)	✓	✓
Li et al. (2014)	China	Intermediate	95	Mandarin	Opinion description	A.30 sec. B.1 min. C.2 min. D.3 min. E.5 min.	✗	MANOVA (Bonferroni correction)	✓	NA

*CAF/ Rater columns ✓= statistically significant differences in scores after planning, ✗ = no statistically significant differences in scores, NA = the research did not feature CAF or Raters.

Table 2 summarises the research findings reported in the TBLT studies. The format is the same as for Table 1 and reports information about the research context, participants, task type, planning time, planning guidance, analytical procedure and planning effects.

The table shows that the research was conducted in a range of national settings including English as a second language contexts (e.g. the UK, Australia, and the USA) and English as a foreign language contexts (e.g. Thailand, Spain, Turkey, Japan, Brazil). The numbers of participants involved in the study range from 17 in Ellis (1987) to 61 in Skehan and Foster (2005). Participants come from a range of L1 backgrounds. In terms of English proficiency, the typical level is described as 'intermediate'. Three studies report proficiency as a score on a test: Yuan and Ellis (2003), Kawauchi (2005), and Mochizuki and Ortega (2008). The majority of the studies use impressionistic terminology such as intermediate, pre-intermediate and advanced. This terminology makes a minimal contribution to the research because the levels are not standardised and there may be little overlap between what is referred to as an 'intermediate' level in different studies.

Task types vary between the studies but a common task is the picture-based narrative task. Ten minutes planning time is frequently used, although some research investigates the impacts of smaller amounts of planning (e.g., Mochizuki and Ortega, 2008, Sasayama and Izumi, 2012) and larger amounts (e.g. Ellis, 1987, Sangarun, 2005). Guidance during planning was investigated in a number of studies (Foster and Skehan, 1996, Kawauchi, 2005, Mochizuki and Ortega, 2008, Sangarun, 2005). In these studies, the guidance involved teacher led tuition of task relevant language.

Variation in the focus and the delivery of the guidance is indicated in Kawauchi (2005) and Sangarun (2005).

The statistical analysis is presented in the following column. ANOVA and MANOVA are frequently used to investigate the impact of multiple variables in the research design (e.g. type of guidance, task type). This column also reports whether the critical significance (alpha) level was adjusted for multiple statistical tests on CAF results. As the final column shows, the effect of the planning variable is consistently reported as positive across the studies. This is despite substantial variation between the studies in terms of the research methods, participants and analytical procedure. However, the widespread absence of alpha correction in both the language testing and TBLT studies is an important limitation that calls in to question the validity of conclusions regarding the impact of pre-task planning on CAF measures (see Section 2.7.2.4).

Table 2 TBLT research into task planning (continued on page 18)

	Setting	Proficiency	Participants	L1	Task type	Planning Time	Guidance	Analysis	Result (CAF)
Ellis (1987)	UK	Pre-intermediate	17	Various	Picture based narrative	1 hour	✗	Chi-square (unadjusted alpha)	✓
Crookes (1989)	Japan	Intermediate/Advanced	40	Japanese	Lego construction Map description	10 minutes	✗	MANOVA (unadjusted alpha)	✓
Foster and Skehan (1996)	UK	Pre-Intermediate	32	Various	Personal Picture based narrative Decision	10 minutes	1.✗ 2. ✓	ANOVA (unadjusted alpha)	✓
Skehan and Foster (1997)	UK	Pre-Intermediate	40	Various	Personal Picture based narrative Decision	10 minutes	✗	ANOVA (unadjusted alpha)	✓
Foster and Skehan (1999)	UK	Intermediate	66	Various	Discussion	10 minutes	✓	ANOVA (unadjusted alpha)	✓
Yuan and Ellis (2003)	China	TOEFL 373-520	42	Mandarin	Picture based narrative	10 minutes and online planning	✓	ANOVA (unadjusted alpha)	✓
Kawauchi (2005)	Japan and UK	Low/High/Advanced TOEFL 420-610	40	Japanese	Picture based narrative	10 minutes	1.Rehearsal 2.Writing 3.Reading	ANOVA (Bonferroni Correction)	✓
Sangarun (2005)	Thailand	Intermediate	40	Thai	Instruction Argument	15 minutes	1.Meaning 2.Form 3.Meaning and form	ANOVA (unadjusted alpha)	✓
Skehan and Foster (2005)	UK	Intermediate	61	Various	Discussion	10 minutes	1.✓ 2. ✓	ANOVA (unadjusted alpha)	✓
Philp et al. (2006)	Australia	Intermediate	42	Various	Information gap, picture description	A.2 minutes B.5 minutes	✗	Friedman test (Bonferroni correction)	✓
Gilabert (2007)	Spain	Low Intermediate	48	Spanish	Picture based narrative	10 minutes	✗	ANOVA (unadjusted alpha)	✓
Mochizuki and Ortega (2008)	Japan	STEP results equivalent to TOEFL 360 - 420	56	Japanese	Picture based narrative	5 minutes	1.✓ 2.✗	MANOVA (unadjusted alpha)	✓
Guara-Tavares (2009)	Brazil	Intermediate	25	Portuguese	Picture based narrative	10 minutes	✗	ANOVA (unadjusted alpha)	✓
Sasayama and Izumi (2012)	Japan	'generally limited' (2012, p. 29)	23	Japanese	Picture based narrative	5 minutes	✗	ANOVA (unadjusted alpha)	✓

Genc (2012)	Turkey	Low-intermediate	60	Turkish	Picture based narrative	10 minutes	✗	t-test (unadjusted alpha)	✗
Geng and Ferguson (2013)	UK	Upper Intermediate	32	Various	Decision making Information-exchange	10 minutes	✓	MANOVA ANOVA (unadjusted alpha)	✓
Nielson (2013)	USA	Intermediate	40	Various	Picture based narrative	10 minutes	✗	ANCOVA MANCOVA (unadjusted alpha)	✓
Pang and Skehan (2014)	Macao	Low/High intermediate	48	Mandarin and Cantonese	Picture based narrative	10 minutes	✗	Descriptive statistics	✓
Bui and Huang (2016)	Hong Kong	Upper Intermediate (B2)	58	Cantonese	Description	10 minutes	✗	ANOVA (unadjusted alpha)	✓

*Continuation of Table 2

*CAF column ✓= statistically significant differences in scores after planning, ✗ = no statistically significant differences in scores.

2.2.1 What do the research findings indicate about pre-task planning?

Comparison of Tables 1 and 2 reveals that the effect of pre-task planning varies substantially between studies with a TBLT focus and studies with a language testing focus. The cause of this variation may be attributable to four main differences between the studies; test taker characteristics, task type, planning time, and measurement.

To outline the first source of variation, in the research test takers come from a range of educational backgrounds. While many of the studies were completed in an English as a foreign language context (e.g. in China, Thailand, or Spain) the vast majority were conducted in English as a second language context (e.g. in the UK, the USA, or Australia). The educational context dictates the frequency with which the language can be used outside the language classroom and the range of communicative experience the participants bring to the study. Language proficiency may also be a key difference between the studies but this is difficult to gauge from the widespread use of vague terminology such as 'intermediate' (Foster and Skehan, 1999).

Another source of variation between the studies is task type. A range of task types has been used to investigate pre-task planning. The majority of the tasks are monologue, although dialogue tasks have also been investigated (e.g. Nitta and Naktsuhara, 2014). Picture-based narrative tasks are common in both the language testing and TBLT research. Another common task type is to provide personal information (e.g. Foster and Skehan, 1996) or give an opinion on a topic (Li et al., 2014). While the impact of planning for such tasks is typically positive in TBLT, studies with a language testing focus have demonstrated less consistent effects.

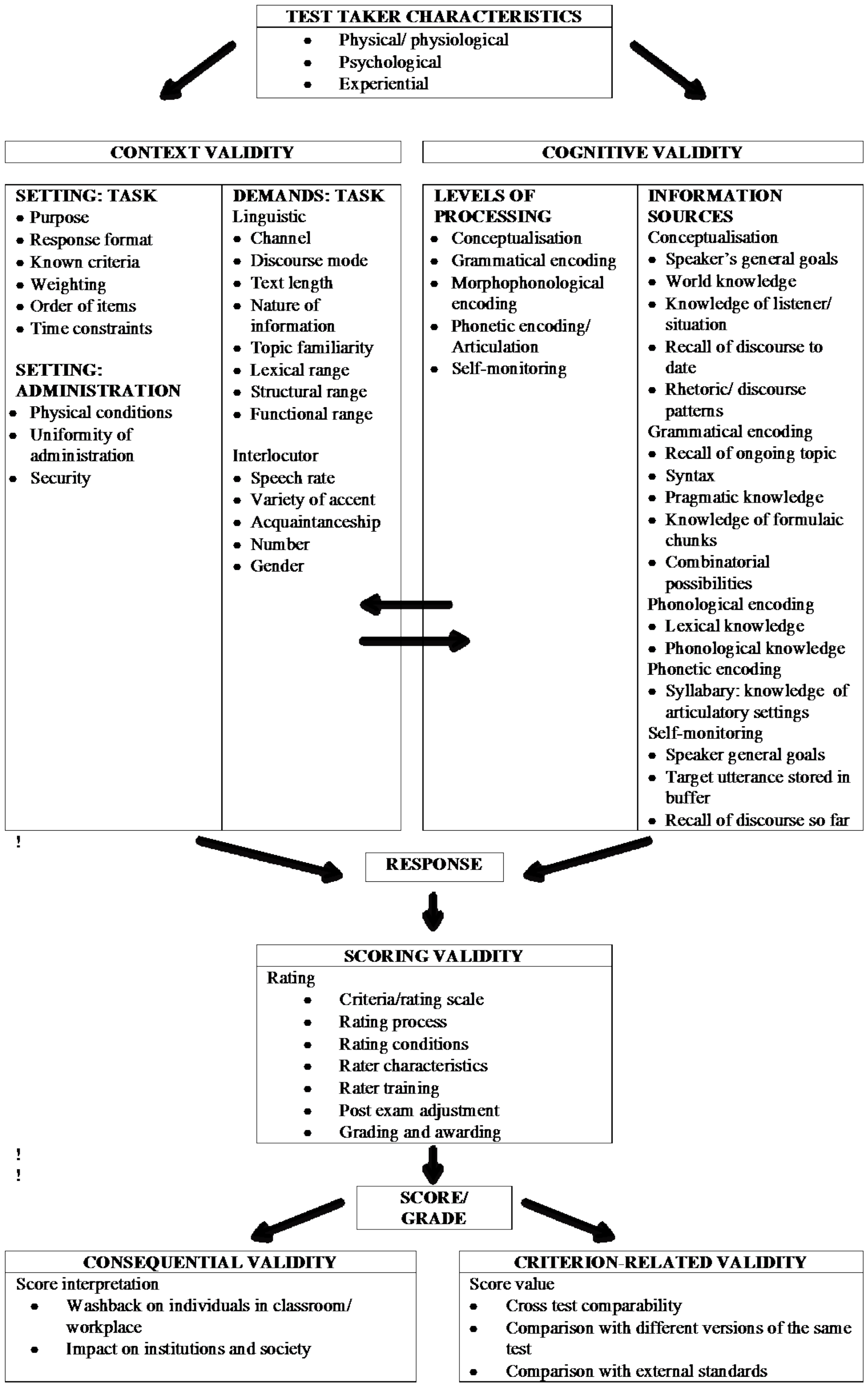
The third source of variation is the amount of pre-task planning time investigated in the research. The most common amount of planning time in TBLT is ten minutes. In contrast, studies with a language testing focus generally involve less planning time, typically from one minute to three minutes. Planning for ten minutes has proved consistently effective in the TBLT literature, whereas less planning time has produced conflicting results in the language testing literature.

The final source of variation between the studies is measurement of task performance. In TBLT the impact of pre-task planning is assessed exclusively with measures of complexity, accuracy and fluency (CAF) (see Section 2.7.2). Language testing studies use rating scales to measure speech performance (see Section 2.7.3) or a combination of a rating scale and measures of CAF. Tables 1 and 2 show that the addition of planning time to a language task has predominantly affected CAF measures. Rating scales have proved less successful in capturing an effect of planning on test scores (Wigglesworth, 1997, Elder et al., 2002, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006).

2.3 The socio-cognitive framework

The organising principle of this study is the socio-cognitive framework proposed by Weir (2005) and adapted by Taylor (2011, p. 28) to further explore L2 speech assessment (see Figure 1). The socio-cognitive framework specifies the types of evidence that are required in the process of language test validation. The framework is *socio-cognitive* because language ability is a cognitive skill that is used to fulfil social purposes (Taylor, 2011). The socio-cognitive framework is widely used to investigate evidence of language test validity, most notably by Cambridge English Language Assessment (Weir and Taylor, 2011). The framework provides a suitable organising principle to investigate issues relating to the implementation of pre-task planning in a test of second language speech.

Figure 1. A framework for conceptualising speaking test validity (Taylor, 2011, p. 28, adapted from Weir, 2005, p. 46)



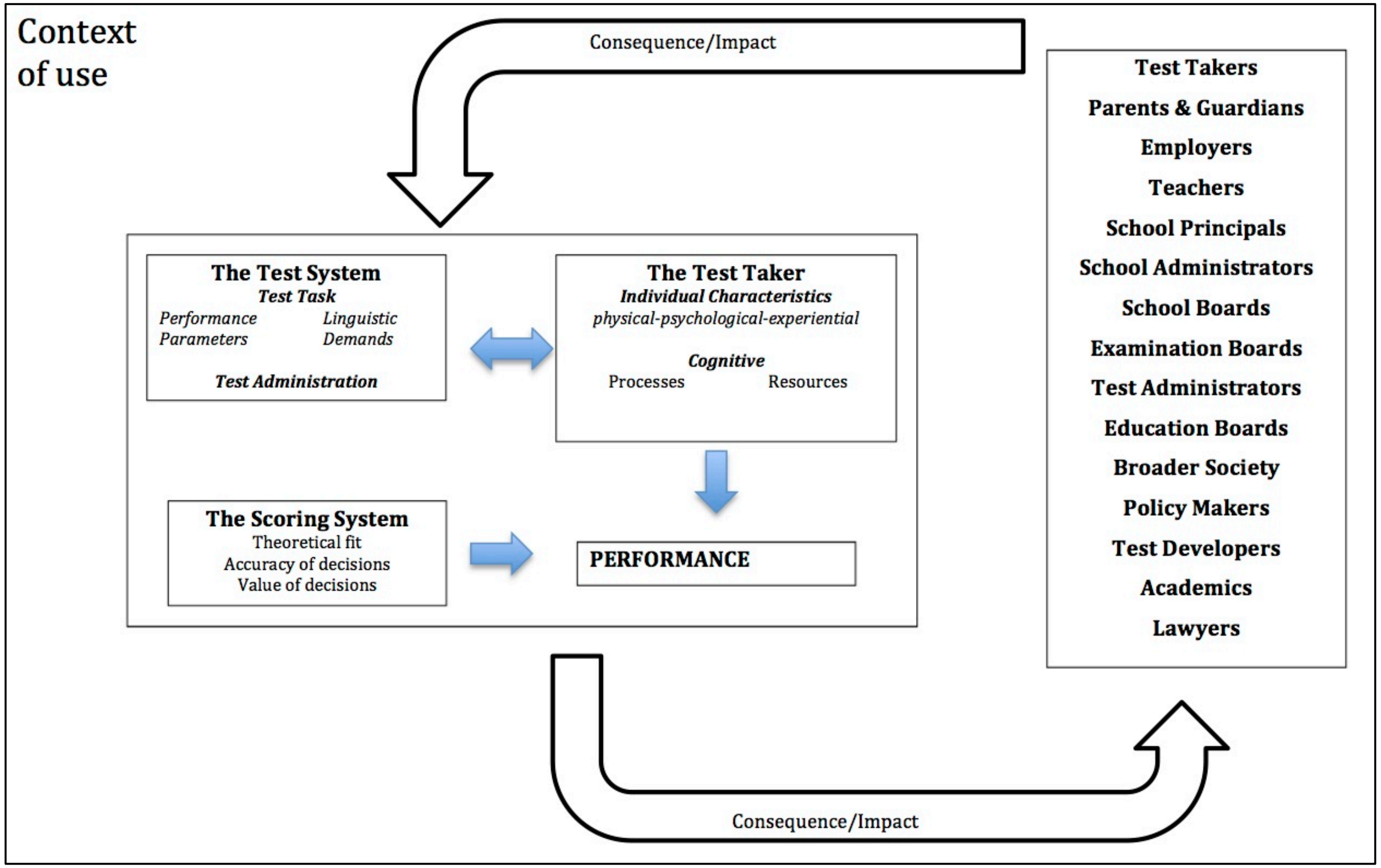
Weir (2005, p. 43) stipulates that evidence of validity should be collected at two stages: before the test '*a-priori*' and after the test '*a-posteriori*'. *A-priori* evidence relates to the extent to which test taker characteristics are accounted for by the test developer (O'Sullivan, 2000a, O'Sullivan and Green, 2011; see Section 2.4), and issues of context validity, and cognitive validity. Context validity is related to the extent to which the language test represents the target situation (see Section 1.2). Cognitive validity is the extent to which the test elicits the same cognitive processes that would be engaged in the completion of tasks in the target situation. Issues relating to context validity are described in Section 2.5. Issues relating to cognitive validity are described in Section 2.6. At the *a-posteriori* stage, evidence-based evaluative conclusions are made concerning scoring validity, consequential validity and criterion-related validity. Scoring validity concerns aspects of the test relating to measurement; analysis of the test scores, consistency of the grading system and the degree of error involved in the measurement. Issues relating to pre-task planning and scoring validity are discussed in Section 2.7. Consequential validity refers to the washback effect of the test: the way the test impacts individuals and society. Criterion-related validity refers to the relationship between information generated by a test and an external measure of the same ability (e.g. a different test, or a different version of the same test) and how well the test can predict future behaviour.

O'Sullivan (2016) points out that the division of validity in to different forms in the socio-cognitive framework (i.e. cognitive validity, context validity, scoring validity, consequential validity, criterion-related validity) does not accord with the commonly accepted unitary construct approach to validation (Messick, 1989). Validity resides in the inferences we are able to make about a test taker's underlying

L2 ability based on test scores (see Section 1.2). This entails that weaknesses in one aspect of the test (e.g. the target language domain is not well represented by test tasks) lead to misguided inferences regardless of the strength of the other aspects. In recognition of this shortcoming, O'Sullivan and Weir (2011) revised the original framework down to three central elements of a unitary validity argument: the test taker, the test task, and the scoring system. O'Sullivan (2016, p. 215) explains that 'the elements of the original frameworks… remain the same' in the adapted framework: 'the complexity of the original has not been undermined in any way'. However, he stipulates that validity evidence is collected during the test development stage rather than after the test has been used, i.e. at the *a priori* stage rather than *a posteriori* stage. This stipulation extends to test consequences, traditionally evaluated at the *a-posteriori* stage in the original socio-cognitive framework, which should be outlined for the various stakeholders before the test is used (see Figure 2).

Taking into account the changes in the socio-cognitive framework identified by O'Sullivan (2016) and the requirement that evidence of validity be collected during test development, this study investigates the impact of including a period of pre-task planning in a speaking test with reference to the context, cognitive and scoring aspects of the validity argument for the university admission test. Following the presentation of the findings, the potential consequences of this investigation for test stakeholders are discussed in the conclusion (see Section 8.3).

Figure 2. Revised test validation model (O'Sullivan, 2016, p. 215).



2.4 Test taker characteristics

This section outlines the importance of accounting for test taker characteristics when conducting language testing and TBLT research. Based on O'Sullivan's (2000a) PhD research, Weir (2005, p. 51) categorises the test taker characteristics that may impact the performance and results of a speaking test as 'physical/physiological', 'psychological', and 'experiential'. The physical/physiological influences on test performance may involve short-term ailments such as toothache, or long-term disabilities such as dyslexia. Experiential influences involve educational background, examination preparedness, examination experience, communication experience, and the experience of residence in the target language country. Psychological influences include personality, motivation, cognitive style, affective schemata, concentration, and emotional state. Experiential influences and the impact of these influences on test motivation and emotional state are discussed in Section 2.4.1. The impact of emotional state on test performance is also discussed in Section 2.5.1.2. Section 2.4.2 discusses psychological influences by reviewing literature that has investigated the interaction between working memory capacity and pre-task planning.

2.4.1 Experiential influences

The test takers that sit the English university admission test typically come from an educational background where English is taught as a foreign language (see Section 1.1). Opportunities to communicate in the L2 outside classrooms are limited. In addition, the centralized university exam in Turkey does not contain an English component and this undermines the importance of the language for many secondary level students (Selvi, 2014). As a result, students leave secondary education with limited proficiency in the target language. The experience of having their speaking ability tested in a high stakes English language test is naturally stressful as test takers have not had the opportunity to develop their speaking skills to a level where they can demonstrate communicative proficiency. In short, prospective students commonly approach the speaking section of the university admission test with anxiety.

The level of communicative experience is a key difference between the participants in this study and those investigated in much of the literature. Much of the pre-task planning research has been conducted in an English as a second language context where language learners also live in the target language environment and commonly use the L2 in their daily lives (e.g. in the UK, Australia, or the USA), or with participants who had been following English-medium instruction for a number of years at the time of the study. The experience of routinely producing the L2 means that the skill is well practiced and language learners are better able to deal with speaking demands (O'Sullivan and Green, 2011). In the current study, the test takers' communicative experience is limited and this may have an important bearing on the impact of pre-task planning (see Section 2.6.2).

2.4.2 Psychological influences

The current section focuses on research into the interaction between pre-task planning time and working memory capacity (Guara-Tavares, 2009, Nielson, 2013). Guara-Tavares (2009, p. 166-7) argued that the outcome of planning for a language task is dependent upon the ‘ability to actually retrieve what was planned into on-line performance’. Variation in working memory capacity is an important factor in this process. Test takers with stronger working memory should be more able to recall the information they planned when completing the language task. However, Guara-Tavares’ (2009) and Nielson’s (2013) research findings showed that although planning did lead to overall gains in measures of CAF, the size of the gains was not impacted by variation in working memory capacity. Nielson (2013, p. 287) speculates that the language tasks used in her study (‘simple’ picture-based narratives) played an important role in this result; differences between the two working memory groups may have emerged after planning if participants had completed challenging tasks. The interaction between task type, planning time and test taker characteristics is an important focus of this study. This interaction is discussed in detail in the following section under the heading the context element of validity.

2.5 The context element of validity

The context element of validity refers to the extent to which the language test accounts for ‘the social dimension of use’ in the target situation (Weir, 2005, p. 19). For a speech assessment task to be contextually valid it must represent the kinds of tasks that test-takers are expected to perform in the target situation or ‘target language use domain’ (Purpura, 2016, p. 193). Reproducing the target situation *in toto* in a language test is not an achievable goal (Weir, 2005). However, the test developer should make efforts to ensure that the test tasks and the test setting are as realistic as possible. During undergraduate study, the target language use situation to which the admission test relates, students are required to participate in a range of situations involving different opportunities to plan speech (Wigglesworth and Elder, 2010). Planning is an important process with potential to determine a student's academic success (undergraduate course grades are more likely to be determined by performance in planned monologues such as academic presentations than in situations where students produce speech spontaneously such as in seminar discussions). In order for the English university admission test to appropriately represent the target situation - to be contextually valid - it must reflect the opportunities for planning that are common in undergraduate study: 'if the test tasks reflect real life tasks in terms of important contextually appropriate conditions and operations it is easier to state what a student can do through the medium of English' (Weir, 1993, p. 28).

2.5.1 Tasks

Language tests are typically made up of a series of tasks that assess different skills, e.g. in tests of English for academic purposes, the ability to take part in a discussion in the second language is often tested through dialogue tasks, the ability to produce a presentation in the second language is tested through monologue tasks (Weir, 2005). A task type commonly used in the literature that has consistently proved amenable to planning is the picture-based narrative (Foster and Skehan, 1996, Nielson, 2013, Pang and Skehan, 2014, Tavakoli and Skehan, 2005). This section discusses whether this task type can be regarded as contextually valid for the university admission English test.

Inoue (2013) suggests that although relating a narrative from a series of pictures is unlikely to be required in most everyday situations (including those encountered in undergraduate study), picture-based narrative tasks do test the test takers' ability to report events. This is certainly a type of discourse that would occur in a university setting, where students are required to use the L2 to present a sequence of events in a research report or summary. In addition to this, Inoue argued that the use of picture-based narratives increases test reliability by constraining the range of topics that can be discussed so that task content is similar for all test takers. Fulcher (2003, p. 70) has stated that picture-based narrative tasks are particularly useful for testing less proficient learners because 'telling simple stories is one of the first things that they are able to do in a second language'. Limitations in the levels of language proficiency within the test-taking population need to be accounted for in this study (see Section 2.4.1). This requirement extends to task selection. In sum, the picture-based narrative task type can be regarded as suitable for this study's test taking population.

On the other hand, Skehan (2009) has argued that standardising the content of the task by using pictures to elicit speech may increase task difficulty, i.e. when test takers do not have the requisite lexis to describe the images (see Section 2.5.1.2). Non-picture-based tasks involving a description of events from experience also tap the ability to report without imposing constraints on the content that is described. They constitute a valid alternative to picture-based narrative tasks that also tests the test taker's ability to report events. This study compares the impact of planning for picture-based narrative tasks and non-picture-based description tasks.

2.5.1.1 Task demands

Weir (2005) categorises task demands in terms of linguistic and interlocutor characteristics. Linguistic demands relate to task characteristics such as the range of lexis and grammatical structures required to complete the task. Interlocutor characteristics relate to variables such as the number of interlocutors involved in the task, their gender and their level of acquaintanceship. This section discusses the task demands that have been investigated in the pre-task planning literature and identifies the specific task characteristics that have proved amenable to planning. Identifying these characteristics is important because in addition to demonstrating the context element of validity, the university admission test should also be shown to bias for the best performance (Swain, 1985). The section begins by discussing task demands in TBLT research (see Section 2.5.1.2) and then discusses task demands in language testing research (see Section 2.5.1.3).

2.5.1.2 Task demands in TBLT

In the TBLT literature, task demands are generally evaluated using models proposed by Skehan (2009) and Robinson (2005). Two influential proponents of the TBLT research approach, Skehan and Robinson argue that task demands are caused by increases in the cognitive processing load required to complete a language task. Skehan and Robinson present their models in the form of a list of task characteristics, which when manipulated impact cognitive load in different ways. In both models, pre-task planning may reduce task challenge by creating opportunities to a) begin cognitive processes that would normally take place during the task (see Section 2.6.1) and b) rehearse planned speech.

Skehan (2009) uses Levelt's (1989) model of speech production (see Section 2.6.1) to describe how particular task characteristics may impact language learners' ability to complete speaking tasks. Skehan's model is presented in Figure 3 (Skehan, 2009, p. 52). In the model, task characteristics may put pressure on, or ease the processes of speech conception (i.e. the generation of ideas), lexical retrieval, and syntactical encoding.

In Skehan's (2009) model, planning plays both a facilitative role (e.g. it aids lexical retrieval and syntactic encoding) and a complexifying role (e.g. by extending the amount of information a speaker generates for communication). In a seminal study, Foster and Skehan (1996) demonstrated that the impact of pre-task planning varied between three tasks. The tasks were systematically designed to include 'progressively less familiar and less predictable information causing an increasingly taxing cognitive load' (1996, p. 306). Drawing from findings in cognitive psychology that attentional resources are limited in capacity (Baddeley, 2007; see Section 2.6.1), it was hypothesized that as the cognitive challenge increased, so would the requirement for attentional resources. Completing relevant processes before the task began would free up attention for use during the task. Pre-task planning would therefore have more of an impact on the tasks requiring more cognitive operations. Three language tasks were designed to test this hypothesis, these included:

- A personal information task: based on very familiar information and requiring least cognitive effort.
- A narrative task: based on visual stimuli that required encoding into linguistic form and involving increased cognitive effort.
- A decision-making task: involving the synthesis and evaluation of new information and requiring the most cognitive effort.

Foster and Skehan (1996) measured the participants' speech complexity, accuracy and fluency (CAF) as they completed the tasks under different planning conditions (a 10-minute condition and no planning). The results demonstrate that pre-task planning positively impacted levels of complexity and fluency on each task. Accuracy scores showed that pre-task planning had very minor, though statistically significant impacts on the results of the personal task and the decision task only. However, overall the largest impacts were observed on the complexity and fluency results of the narrative task. Foster and Skehan (1996, p. 316) conclude that pre-task planning 'does not operate in the same way with all tasks'. The authors suggest that 'encoding new, visual information into linguistic form' in the narrative task imposed a heavy cognitive burden that was reduced by the introduction of planning (Foster and Skehan, 1996, p. 307). This conclusion is supported in much of the TBLT research where picture-based narrative tasks are commonly used and planning leads to statistically significant differences in CAF. In short, Foster and Skehan's (1996) research findings indicate that generating language to describe a series of images in the second language is a cognitively challenging process that may be eased through pre-task planning.

In combination with planning, the type and amount of task information is predicted to have an impact on task performance (Skehan, 2009, Weir, 2005). Abstract, dynamic information is less predictable and causes more of a processing challenge than concrete, static information. The extra cognitive resources that planning makes available may be used to process abstract information that is otherwise difficult to comprehend. In contrast, as speaking about concrete information is predicted to require less cognitive effort, the extra resources that planning makes available may be spent on focus on language forms. In short, the process of attentional allocation during a speaking task varies according to the type of task information involved.

The amount of information involved in the task also plays an important role in attentional allocation (Skehan, 2009). For instance, a task that requires a range of lexis (because there is a large amount of different information to be discussed) is predicted to assert pressure on memory resources and increase task challenge. Pre-task planning time provides an opportunity to carry out unpressured lexical searches, which may have the effect of boosting complexity, accuracy and fluency. Findings in Foster and Skehan (1996), and Foster and Skehan (1999) indicate that this effect is stronger on picture-based tasks than non-picture-based tasks. Picture-based tasks contain content that must be described and this forces language learners in to a) attempting the use of potentially unfamiliar lexis or b) generating strategies such as approximation and circumlocution to complete the task. In contrast, tasks that do not contain obligatory content free up the language learner to use lexis with which they are confident. The 'non-negotiability' (Skehan, 2009, p. 5) of the task is a key source of task challenge and planning is likely to have an important impact on the completion of non-negotiable tasks.

Skehan (2009) identifies task structure as a cause of task challenge. When a task is well structured, the attention required to create connections between the various task elements is minimized. However, when the task is unstructured, attentional resources are consumed by making links between the various task elements. Empirical research has supported this claim. Tavakoli and Skehan (2005) predicted that narrative tasks that are based on a clear, sequential structure (e.g. problem and solution) would require less cognitive processing than those without a clear structure. Their findings demonstrated that structured tasks were completed with greater CAF than unstructured tasks. However, the CAF results did not reveal a statistically significant interaction between task structure and planning. Contrary to Skehan's model, pre-task planning did not impact the speakers' ability to process less structured tasks in a way that affected CAF.

Figure 3. Skehan’s model of task characteristics (2009, p. 52)

Complexifying/Pressuring		Easing/Focusing
	Conceptualizer	
• Planning: extending		• Concrete, static information
• More complex cognitive operations		• Less information
• Abstract, dynamic information		• Less complex cognitive operations
• Greater quantity of information		
	Formulator: Lemma Retrieval	
• Need for less frequent lexis		• Planning: organization of ideas
• Non-negotiability of task		• Dialogic
	Formulator: Syntactic Encoding	
• Time pressure		• Planning: rehearsing
• Heavy input presence		• Structured tasks
• Monologic		• Dialogic
		• Post-task condition

The second model that is discussed in this section was developed by Robinson (2005). Robinson argues that task demand is caused by an interaction between three factors; task complexity, task conditions and task difficulty. The task characteristics that contribute to this interaction are presented in Figure 4 (Robinson, 2005, p. 5).

Task complexity refers to cognitive factors involved in the task. Cognitive factors may be resource directing (i.e. toward language forms), or resource dispersing (i.e. dividing attentional resources). For instance, increasing the need for reasoning demands (a resource directing factor) is predicted to foster advanced language: ‘so, because, therefore’ and cognitive state verbs such as ‘think, believe, know’ involve a necessary degree of grammatical subordination (e.g. ‘she thinks that…’). In addition, increasing the number of task elements may require the speaker to elaborate their language to differentiate between various task elements (e.g. the need to distinguish characters in a narrative task might involve the use of identifying relative clauses: ‘the man that is wearing the t-shirt’). Empirical research provides mixed support for this claim.

Xi (2010) found that when a task was made complex along resource directing lines (i.e. the number of visual elements in a line graph was increased) and pre-task planning time was provided, test takers recorded higher scores on an organizational rating scale than when the task contained few visual elements. In contrast, Sasayama and Izumi (2012) did not uncover any interaction in CAF results between pre-task planning and increases in the number of characters in a narrative task. The conflicting findings seem to indicate that the scoring method adopted in each study was an important variable. Whereas Xi (2010) reported a significant impact of planning on test scores involving an organisational rating scale, Sasayama and Izumi (2012) used CAF measures that did not provide an indication of variation in the organisational features of the speech (see Section 2.7.2.4).

Robinson (2005, p. 4) suggests that tasks can be made more complex by specifying the perspective the language learner adopts: ‘here and now’ and ‘there and then’. The ‘there and then’ perspective requires reference to the past (i.e. through past tense morphology and grammatical aspect) and fosters a focus on form. This focus requires attentional resources and increases the complexity of the task. Gilabert (2007) found that speech on ‘there and then’ tasks involved higher levels of grammatical accuracy than ‘here and now’ tasks, indicating that his participants had focussed on form under the ‘there and then’ condition. In contrast, the ‘here and now’ tasks were completed with more fluency and lexical complexity. These findings suggest that having to focus on specific language forms (past reference) may decrease fluency and the use of complex lexis.

Robinson’s (2005, p. 5) concept of ‘prior knowledge’ refers to familiarity with task content. Without prior knowledge of the task content (‘background knowledge’ and ‘subject matter knowledge’; Weir, 2005, p. 75), test takers may be ‘put off by the topic’ (Lumley and O’Sullivan, 2005, p. 432), and generating information may require increased attention. Bui and Huang (2016) found that variation in language learners’ levels of familiarity with task topics was an important factor in the results of planning for language tasks. Their research involved two tasks in which participants described the processes of viral infections and treatments in the human body and in computers. The tasks were developed for undergraduate students from computer and nursing departments who had different levels of background knowledge on the task topics. Planning time was included as a between subjects variable; half of the group completed the tasks after a 10-minute planning condition and half without any planning time. The researchers analysed the participants’ speech fluency on the tasks. Results showed that pre-task planning had more of an impact when participants spoke about unfamiliar topics than when they were familiar with the task content. This indicates that planning helped compensate for limitations in the participants’ ‘affective schemata’ (Weir, 2005, p. 51) with the effect that speech fluency increased.

The second and third rows of Robinson's (2005) model describe the social factors that may impact on speech performance. In the second row of Robinson's model are task conditions. Task conditions are interactional factors and are divided into participation variables and participant variables. Most of the pre-task planning research has been conducted using monologue tasks rather than dialogue tasks (see Section 2.2). As this research also investigates the impact of planning for monologue tasks, participation variables are not discussed at length here.

O'Sullivan (2000a) and Weir (2005) describe interlocutor characteristics (referred to as 'participant variables' in Robinson, 2005, p. 5) and the influence these characteristics have on speech performance. For example, research has shown that the interaction between examiner gender and test taker nationality may be an important variable in the results of speaking assessment. In O'Sullivan (2000b) Japanese test takers received higher test scores when interviewed by female examiners, whereas in Porter (1991) Arabic test takers received higher scores when interviewed by male examiners. In addition, the degree of familiarity between interlocutors (O'Sullivan, 2002) and the extent to which one interlocutor holds more authority (i.e. in an assessor-test taker relationship, the assessor is typically an expert of the skill being tested) may influence the test taker's use of language (e.g. in terms of politeness and deference) and their levels of confidence and anxiety (Porter, 1991). No research in to the interaction between such factors and pre-task planning has been published.

The third row of Robinson's (2005) model, task difficulty, refers to learner factors that impact on task challenge and is divided into affective and ability variables. Affective variables refer to the language learner's level of anxiety, confidence and motivation for the task. In Elder and Wigglesworth (2006), participants indicated that planning decreased their levels of test anxiety and increased motivation and confidence for the test task. Instilling confidence and motivation is desirable because language tests should aim to elicit the best possible performance: 'bias for best' (Swain, 1985, p. 42; see Section 1.3). Ability variables refer to intelligence, working memory capacity and aptitude. Research has shown that participants with larger working memory capacity do record higher CAF results when completing language tasks (Guara-Tavares, 2009, Nielson, 2013). However, this research has not demonstrated an interaction between working memory capacity and planning (see Section 2.4.2).

Figure 4. Robinson's model of task characteristics (2005, p. 5)

Task complexity (cognitive factors)	*Task conditions* (interactional factors)	*Task difficulty* (learner factors)
(a) resource-directing e.g.+/-few elements +/- here and now +/- no reasoning demands	(a) participation variables e.g., open/closed one-way/two-way convergent/divergent	(a) affective variables e.g., motivation anxiety confidence
(b) resource-dispersing e.g. +/- planning +/- single task +/- prior knowledge	(b) participant variables e.g., same/different gender familiar/unfamiliar power/solidarity	(b) ability variables e.g., working memory intelligence aptitude

To summarise, empirical evidence of the interactions between planning and task characteristics proposed by Skehan (2009) and Robinson (2005) is inconclusive. This may be due to an important limitation in the TBLT approach. In the TBLT literature (e.g. Bui and Huang, 2016, Foster and Skehan 1996, Nielson, 2014, Sasayama and Izumi, 2012, Skehan and Foster, 1997) the focus of the research is the extent to which limitations in the language learner's cognitive processing capacity (Baddeley, 2007) can be overcome through systematic task manipulation (i.e. by engaging in pre-task planning) with the effect that speech production improves. This emphasis on the internal processing of the speaker means that participant characteristics and the research context are generally not well defined with the result that TBLT researchers neglect the social aspect of speech production. This approach does not fit with the socio-cognitive approach proposed by Weir (2005) and O'Sullivan (2016) that posits cognition as situated and thus influenced by contextual factors. While Robinson's (2005) model makes efforts to define the interactional and learner factors that influence task completion, the following section (2.5.1.3) shows that both Skehan's and Robinson's models fall short of an adequate framework for evaluating task performance for purposes of language test development.

Despite limitations in the TBLT approach, the research indicates that the extent to which a task obliges test takers to discuss unfamiliar topics may be an important source of task challenge (Bui and Huang, 2016, Foster and Skehan, 1996, 1999). This indicates that task challenge is not a property of a language task but rather a result of the interaction between the task and test taker characteristics such as familiarity with task content (O'Sullivan, 2000a, Weir, 2005). Planning compensates for limited background knowledge and allows test takers to generate language to discuss obligatory task contents.

2.5.1.3 Task demands in Language Testing

Application of the cognitive models in language testing research has produced results that offer little support to the claims made by Skehan (2009) and Robinson (2005). Iwashita, et al. (2001) found a minimal impact of task manipulation along the lines identified in the two models in a research study that involved both measures of CAF and scores on an analytic rating scale comprising descriptors of complexity, accuracy and fluency (see Section 2.7.3). Iwashita et al. (2001, pp. 414-415) operationalized task challenge as variation in demands associated with 'perspective' (telling the story from one's own perspective, and telling the story from someone else's perspective), 'immediacy' (telling the story directly from a series of images, and telling the story from memory), 'adequacy' (telling the story from a complete set of images, and telling the story from an incomplete set of images) and 'planning' (three minutes and 30 seconds pre-task planning, and 30 seconds pre-task planning). The results showed that only the immediacy condition impacted test scores and the effect was in the opposite direction from the prediction made by the researchers. Telling the narrative from memory (i.e. without the pictures) increased test scores (the score difference is not reported) and accuracy measures (from 67.66 per cent to 74.26 per cent error free clauses). By way of conclusion the authors discuss the possibility that, 'speaking is more of a unitary skill, deployed in similar ways across different task conditions' (2001, p. 428). However, they are quick to acknowledge that this interpretation does not account for previous findings in the literature (TBLT).

Khabbazbashi (2017) found that variation in topic familiarity did have a statistically significant impact on scores of the IELTS speaking test but that this impact was not large enough to have a practical, meaningful effect on the test scores. Test takers were placed into the same band on the IELTS rating scale regardless of their familiarity with the task topic. Khabbazbashi concludes that differences between tasks do not automatically lead to differences in test scores. This is an important observation for this study as it suggests that different task characteristics may not have as much of an influence on test scores as they do on CAF results (see Section 2.7.2.4).

The influence of the test taker's cultural background on task performance is not accounted for in the cognitive models of task characteristics (Skehan, 2009, Robinson, 2005). It has been suggested that test takers may find a task challenging if it is based on culturally unfamiliar information or requires culturally unfamiliar procedures (Fulcher, 2003, Weir, 2005). Fulcher and Marquez Reiter (2003) investigated the impact of test takers' cultural background on speaking test performance. The test required test takers to make requests involving different degrees of imposition (e.g. from borrowing a book to borrowing a laptop). The level of imposition was predicted to vary according to the participants' cultural backgrounds (English and Spanish). The findings suggest that test taker variables such as cultural background may cause deviations from expected routines (e.g. not acknowledging the magnitude of a request), which may have a bearing on test scores.

Fulcher (2003, p. 67) argued that it is not clear what makes a test task demanding because challenge does not reside in the task but rather in the 'interaction of tasks, conditions and test takers' (see also Bachman, 2002). Task difficulty can therefore only be discussed 'in relation to specific speakers' (Fulcher, 2003, p. 66).

Bachman (2002) makes this clear in an analogy between language test tasks and the high jump competition in athletics; members of a high school athletics club are likely to find a particular height setting more difficult than Olympic athletes. Essentially, Bachman's argument is that experts will complete a task with relatively little effort where novices would struggle. In their review of the literature, Fulcher and Marquez Reiter (2003, p. 326) reach a similar conclusion: 'learner ability accounts for most score variance in these studies, and task difference, even if significant, accounts for only a small part of score variance'.

The language testing literature emphasises test taker characteristics (i.e. ability/proficiency; see Section 2.6.2) as an important factor in task challenge. The L2 ability of the participants in this study is limited and they have little opportunity to use English outside classrooms (see Section 2.4.1). Speech production is therefore difficult and these test takers require support to successfully complete speaking tasks. Pre-task planning may provide such support although questions remain concerning the value of planning in terms of influencing test scores (Iwashita et al., 2001).

2.5.1.4 Summary

In summary, the evidence suggests that the more challenging test takers find a language task, the more benefits can be derived from pre-task planning. However, the nature of task challenge is hard to pin down and predictions made by Skehan (2009) and Robinson (2005) have not been consistent with research findings. Nonetheless, one important generalisation can be made from the literature. Picture-based narrative tasks constrain the content that test takers describe, which may cause problems when the necessary language is unfamiliar (Bui and Huang, 2016, Foster and Skehan, 1996, Skehan, 2009). Picture-based narrative tasks may therefore be more challenging than tasks that are based on familiar information in which the test takers are free to determine the language they use. Both picture-based narrative tasks and non-picture-based description tasks test the test takers' ability to report in the second language, which is a common requirement at the undergraduate level (Inoue, 2013). For the current study, both picture-based narrative tasks and non-picture-based description tasks can be seen as contextually valid task types that pose different levels of challenge to test takers (see Section 2.5.1).

2.5.2 Planning time

An important consideration in the development of speaking test tasks is the length of planning time that is provided before the task. In the literature, the length of planning time is a key difference between studies with a task-based language teaching (TBLT) focus and those with a language testing focus. TBLT studies frequently report gains in complexity, accuracy and fluency (CAF) under a ten-minute planning condition (Bui and Huang, 2016, Foster and Skehan, 1996, 1999, Geng and Ferguson, 2013, Kawauchi, 2005, Pang and Skehan, 2014, Skehan and Foster, 1997, 2005, Yuan and Ellis, 2003). Gains in CAF have also been reported under a five-minute pre-task planning condition (Mochizuki and Ortega, 2008, Philp et al., 2006, Sasayama and Izumi, 2012) and under a fifteen-minute planning condition (Sangarun, 2005) with similar results to studies that use ten minutes planning time.

In the language testing literature, the amount of planning time is consistently less than ten minutes (one minute has featured most frequently, Li et al., 2014, Weir et al., 2006, Wigglesworth, 1997, Xi, 2005, 2010) and the effect of the pre-task planning on the test result is inconsistent. Researchers have argued that the inclusion of planning time in a language test has a bearing on test practicality (Elder and Wigglesworth, 2006). Large scale, high stakes testing is often carried out under stringent time constraints. Including a period of ten minutes planning per language test would increase the amount of time required to assess every test taker. The studies with a language testing focus tend to account for this by investigating the impact of short amounts of planning time. However, short amounts of planning time have not been shown to have as substantial an impact on results as planning for ten minutes. This suggests that there may be an optimal amount of planning time that needs to be included in the test to bias for the best performance (Swain, 1985).

Despite variation between research findings, relatively few studies have compared task performance after different amounts of pre-task planning time: Elder and Wigglesworth (2006), and Li et al. (2014). These studies present conflicting accounts of the effect of variation in planning time. Elder and Wigglesworth (2006) found no difference between test scores (rater scores and CAF) after no pre-task planning, under a one-minute planning condition, and under a two-minute planning condition. However, in a questionnaire about their use of the pre-task planning time, test takers expressed a preference for planning. The researchers suggest that as two minutes did not benefit the test takers in terms of test scores, extending planning time on the IELTS exam beyond one minute would be redundant.

Li et al. (2014) used measures of CAF (see Section 2.7.2) to compare the effect of five periods of planning time: 30 seconds, one minute, two minutes, three minutes, and five minutes. Overall, the increases in planning time led to progressively more accurate language (i.e. more error free analysis of speech units (AS-units) and fewer errors per AS-unit). The five-minute planning condition resulted in the most accurate performances overall. However, the largest gains were made between the 30 seconds and one-minute planning conditions, where the number of error free AS-units increased from .48 to .60. Fluency was measured with mean length of run (mean number of syllables supplied between pauses above 0.28 seconds), and speech rate A (syllables per minute) and speech rate B (meaningful syllables per minute). Mean length of run and speech rate A showed incremental increases with every addition of extra planning time up to three minutes. The increases peaked at three minutes and results were lower under the five-minute planning condition. The researchers suggest that this result is evidence of an optimal planning condition for fluency (i.e. three minutes) which, if exceeded causes the effect of pre-task planning to decrease. This conclusion was not confirmed by the speech rate B results, which increased with every addition of extra planning time, i.e. five minutes led to the highest results when only the ‘meaningful’ syllables were calculated in the analysis (Li et al., 2014, p. 46). Complexity results showed that 30 seconds planning led to the highest levels of syntactic complexity and one minute led to the highest level of lexical complexity. This was a surprising result. The researchers do not provide any explanation for the results of the lexical analysis. However, based upon the results of previous research findings (e.g. Crookes, 1989, Yuan and Ellis, 2003), Li et al. suggest that their planning conditions did not provide sufficient time to raise syntactic complexity, and that ten minutes pre-task planning may be necessary for increases in syntactic complexity to occur.

In sum, the findings reported in Li et al. (2014) indicate that increasing the amount of planning time does not cause systematic increases in all measures of CAF. Perhaps most importantly, the researchers suggest that three minutes is the optimal period of planning for eliciting high levels of fluency. Planning for periods in excess of this amount may cause the planning impact to diminish. This finding does not correspond to the broader research findings reported in the TBLT research, which show that ten minutes planning consistently leads to increases in measures of speech fluency. Nor does it support the results of language testing research that has investigated planning for three minutes (Elder et al., 2002, Elder and Iwashita, 2005, Nitta and Nakatsuhara, 2014). In the literature, the optimal amount of planning time is ten minutes (Ellis, 2009).

The inconsistency in the reported findings of Elder and Wigglesworth (2006) and Li et al. (2014) may perhaps have resulted from differences in the research settings. The former was conducted in Australia (i.e. an English as a second language context; see Section 2.4.1), whereas Li et al. was conducted in China (i.e. an English as a foreign language context). Li et al. involved a homogenous group of participants with similar levels of ability in English that did not have the opportunity to use the language frequently and so develop spoken language proficiency. Proficiency has been identified as a key variable in the effect of pre-task planning (see Section 2.6.2). Additionally, it is not immediately clear that the research participants in Li et al. were aware that they were being tested. The study was conducted in a language laboratory with participants speaking directly to a computer and there is no indication in the research that the participants were informed that the speaking samples would be used to assess their L2 ability. In contrast, Elder and Wigglesworth investigated the impact of planning on IELTS, a high stakes exam used to determine eligibility to follow English-medium education and for purposes of immigration. The literature indicates that the impact of planning on spoken performance varies substantially under examination and non-examination conditions (see Tables 1 and 2): benefits are more frequently observed when the participant is not being tested. In short, test taker characteristics and the social setting may account for much of the disparity between the studies.

A major gap in the literature is the absence of research both in language testing and TBLT that compares the effect of ten minutes planning with other lengths of planning time. This is problematic; ten minutes has most consistently led to positive results in the TBLT research. The most crucial comparison to make is between test scores under a ten-minute planning condition and under a one-minute planning condition. This is because the studies that investigate the impact of planning for one minute have been conducted with a language testing focus and show inconsistent results. According to this review, another common length of planning time is five minutes. This length of planning time increased CAF in Tavakoli and Skehan (2005), Li et al. (2014), Mochizuki and Ortega (2008), and Sasayama and Izumi (2012). Five minutes may be a sufficient increase over three minutes, which has consistently been shown to have a minimal impact on test scores and CAF measures in the literature (Elder et al., 2002, Elder and Iwashita, 2005, Iwashita et al., 2001, Nitta and Nakatsuhara, 2014). To obtain results after a short period of planning, the research commonly uses 30 seconds planning time, which is generally deemed sufficient for test takers to familiarise themselves with task demands (Elder et al., 2002, Iwashita et al., 2001, Li et al., 2014).

2.5.2.1 Summary

This section has described research that compares the effects of different lengths of pre-task planning time. To sum up, planning for ten minutes before a language task most consistently leads to high levels of CAF in TBLT. This is in contrast to studies with a language testing focus, which due to practicality constraints, investigate the effect of less planning time (typically one minute), and report inconsistent results. Elder and Wigglesworth (2006) show that test scores are similar after no planning, one-minute, and two-minute planning conditions. However, Li et al. (2014) report differences in CAF after minor changes to planning conditions (e.g. from 30 seconds to one minute). The amount of planning time may be an important variable in the results of the research. The present study investigates the impact of four pre-task planning conditions on test scores: 30 seconds, one minute, five minutes, and ten minutes.

2.6 The cognitive element of validity

The cognitive element of validity refers the extent to which the test tasks identified to represent the target situation (see Section 2.5) elicit cognitive processes from test takers that would naturally occur in that environment (Field, 2011, O'Sullivan and Weir, 2011, Weir, 2005). To elicit these cognitive processes, Weir

(2005, p. 18) explains that test developers ‘need to be aware of prevailing theories concerning the language processing which underlies the various operations required in real-life language use’. Evidence of the cognitive element of validity is gathered from theory about the cognitive processing underpinning the skill being tested, which is integrated with relevant information about the target situation (Field, 2011). As planning is commonly part of the target situation (see Section 2.5), it is crucial for the test developer to make efforts to elicit the cognitive processes involved in recalling speech from a plan. Therefore, the following section describes Levelt’s (1989) model of speech production to explain the impact that pre-task planning is believed to have on the cognitive operations involved in L2 speech production.

2.6.1 Speech production and planning

In both TBLT and language testing the process of speech production is commonly understood in terms of Levelt’s (1989) model of the first language (L1) speaker (Field, 2011, Skehan, 2009, Weir, 2005). The model describes a modular process of speech production involving three stages: conceptualization, formulation and articulation. Conceptualization is a process in which the speaker generates a pre-verbal message. Once generated, the pre-verbal message is passed on to the formulator, which identifies and retrieves relevant information from the speaker’s lexicon to create a morpho-syntactic plan. The morpho-syntactic plan goes through a process of phonological encoding in which it is converted into a phonetic plan. The phonetic plan is then sent to the articulator for the production of overt speech. In L1 speech, these processes proceed in parallel. This is due to the extensive competence people obtain in their first language through years of immersion and use. This level of competence allows for the effortless retrieval and assembly of language that facilitates fluent speech.

During L2 acquisition, the speech of a second language learner draws on less extensive competence than is available in the L1. Whereas L1 speech production is a relatively automatized process, L2 speech utilises a less refined lexicon and requires 'conscious attentional control' in order to proceed (Kormos, 2006, p. 166). However, attentional resources are limited in capacity (Baddeley, 2007) and attention devoted to one stage of Levelt's model compromises the ability to simultaneously carry out another (Skehan, 2009). Under these circumstances, speech production proceeds serially. This form of processing yields speech that is effortful and slow (Kormos, 2006, Skehan, 2014, 2016). Efforts to maintain parallel processing create competition for attentional resources. Therefore, the novice language learner is reliant on familiar forms that have become automatized through regular use (Ellis and Barkhuizen, 2005). These forms represent the language learner's early approximation of the target language system and frequently contain instances of non-standard usage. Limited attentional resources are available for speech monitoring so if an erroneous form is produced, it may not be corrected.

The assumption in second language acquisition research is that planning impacts the language learner's processing capacity by freeing up attentional resources for speech production (Skehan, 2009). This begins with conceptualisation (Ellis, 2005). Engaging in conceptualisation during planning allows the learner to consider content, pre-empt problems related to gaps in language knowledge and develop potential solutions. This may involve the development of suitable communicative strategies that can be used during the task to compensate for limited linguistic knowledge (Ellis and Barkhuizen, 2005). Pre-task conceptualisation reduces competition for limited attentional resources. Therefore there is less need for conscious attentional control and serial processing. Once the pre-verbal message has been processed, the language learner may then begin to consider suitable forms to express it. Planning may thus facilitate focus on form without the dual pressure of having to simultaneously produce speech (Ellis, 2005, 2009). There is less reliance on automatized forms and the language learner may produce more advanced language that is less immediately available in the lexicon.

2.6.2 Second language proficiency and planning

In light of the previous section, this review identifies language proficiency as a key variable in the pre-task planning literature. As described in Section 1.1, this study is designed to assess the impact of pre-task planning with language learners that lack high levels of proficiency in the target language. This section discusses the findings that have been reported in the literature with regards to proficiency and planning. These findings indicate that planning may only impact speech when language learners have a sufficient amount of language proficiency.

Relatively few studies have investigated L2 English proficiency as an independent variable in the planning research. In Kawauchi (2005) high-level language learners (TOEFL group mean = 545, range = 510-580, IELTS mean = 6, range = 5.5-6.5) benefitted more from the opportunity to plan than advanced learners (TOEFL group mean = 588, range = 550-610, IELTS mean = 6.7, range = 5.5-7.0) and low-level learners (TOEFL-equivalent group range = 420-480). Kawauchi argues that the advanced learners were constrained by a ceiling effect. The advanced learners' language proficiency was so developed that planning had little impact on their task performance as measured by CAF. Regarding the low-level group, Kawauchi speculates that while planning did lead to minor gains in accuracy and fluency, complex L2 forms had not yet been acquired and were unavailable during planning. To extrapolate, planning before a language task may not lead to more complex language if complex forms are not available in the language learners' repertoire.

Kawauchi's observation that planning has a limited impact on the speech of low proficiency language learners is supported by the findings of two research studies. In Mochizuki and Ortega (2008), pre-task planning did not enhance elementary level learners' language complexity because complex language forms had not been acquired. Secondly, Genc (2012) found that pre-task planning did not impact levels of accuracy in the speech of low-intermediate Turkish learners of English. She speculates that low proficiency language learners may be preoccupied with the generation and organisation of ideas during pre-task planning and have limited resources remaining to focus on form. Pre-task planning may thus have no effect on speech accuracy at low levels of proficiency. This research suggests that at lower levels of proficiency, learners do not have the L2 resources available to benefit from pre-task planning.

Wigglesworth (1997) investigated potential interactions between planning and proficiency: her participants were classified as levels 3 and 2 on *access* (Australian assessment of communicative English skills). Brindley, Hood, McNaught and Wigglesworth (1997, p. 36) describe proficiency at level 3 as the ability to ‘speak English well enough to handle basic communication in everyday situations but you make a lot of errors’ and at level 2 as the ability to ‘speak enough English to have a very elementary conversation but with many errors and hesitations’. Wigglesworth’s (1997) results showed that all participants benefitted from planning. When the proficiency levels were compared, results showed the higher level participants had benefitted from planning more than the lower level participants for measures of complexity and fluency. This was particularly evident on Wigglesworth’s complicated tasks (based upon test scores). In the case of lower level test takers, Wigglesworth argues that test takers either do not make effective use of the planning time or are engaged in formulating the content of the speech, which has little effect on speech quality: ‘It may be that at different levels of proficiency candidates undertake different activities during planning time and focus on the different requirements of the task’ (1997, p. 102).

2.6.3 Generalizing about the relationship between proficiency and planning

The literature indicates that the levels of proficiency in the participant sample will likely play a major role in a language test that involves pre-task planning. However, generalizing about the interaction between proficiency and pre-task planning from the research done so far is difficult because of the absence of a systematic approach in the literature to measure proficiency according to common criteria. Specifically, research that has formally investigated task results as a product of planning and proficiency relies upon subjective terminology such as pre-intermediate, intermediate and advanced. In short, it is uncertain whether Kawauchi's (2005) low-level group is similar to the 'low-level' groups discussed elsewhere. This makes it difficult to generalise about the relationship between planning and proficiency based upon findings in the literature. However, Kormos (2006) and Skehan (2009, 2014) indicate that as proficiency develops, the amount of L2 resources that are accessible during planning also increases. Therefore, in order for planning to be effective, test takers must have reached a level of language proficiency in which linguistic resources are available during the planning stage.

2.7 The scoring element of validity

In the socio-cognitive framework, the scoring element of validity relates to all aspects of test scores (Weir, 2005). In order to demonstrate the scoring element of validity, the test developer must provide evidence for an absence of error and bias in the scores, consistency in scoring, and dependability of decisions relating to test taker performance. In tests of spoken ability, issues of scoring relate to the rating criteria or rating scale, the rating procedures (e.g. training and standardisation), the raters, and the statistical analysis of the results. Given the differences in approaches to scoring between TBLT and language testing (see Section 2.2), this section discusses these issues under the broad heading of measurement.

2.7.1 Measurement

This section reviews the approaches to the measurement of task-based language performance in the literature by discussing measures of complexity, accuracy and fluency (CAF) and rating scales. This review demonstrates that measurement is a key issue in the pre-task planning literature. The TBLT research invariably assesses the effect of pre-task planning on CAF measures and provides consistent evidence of increases in results. Studies with a language testing focus commonly assess the impact of pre-task planning with reference to a rating scale. Measures of CAF have also featured as dependent variables in the language testing research. As discussed in the Section 2.2, planning has not been shown to have a consistent impact in the language testing studies. Therefore, the following section discusses the fundamental approaches toward and differences between measurement in the TBLT and language testing research.

2.7.2 Complexity, Accuracy, Fluency

This section describes the measurement of complexity, accuracy and fluency (CAF) in the pre-task planning literature. CAF is designed to provide an 'objective, quantitative and verifiable' measure of L2 proficiency and use (Housen, Kuiken, Vedder, 2012, p. 2). CAF commonly features as dependent variables in second language acquisition research that evaluates the effect intervention on the language acquisition process (e.g. instructional approaches). The measures feature in the pre-task planning research as a way to identify the impact that planning has on speech performance. This review discusses each component of the CAF triad with the aim of generating functional definitions and reviews the conclusions that have been made in the literature regarding CAF and planning. Each section begins with a definition provided by Housen et al. (2012) and discusses the use of relevant measures in the planning research.

2.7.2.1 Complexity

Complexity is defined as ‘the ability to use a wide range of sophisticated structures and vocabulary in the L2’ (Housen et al., 2012, p. 2). Measures of speech complexity typically fall into two categories: measures of syntax, and measures of lexis. Measures of syntax are designed to assess the extent to which a text varies structurally in terms of tense, aspect, modality, voice, subordination and coordination. Measures of lexis aim to identify the extent to which the task performance varies in terms of lexical density and lexical sophistication. This review begins by discussing measures of syntactical complexity.

Measurement of syntactical complexity is typically obtained through analysis of the amount of subordination and coordination in the speech sample. Greater amounts of subordination and coordination indicate more structure in the speech (Housen et al., 2012). In the pre-task planning research, syntactical complexity is commonly calculated as the average number of clauses per speech unit (Foster, Tonkyn and Wigglesworth, 2000). A speech unit is a multi-clausal unit of spoken discourse. The speech units that have been investigated are the T-unit, the C-unit and the AS-unit. However, it has been argued that the T-unit and C-unit have not been well defined and their usage varies dramatically across the literature (Foster et al., 2000). Foster et al. (2000) argue that research involving measures of T-units and C-units may not realistically represent spoken language. For this reason, the AS-unit is frequently used in contemporary research.

The AS-unit is a ‘single speaker’s utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with either’ (Foster et al., 2000, p. 365). For example, in the following extract from a speech given by Tony Blair in 2014, the AS-unit is made up of three clauses, two of which are subordinate clauses.

‘Indeed, you can see what happens	
when you leave the dictator in place	**Subordinate Clause**
as has happened with Assad now’	**Subordinate Clause**

(Tony Blair, quoted on http://www.bbc.com/news/uk-27852832 accessed 9, 11, 2014)

The number of clauses in the AS-unit is 3. As the speaker continues, AS-units accumulate in the discourse and the mean number of clauses per AS-unit for the entire speech can be calculated. Mean number of clauses per AS-unit has been used to measure overall structural complexity (Skehan and Foster, 2005) and commonly features in the literature.

In the pre-task planning literature, pre-task planning has been shown to impact syntactical complexity in the following ways: by increasing the mean number of clauses per C-unit (Foster and Skehan, 1996, 1999, Skehan and Foster, 1997), by increasing the mean number of clauses per T-unit (Kawauchi, 2005, Mochizuki and Ortega, 2008, Nielson, 2013, Sangarun, 2005, Sasayama and Izumi, 2012, Yuan and Ellis, 2003), by increasing the mean number of clauses per AS-unit (Elder and Wigglesworth, 2006, Geng and Ferguson, 2013, Skehan and Foster, 2005, Tavakoli and Skehan, 2005), and by increasing the total number of relative clauses (Mochizuki and Ortega, 2008). For an example of the kind of difference planning makes to levels of syntactical complexity, Foster and Skehan (1996) report that planning increased the number of clauses per T-unit from 1.20 to 1.43 on a picture-based narrative task. In addition, Mochizuki and Ortega (2008) designed a six point scoring scheme for the efficacy of relative clause use. Their results showed that while planning did increase the amount of relative clauses, the perceived quality of the relative clauses was not affected. In general, the research shows that planning has a consistent impact on the amount of subordination and coordination in the speech. However, the difference is relatively minimal and may not affect rater perceptions (Mochizuki and Ortega, 2008).

There are two forms of lexical complexity: lexical density; the range of different words that occur in the speech sample, and lexical sophistication; the relative frequency of lexis. Lexical sophistication is discussed in detail later in this section. A common approach to the assessment of lexical density is with a type-token ratio (TTR). TTR involves calculation of the number of words in the text (tokens) and the frequency with which each word occurs (types). For example, in the sentence ‘It was the best of times, it was the worst of times’ (Dickens, 1955, p.1), there are twelve tokens but only seven types as the tokens *it, was, the, of* and *times* are repeated. This sentence returns a result of 58 per cent for TTR: 42 per cent of the text involves lexical repetition. In the pre-task planning literature, the introduction of pre-task planning to a task of second language speaking has been reported to increase TTR (Crookes, 1989, Wigglesworth, 1997). For example, Crookes found that planning for ten minutes increased the mean TTR value from 3.20 to 3.57.

The use of TTR has recently been questioned with critics arguing that it is sensitive to text length (Kuiken and Vedder, 2007). The longer a text goes on the more likely TTR will decrease as repetitions naturally accumulate to form and maintain cohesion and coherence. An alternative measure put forward by Yuan and Ellis (2003) that accounts for repetition is the mean segmental type-token ratio. The mean segmental type-token ratio provides the mean TTR per 40 words. It is not heavily influenced by text length and may be a more reliable measure of lexical density than TTR. In their study however, Yuan and Ellis (2003) report that pre-task planning did not statistically significantly impact the mean segmental type-token ratio.

Three other measures of lexical density have been used in the literature. These measures involve more complicated procedures of calculation than TTR and mean segmental type-token ratio. These measures are Guiraud's index of lexical richness, D value and the measure of textual lexical density (MTLD). Guiraud's index of lexical richness is an adaptation of TTR in which the type total is divided by the square root of the token total to account for variation in the length of the text. Gilabert (2007) reports that Guiraud's index increased after pre-task planning. D value is calculated through random samplings of TTR in a text to establish an empirical curve based on TTR means. The D coefficient is used to create a theoretical curve that matches the empirical curve. The calculation is repeated three times (to account for the random sampling) and the result is then used to report lexical density (McCarthy and Jarvis, 2010). D-value was shown to improve after planning in Li et al. (2014). MTLD is calculated by creating a sequential string of words that share the same TTR value. Once a word is repeated, the value is reset and a factor is added into the MTLD equation. The final value represents the amount of lexical density as the sum of the total number of words divided by the factor value (McCarthy and Jarvis, 2010). Pre-task planning did not affect MTLD in Nitta and Nakatsuhara (2014).

Measures of lexical sophistication are obtained by calculating the frequency with which the words occurring in the text also occur in a particular corpus such as The British National Corpus. The rationale for this approach is that the less frequent words in the corpus are more advanced, technical, concise and sophisticated and the ability to use this vocabulary is indicative of a more developed L2 system. A way of obtaining a measure of lexical sophistication is the VocabProfile program (Cobb, n.d) [accessed 1 October 2015 from http://www.lextutor.ca/vp/]. This is a computer program that categorizes the vocabulary in a text into three word lists: the first and second 1,000 most frequent words and the university word list (academic word list). The greater the proportion of the text that belongs to the second and third lists, the greater the level of lexical sophistication. This measure has not been used to assess the impact of pre-task planning on lexical sophistication.

In sum syntactical complexity, as measured by the number of clauses per speech unit has frequently been shown to increase with the addition of pre-task planning to a language task. Lexical density, measured by type-token ratio, Guiraud's Index and D-value has also been shown to increase with planning. The interaction between planning and lexical sophistication has not yet been investigated.

2.7.2.2 Accuracy

Accuracy is defined as 'the ability to produce target-like and error free language' (Housen et al., 2012, p. 2). Measures of accuracy assess the extent to which language learners produce speech that adheres to the grammatical conventions of the target language. There are two broad categories of accuracy measures: measures of global accuracy and measures of accuracy of specific language forms. This review begins by discussing measures of global accuracy.

Measures of global accuracy are designed to provide an overall indication of the proportion of language use during a language task that is grammatically accurate. In the pre-task planning literature, the process of assessing global accuracy has involved the identification and quantification of measures such as error free clauses (Skehan and Foster, 1997, Foster and Skehan, 1999) or the percentage of error free speech units (T-units in Crookes, 1989, C-units in Elder and Iwashita, 2005, AS-units in Elder and Wigglesworth, 2006). To calculate the number of error free AS-units, first calculate the number of AS-units, then calculate the number of AS units that contain grammatical errors. The result is the number of error free units divided by the total number of units. For example, if ten AS-units are produced, and six of these units contain errors, the percentage of error free AS-units is 40 per cent. In the pre-task planning literature, planning has not consistently impacted the number of error free speech units (Crookes, 1989, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006). However, Skehan and Foster (1997) and Foster and Skehan (1999) demonstrate increases in the percentage of error free clauses after planning. For example, Skehan and Foster (1997) report that planning increased the percentage of error free clauses on a picture-based narrative task from 53 per cent to 69 per cent.

The calculation of error free AS-units provides an indication of the overall grammatical accuracy of a speech sample. However, there is a shortcoming with this measure. An AS-unit that contains multiple grammatical errors returns the same result as an AS-unit that contains only one (i.e. this AS-unit is/is not error free). It is necessary to calculate an additional measure of global accuracy that accounts for this shortcoming. Li et al. (2014) propose the mean number of errors per AS-unit. Used in combination with the calculation of error-free AS-units, the mean number of errors per AS-unit describes the extent to which grammatical errors are common within AS-units. Li et al. (2014) demonstrated that planning enhanced grammatical accuracy by reducing the mean number of errors per AS-unit from .68 after no pre-task planning to .42 under their five-minute pre-task planning condition.

Measures of the accuracy of specific language forms focus on a particular area of the second language such as the article system. This level of analysis is specific and provides detailed information about the kind of inaccuracies that appear in a spoken sample. For example, to measure the accuracy of article usage, first the number of 'obligatory occasions' (Ellis and Barkhuizen, 2005, p. 80) for articles is calculated, then the number of correctly supplied articles is calculated, the number of correctly supplied articles is then divided by the number of obligatory occasions and the percentage that was correctly supplied is reported. Research that uses specific measures includes Yuan and Ellis (2003) who report no difference in correct verb forms after pre-task planning. Wigglesworth (1997) reports that planning increased the accuracy of articles but had no effect on plural usage and verbal morphology. Neilson (2013) found that planning had no impact on subject verb agreement. Crookes (1989) found no difference in the accuracy of article use and plural noun endings. In sum, the findings indicate that pre-task planning makes little difference to the results of measures of the accuracy of specific language forms.

One drawback of the measurement of the accuracy of specific language forms is that the outcome is heavily reliant on the individual test taker who may overuse or avoid a given structure. Specifically, in Crookes' (1989) study, the degree of accurately supplied articles was heavily dependent upon the language proficiency of the participants. Crookes argues that the definite article is a language form that is generally acquired at later stages of the acquisition process by Japanese learners of English. Seeking to generalize about the level of accuracy with such a measure may therefore produce a questionable outcome. As such, the solitary usage of specific measures for assessment of accuracy may prove problematic and is not recommended. A combination of global and specific measures allows for a richer and fuller description of the grammatical accuracy of test performance.

Researchers have commented upon the limited effect that planning seems to have on the grammatical accuracy of speech (Ellis, 2005, 2009, Foster and Skehan, 1996, 1999, Skehan, 2009, Skehan and Foster, 1997, Yuan and Ellis, 2003). It is widely acknowledged that of the three areas of speech production (CAF), accuracy is least consistently increased by pre-task planning (Ellis, 2009). Yuan and Ellis (2003) suggest that accuracy is more related to opportunities to engage in online planning (i.e. planning during the task) than pre-task planning. In their research, the findings indicate that when a task involves time pressure, levels of accuracy diminish.

2.7.2.3 Fluency

Fluency is defined as ‘the ability to produce the L2 with native-like rapidity, pausing, hesitation, or reformulation’ (Housen et al., 2012, p. 2). The term fluency encompasses both a general, popular understanding akin to proficiency or competence and a more specific, specialist interpretation within Applied Linguistics that views the concept as a temporal construct. Within this specialist interpretation (i.e. the CAF framework), fluency is distinguished from speech accuracy and speech complexity and comprises characteristics of the speech such as speed, length, instances of repetition, and the duration and position of pauses and hesitations.

Measurement of fluency involves analysis of the speed of speech delivery and the extent to which breakdown fluency (pausing) and repair fluency (false starts, reformulation) occur in the speech (Skehan, 2009). Kormos (2006) discusses two measures of fluency that frequently appear in the second language acquisition (SLA) literature and are particularly relevant for this study: speech rate and phonation time ratio. Speech rate is the total number of syllables/words produced in the text divided by the total speaking time expressed in seconds. This number is then multiplied by sixty to represent syllables/words per minute (Kormos, 2006). It provides an indication of the speed with which speech is produced. Speech rate increased after pre-task planning in Guara-Tavares (2009), and Li et al. (2014). For an example of how pre-task planning affects speech rate, Li et al. (2014) report that the mean number of syllables increased from 112.49 after planning for 30 seconds to 134.86 after planning for five minutes. Research has also used pruned speech rate as an indication of speech fluency. This is a similar measure to speech rate that involves removing all pauses and hesitations from the final calculation. Pruned speech rate was shown to increase after pre-task planning in Sangarun (2005), Gilabert (2007), Guara-Tavares (2009), Geng and Ferguson (2013) and Nielson (2013).

The second measure discussed by Kormos (2006) is phonation-time ratio, which is a measure of breakdown fluency. This measure compares the total amount of time spent speaking with the total amount of time spent pausing. It provides a strong indication of the amount of task time that the speaker spends producing speech. To calculate this measure, identify the amount of silent task time in seconds and subtract this number from the total task time. The result of this calculation is divided by the total task time and reported as a percentage. Bui and Huang (2016) report that pre-task planning increased phonation-time ratio from 76 per cent to 82 per cent.

Speech rate and phonation-time ratio can be calculated with an acoustic analysis of the speech sample. However, Field (2011) suggests that a more thorough, qualitative analysis is required to generate a comprehensive measure of fluency. Pauses, gaps in speech that occur between syntactic boundaries, must be distinguished from hesitations, gaps that occur within syntactic boundaries. This distinction is important because pauses are a common feature of all speech and fulfill a necessary role of allowing the speaker to generate content. In contrast, hesitation shows that some form of extra effort and attention is required to complete an utterance. Excessive hesitation is caused by gaps in language proficiency and is more likely to effect an interlocutor's impression of fluency than pauses (Field, 2011). Kormos (2006) describes pauses and hesitations as a period of silence in excess of .25 seconds, whereas Foster and Skehan (1996) set the criteria as silence in excess of one second. Calculations of the number of pauses and hesitations using Foster and Skehan's criteria (1996) have figured frequently as measures of fluency in the pre-task planning literature (Foster and Skehan, 1996, Foster and Skehan, 1999, Nitta and Nakatsuhara, 2014**,** Skehan and Foster, 1997, Skehan and Foster, 2005, Tavakoli and Skehan, 2005). This research has produced consistent evidence that the number of pauses and hesitations decreases after a period of pre-task planning. For example, Skehan and Foster (1997) demonstrated that the number of pauses decreased from 23.8 to 6.0 after pre-task planning on a picture-based narrative task.

Pauses and hesitations may involve periods of silence during the task or may be filled with fillers such as *erm*, *um,* or *mmm*. Filled pauses and hesitations are very common in spoken discourse and serve the same purpose as unfilled pauses and hesitations (Kormos, 2006). In the pre-task planning literature, Skehan and Foster (2005) report that the number of filled pauses decreased after pre-task planning.

Manual identification of pauses and hesitation allows for additional, more detailed phonological analysis of the speech such as mean length of utterance, which is a measure of the number of words produced between filled/unfilled pauses and hesitations. The ability to produce speech without having to pause or hesitate is a key indicator of language proficiency. Field (2011) describes the mean length of utterance as a measure of the extent to which the processes of language retrieval and encoding of the speech have become proceduralised. In the literature, mean length of utterance (referred to as ‘mean length of run’) increased in Li et al. (2014, p. 9) from 5.76 without planning to 6.94 after three minutes planning.

Limitations in the study of CAF are discussed in the following section (Section 2.7.2.4). However, Fulcher (2015) specifically discusses limitations in the use of the kinds of fluency measures discussed in the previous paragraphs and is therefore included in this section. Fulcher views fluency as context dependent and as a quality that the listener is attuned to. Fulcher (2015, p. 76) stresses that fluency measures do not account for the influence of the environmental context on language use: ‘what then can be the purpose of simply counting pauses, or measuring pause length or speech rate, when these vary for a variety of reasons, only some of which are related to L2 language proficiency’. Fulcher (2015, p. 80) concludes by stating ‘with the help of a rating tool, the most reliable and valid measures of spoken fluency come from human judges’. Though language learners do pause for reasons of pragmatics, at the early stages of L2 proficiency, a great deal of cognitive effort and attention is required to produce speech (Field, 2011). This cognitive effort impacts the degree to which pauses and hesitations are necessary; more effort requires more hesitation. Measures of speech rate and phonation time ratio supply important information about the degree of effort and attention required to communicate a message, which is a key indication of language proficiency. As the pre-task planning literature consistently shows that planning impacts speech fluency, it is important to investigate the impact of planning on fluency in the present study.

2.7.2.4 Limitations of CAF

Applied linguists have described various methodological shortcomings in the study of CAF (Fulcher, 2015, Lambert and Kormos, 2014, Housen et al., 2012, Housen and Kuiken, 2009, Pallotti, 2009, Norris and Ortega, 2009). Primarily, the CAF approach assumes that the results of CAF measures are directly related to the quality of speech. Implicit in much of the research is the notion that more is better. For example, if a speaker demonstrates a high speaking rate he is deemed to be fluent regardless of the effect that the speech speed might have on the listener. Likewise, speech that is lexically diverse is complex regardless of the extent to which variety is appropriate to the specific context. In short, an entirely quantitative approach based on CAF measures is not able to evaluate the social aspect of language use in terms of adequacy and appropriacy (Hymes, 1972). As Pallotti (2009, p. 596) writes:

‘If in an information gap task a learner were to utter unhesitatingly *colorless green ideas sleep furiously on the justification where phonemes like to plead vessels for diminishing our temperature*, her production would score extremely high on CAF, in spite of being completely irrelevant, and probably counterproductive, for task success.’

Research into pre-task planning has found little relationship between the results of CAF analysis and test scores (Wigglesworth, 1997). Wigglesworth reports that CAF results increased with planning but the results of the test scores remained the same. It was clear that participants had benefitted from the planning time but these gains were only evident in the results of CAF analysis. These findings suggest that increases in CAF results may have little bearing upon rating scale results. This calls into question research conducted by Tavakoli and Skehan (2005) and Li et al. (2014) who refer to increases in CAF in their studies to suggest that planning is likely to make a difference to scores on language tests.

One important limitation of using multiple CAF measures is the increased chance of committing a type one error in tests of statistical significance. As discussed in Section 2.2, calculating multiple statistical tests simultaneously increases the likelihood of obtaining a statistically significant result by chance (Panchin and Tuzhikov, 2017). To reduce the chances of committing a type one error, the critical significance level (alpha) may be corrected using a method such as Bonferroni correction (Brown, 1990). Table 3 presents the studies that have reported positive impacts of pre-task planning on multiple CAF results at p = .05. The table reports the number of tests that were conducted and the alpha value that was applied. Following this, the number of CAF results that were statistically significant at p = .05 are stated. In the following column, the adjusted alpha using Bonferroni correction is reported. The CAF results that meet the new alpha level are reported in the final column.

Three of the studies included in the table report results that do not meet the Bonferroni correction (i.e. Ellis, 1987, Crookes, 1989, Sasayama and Izumi, 2012). In nine of the studies, the number of statistically significant results decreases. For instance, in Foster and Skehan (1996) using an adjusted alpha level of (.05/10) p = .005, rather than the reported five results, two results reach statistical significance: the number of pauses and amount of task silence. Wigglesworth (1997) discusses statistically significant impacts of the planning variable on complexity and accuracy.

However, the only result that reaches statistical significance at (.05/6) *p* = .008 is the number of self-repairs. Increases in the use of relative clauses in Mochizuki and Ortega (2008) are statistically significant using the adjusted alpha but this may be due to instructions provided to participants about how to structure relative clauses during the planning time. Furthermore, the quality of relative clause use was not significant at *p* = .008, indicating that the planning time did not affect how well the participants used relative clauses but rather how frequently. In short, form focused planning did not lead to greater accuracy with the form.

When an alpha correction is applied, fluency measures are most consistently statistically significant when no guidance is provided during the planning stage. However, even in these cases, the number of statistically significant results is lower following alpha correction than was reported by the researchers. For example, in Bui and Huang (2016) ten of the 14 results are statistically significant after the alpha correction is applied. However, many of the measures assess much the same thing; mid-clause pause length, mid-clause silence total and phonation time ratio are all measures of the amount of silence that occurs during the task. Finding one statistically significant result among these measures suggests that the others are also likely to be statistical significant. When corrections are made to account for the risk of type one error, conclusions concerning the effects of pre-task planning on CAF outcomes appear less conclusive. Limitations in the statistical analytical approaches adopted in TBLT constitute a major shortcoming in the pre-task planning literature that means the reported findings must be interpreted with caution.

Table 3 Statistically significant effects of the planning variable with and without alpha correction (continued on page 74)

Study	Number of dependent variables	Unadjusted alpha value	Number and nature of results that meet unadjusted alpha value	Alpha value after Bonferroni correction	Number of results that meet adjusted alpha value
Ellis (1987)	3	.05	1: Accuracy of past copular	.016	0
Crookes (1989)	16	.05	4: Words per utterance, sub clauses per utterance, s-nodes, words per subordinate clause	.003	0
Foster and Skehan (1996)	10	.05	5: number of replacements, number of hesitations and repetitions, accuracy of past tense, number of pauses, amount of silence.	.005	2: number of pauses, amount of silence
Skehan and Foster (1997)	3	.05	3: number of pauses, error free clauses, number of clauses per c-unit.	.017	3
Wigglesworth (1997)	6	.05	4: number of subordinate clauses, type-token ratio, number of self repairs, accuracy of articles.	.008	1: self repairs
Foster and Skehan (1999)	8	.05	4: clauses per c-unit, error free clauses, number of pauses, turn length	.006	1: Turn length
Yuan and Ellis (2003)	7	.05	2: pruned speech rate, clauses per t-unit	.007	1: clauses per t-unit
Sangarun (2005)	7	.05	7: s-nodes per t-unit, clauses per t-unit, percentage of error-free clauses, number of errors per 100 words, speech rate, pruned speech rate, percentage of total pausing time	.007	4: s-nodes per t-unit, percentage of error free clauses, number of errors per 100 words, percentage of total pausing time
Tavakoli and Skehan (2005)	12	.05	7: total silence, length of run, pause length, speaking time, speech rate, error-free clauses, clauses per AS unit	.004	5: total silence, length of run, pause length, speaking time, error-free clauses
Gilabert (2007)	4	.05	2: pruned speech rate, Guiraud’s Index	.013	2

Mochizuki and Ortega (2008)	6	.05	3: quality of relative clauses, number of relative clauses, relative clauses per t-unit	.008	2: number of relative clauses, relative clauses per t-unit.
Guara-Tavares (2009)	4	.05	3: errors per 100 words, speech rate, pruned speech rate	.013	3
Sasayama and Izumi (2012)	7	.05	2: clauses per t-unit, number of repetitions	.007	0
Geng and Ferguson (2013)	3	.05	3: speech rate, clauses per AS-unit, errors per 100 words	.016	3
Nielson (2013)	4	.05 (3 parametric tests) .05 (1 non-parametric test)	2: pruned speech rate (parametric test), clauses per t-unit (non-parametric test)	.016 (3 parametric tests) .05 (non-parametric)	2: pruned speech rate (.016), clauses per t-unit (.05)
Nitta and Nakatsuhara (2014)	7	Not specified	3: number of words per second, length of pauses per second, number of words per turn	.007	2: number of words per second, length of pauses per word
Bui and Huang (2016)	19	.05	14: speech rate, pruned speech rate, phonation time ratio, number of mid-clause pauses, mid-clause length, mid-clause silence total, number of independent clause pauses, independent clause pause length, independent clause silence total, dependent clause pause length, number of pseudo filled pauses, false starts, reformulations, repetitions	.003	10: pruned speech rate, phonation time ratio, mid-clause pause length, mid-clause silence total, number of independent clause pauses, independent clause silence total, number of pseudo filled pauses, false starts, reformulations, repetitions

* Continuation of Table 3

2.7.3 Rating scales

In contemporary language testing, examiners (referred to as raters) use rating scales to assign scores to test takers. Rating scales 'consist of graded descriptions' of language ability that are 'intended to characterise different levels of performance' on a language test (Green, 2014, p. 144). An important consideration in the development of speech assessment is therefore the design of the rating scale and its content. A number of studies have investigated the impact of planning on the results of test scores (Elder et al., 2002, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006, Iwashita et al., 2001, Nitta and Nakatsuhara, 2014, Weir et al., 2006, Xi, 2005, 2010). All of these studies were conducted in a language testing context and report findings that are relevant for this study. The findings are reviewed below.

Iwashita et al. (2001) developed an analytic rating scale to describe complexity, accuracy and fluency at five levels of language ability (see Appendix 1). The scale content ranges from descriptions of beginner language learners to advanced users, i.e. 'similar to a native speaker' (2001, p. 435). The authors do not describe the process of creating the scale. However, it seems likely that the scale content was generated according to the authors' theoretical understanding of complexity, accuracy, and fluency at five levels of language ability. This method of scale development is referred to as intuitive; it is derived from theoretical application rather than data analysis (Fulcher, 2003). Intuitive scale development is discussed at length in Section 2.7.3.1.

The findings of the studies that use this scale offer little evidence of any pre-task planning impact on the test scores. In Iwashita et al. (2001) and Elder et al. (2002), increasing planning time on a narrative task from 30 seconds to three minutes and 30 seconds did not make a difference to the scores on the scale in relation to performance by the same test takers on otherwise equivalent tasks. Elder and Iwashita (2005) also found no significant difference in scores on a narrative task after an extra three minutes of pre-task planning was included in the test. The findings of these studies indicate that planning made no difference to the results measured by this particular scale. Nitta and Nakatsuhara (2014) modified the Iwashita et al. (2001) scale to include more levels of ability by placing a band level in between each original descriptor. This had the effect of increasing the number of bands from five to nine for each category (complexity, accuracy, and fluency). The researchers used the adapted scale to assess performance on a discussion task after three minutes planning. Statistically significant increases in scores were observed in complexity (an increase of .18) and fluency (an increase of .44) after three minutes planning time. In sum, when the scale was adjusted to include more levels, planning caused a minimal increase in scores.

Xi's research (2005, 2010) was conducted with the Speaking Proficiency English Assessment Kit (SPEAK) produced by the Educational Testing Service (ETS). The SPEAK exam is a tape mediated test of spoken proficiency that is commonly used to measure the language proficiency of prospective international teaching assistants in American universities. The SPEAK rating scale contains five levels of language ability ranging from beginner to advanced. Using the scale, raters consider linguistic competence, discourse competence, functional competence, and sociolinguistic competence to distinguish between the levels. Xi (2005) used the scale to assess test performance on a series of graph description tasks. Results showed that allocating one minute for pre-task planning increased mean scores by 1.31 on the scale.

Building upon the earlier study, Xi (2010) developed an analytic rating scale to assess fluency, organization, and content at five levels of second language ability. While Xi's approach to the construction of the scale was intuitive in the sense that the content was not derived from empirical analysis of test performance, the basis for the scale content was drawn from previous findings (Xi, 2005). Xi predicted that planning would have more of an effect on the fluency, organization and content categories of the SPEAK scale than the sociolinguistic competence category of the scale. This allowed Xi to identify, and construct the scale around the features of speech that had been most impacted by planning in the earlier research. The results showed increases of .23, .31, .24 of a level for fluency, organization and content respectively after one minute of pre-task planning. Although it was clear that planning had impacted test scores, the differences in scores were rather small.

Both Weir et al. (2006) and Elder and Wigglesworth (2006) investigated the impact of pre-task planning on the speaking scale of the International English Language Testing System (IELTS). The scale contains band descriptors for nine levels of ability in four categories: fluency and coherence, lexical resources, grammatical range and accuracy and pronunciation. Weir et al. report that scores increased by .32 of a band after planning for one minute. However, Elder and Wigglesworth found that test takers did not receive higher grades after planning for either one minute or two minutes. Despite similarities between the studies in the use of the same scale and the same amount of planning time, there were still differences in the results.

Researchers have suggested that the absence of a clear impact of planning in the language testing literature may be due to the use of rating scales. Wigglesworth (1997) found that planning improved the results of measures of CAF, but made no difference to rater scores (see Section 2.7.2.4). She suggests that this may be because:

- The increases in CAF were too minor to be noticed.
- The increases were not considered sufficiently important to affect the grading.
- The rating instrument did not draw raters' attention toward the elements of the speech that improved with planning.

Comparable claims have been made by Elder and Wigglesworth (2006) who suggest that there may be a mismatch between what the IELTS raters valued and the improvements that planning brought about. This interpretation is partially contradicted by findings reported in Weir et al. (2006) who report an increase of .32 of a band on the IELTS rating scale after pre-task planning. However, the picture that emerges from the review of the literature is that when pre-task planning does impact rating scale scores, the increase is generally minimal (Nitta and Nakatsuhara, 2014, Xi, 2010). The notion that the absence of consistent planning impacts in the literature is due to the rating scale deserves further consideration.

2.7.3.1 The importance of context in rating scale development

As early as 1920, Yerkes (1920) emphasised the importance of accounting for contextual factors in rating scale development. Fulcher (2003, p. 19) argued that rating scales should refer to specific contexts of use: ‘we should not assume that any description, any rating scale, captures some psychological reality that exists in the language competence of all speakers for all time in all contexts’. Researchers generally recognise the importance of context in rating scale development but also acknowledge the need for general-purpose scales, which facilitate the generalizability of test scores beyond specific contexts. The conflict is best exemplified in recent discussions (Alderson, 2007, North, 2007, Hulstijn, 2007) of the use of the Common European Framework reference level descriptors (Council of Europe, 2001). For example, North (2007, p. 658) discussed the need for the reference level descriptors of the Common European Framework ‘to be context-free in order to accommodate generalizable results from different specific contexts, yet at the same time the descriptors on the scale need to be context-relevant’. To varying degrees, researchers have suggested that language use is contextually dependent and scale developers should aim to describe language use in specific contexts (Turner and Upshur, 1996). This can best be achieved with empirical, ‘data-based’ or ‘data-driven scale development’ whereby the ‘key features of performance’ within a specific test taking population are observed and referenced in the scale (Fulcher, 2003, p. 92). This ensures that scale content is relevant for the context of use.

Fulcher (2012, p. 383) contrasts the empirical approach with the intuitive, ‘armchair’ approach: informed by theory, experience of teaching and testing the intended test taker population, consultation with existing scales, or institutional objectives (Luoma, 2004). Intuitive scales ‘may not characterize actual language use’ in the language test (Fulcher and Davidson, 2007, p. 94) and describe a range of language characteristics and abilities that are not represented in the test taking population. Fulcher, Davidson and Kemp (2011, p. 8) refer to this as ‘descriptional inadequacy’. If a scale is inadequate for its intended purpose and context, the precision of the measurement may be affected negatively and the inferences that are drawn from test scores may be unsound. This is an example of construct under-representation and is a threat to the validity of the test (Messick, 1989). Tests in which the construct (e.g. L2 spoken proficiency) is under-represented generate test scores that are not an adequate representation of a test taker’s ability and this may have consequences for various stakeholders. In this research context, test takers may be either denied or granted access to educational opportunities based on inaccurate information pertaining to their ability to cope in an English-medium educational environment.

Fulcher’s (2003) argument that scales should be context specific is best illustrated in the pre-task planning literature in the use of the Iwashita et al. (2001) scale. The scale is designed to function as a general-purpose scale that describes a very broad range of language ability (i.e. from novice to advanced user; see Appendix 1). Including such a broad range of proficiency within one scale means that the precision of description is necessarily compromised. The scale is unlikely to adequately describe the nuanced variations in language performance within a group of language learners that share a limited range of proficiency (e.g. because they come from a similar secondary education background). The impact of pre-task planning on speech performance with these test takers needs to be very large to make a difference to test scores on this scale.

In the pre-task planning research, there is substantial mismatch between the range of language proficiency described in the Iwashita et al. (2001) analytic scale and the research participants' levels of language proficiency. Iwashita et al. (2001), Elder et al. (2002) and Elder and Iwashita (2005) report that the range of proficiency levels in their studies was 427-670 on the TOEFL paper based test. These test scores represent levels of language ability that are well beyond elementary and beginner levels. However, the scale makes reference to performance at very low levels of ability, which is both redundant and unlikely to contribute to the measurement (Iwashita et al., 2001, pp. 435-436):

- 'Clear lack of linguistic control even of basic forms' (Accuracy level 1)
- 'Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect ideas across clauses' (Complexity level 1)
- 'Speech is quite disfluent due to frequent and lengthy hesitations and false starts' (Fluency level 1)

Nitta and Nakatsuhara (2014, p. 151) report that the average level of proficiency in their study was '476.41' on the TOEFL exam. The range of language proficiency in their sample was limited. However, as discussed earlier in this section, Nitta and Nakatsuhara adjust the Iwashita et al. (2001) scale to include nine levels of language ability. The result of the planning variable is very minor increases in scores for complexity and fluency.

The Iwashita et al. (2001) scale contains content that is unlikely to be relevant to a population of test takers that share similar levels of ability. A data based approach, in which test samples are used to produce scale content, may produce band descriptors that are more relevant to the test context. Context specific scales are more likely to lead to precise measurement and discriminate between test performances. Precision of measurement is crucial for this study to identify how pre-task planning affects test performance. Therefore this study uses a rating scale that is tailored to the specific context and represents the language use of the test taking population. This is discussed at length in Section 2.7.3.3

2.7.3.2 Raters

When using rating scales it is necessary to first train raters in the use of the scale (O'Sullivan, 2012). Raters need to be informed what to look for in a speaking test sample in order to make valid decisions about language proficiency. Davis (2016, p. 119) explains that rater training is carried out to enhance test reliability, 'to reduce the differences in scores from different raters' and validity, 'to lead raters to an understanding and application of the scoring criteria that accurately reflects the language abilities the test is intended to measure'. However, when scales are developed through the intuitive methods discussed by Fulcher (2003), the scale may be inadequate for the test purpose and context. The training serves to enhance the reliability of the test but the construct may still be under-represented and the results may lack validity.

Fulcher (2003, p. 97) suggests that the process of rater training may ‘mask problems with the wording of bands in the scale by creating the illusion of psychological reality through high rater reliability’. This is not an uncommon account of the flaws involved in the use of rating scales in language assessment. Harding (2016, p. 13) writes that rating scales ‘have a limited capacity for ensuring valid interpretation and consistent application among raters’. Wisniewski (2017) identifies multiple sources of variation in rating scale-based judgments including raters placing greater value on certain language features than others, evaluating aspects of performance that are not described in the scale, and ignoring the scale completely. In the literature, research findings have been presented that show substantial variation between raters in adherence to and interpretation of the rating scale content on the Cambridge FCE exam (Orr, 2002) and institutional examinations (May, 2006).

One particularly revealing example of rater variability is Brown (2006). In her study, Brown describes the persistence of rater variation in the IELTS speaking exam. IELTS is a high stakes exam that serves a gate-keeping function for educational opportunities in English speaking countries and immigration (Merrylees, 2003). IELTS speaking examiners are naturally required to attend frequent, rigorous training sessions in the use of the IELTS rating scale. However, Brown (2006) found that even after this training, raters ‘appeared to interpret the criteria differently and included personal criteria not specified in the band scales (in particular interactional aspects of performance, and fluency). In addition, it appeared that different criteria were more or less salient to different raters’ (2006, p. 2). Brown’s research demonstrates that individual differences between raters commonly persist despite thorough rater training.

The rater has considerable impact on the results of the test (Brown, 2006, Fulcher, 2003, Wisniewski, 2017). In addition to representing specific groups of test takers, rating scales should be designed with a population of raters in mind. The scale developer must seek to represent in the rating scale the features of the test performance that raters regard as salient. This procedure supports the use of the rating scale and ensures that consistent conclusions are made about the way pre-task planning affects the test scores.

2.7.3.3 EBB scales: a solution?

An approach to scale design that has not featured in the pre-task planning literature is Turner and Upshur's (1996) EBB method. This method is an example of a data based approach to rating scale construction that attempts to resolve some of the issues relating to traditional rating scales identified in the previous section. 'The scale is *empirically* derived, requires *binary* choices by raters, and defines the *boundaries* between score levels (EBB)' (Turner and Upshur, 1996, pp. 60-61). In the EBB method, the features of language performance on a specific task that are most salient to the raters are identified and used as the basis for the scale content. EBB scales are assessor oriented: the rater's rationale for scoring test samples is at the center of the rating process. Rating criteria are presented as a series of binary distinctions that represent boundaries between the levels of ability in the test taking population. Turner and Upshur (1996) describe the procedure for creating EBB scales as follows:

- A series of task samples representing the range of ability is selected and presented to a group of raters who are familiar with the student profile and task.

- The group then rank-orders the samples and decides how many levels of proficiency are present in the samples.

- The samples are divided into two groups: high-level proficiency and low-level proficiency. A feature that is common to the performances in one half of the sample is identified, e.g., 'Variety of structures (2+ sentences patterns) with expansions' (Turner and Upshur, 1996, p. 67). This is then formulated as a binary yes/no question.

- This process is repeated until each level has been distinguished with similar binary questions.

- A descriptive summary of language ability is composed for each level on the scale for stakeholder feedback purposes.

Research has shown that the use of EBB scales leads to high levels of test reliability in terms of inter-rater agreement and high discrimination between test takers' levels of speaking proficiency (Hirai and Koizumi, 2013, Turner and Upshur, 1996, Upshur and Turner, 1995). In an EBB scale validation study conducted in Japan, raters were asked to grade a series of spoken samples using both an EBB scale and an analytic scale containing the same descriptors for five levels of proficiency.

The EBB format was shown to foster higher levels of rater agreement and rater consistency than the analytic format (Hirai and Koizumi, 2013). Discussing rater evaluations and comparisons of the scales, the researchers write that the analytic scale exposed raters to all of the scale criteria at once and may thus have 'created too much of a cognitive demand on the raters, which may have led to fluctuating ratings across the five levels' (2013, p. 409).

In sum, EBB rating scales have the advantage of being referenced to a specific population and task. EBB scales are assessor oriented and reflect the raters' criteria for making proficiency related decisions. In contrast, the most frequently used scale in the pre-task planning literature, the Iwashita et al. (2001) scale, is general-purpose and seeks to describe a broad range of language proficiency. Iwashita's approach to scale design is ambitious in the range of proficiency it seeks to describe but compromises the precision of measurement and in turn the validity of the test scores. In order to measure the impact of pre-task planning on test scores, the measurement tool must be precise. The contrast between EBB and the analytic, Iwashita et al. (2001) scales is an important one for this study. The characteristics of each scale are summarised in Table 4.

Table 4 Features of EBB scales and the analytic scale

EBB scale (Turner and Upshur, 1996)	Analytic scale (Iwashita et al., 2001)
Defines boundaries between performance levels as a binary distinction.	Grades performance levels on a five-point scale.
Designed to reflect language use within a specific context by a specific test taking population.	Intended as a general-purpose scale for all contexts and users.
Empirical: raters provide rating criteria.	Intuitive: rating criteria are informed by theory.

2.7.4 Summary

This section has described approaches to measurement in the pre-task planning literature. To sum up, the results of research that involve rating scales have been inconsistent with regard to the pre-task planning impact. This may be due to the rating scale content, which may not provide sufficient description of language performance within the test taking population or adequately reflect the criteria that the raters regard as salient to their decision making process. In contrast, CAF measures have recorded consistent impacts of planning on test performance. These impacts are most evident in the complexity and fluency of the speech, although increases in accuracy have also been reported. However, there are a series of limitations in the use of CAF. Firstly, the relationship between CAF measures and test performance has been questioned on the basis that CAF does not adequately represent language use in context. In addition, increases in CAF have not been shown to correspond to increases in test scores when trained raters make judgments about language proficiency. Thirdly, the absence of alpha correction in the analysis of multiple CAF results is a shortcoming that detracts from the researchers' conclusions.

3 Research questions

This chapter begins by summarising the key issues relating to pre-task planning that were discussed in the literature review and identifies gaps in the literature relating to the measurement of test performance, task type, test taker proficiency, and different amounts of planning time. Following this, the research questions are stated.

The literature review indicates that conflicting accounts of pre-task planning may be attributable to the measurement of speech that is adopted in the research. There are broadly two approaches to the measurement of speech in the pre-task planning literature. The first approach involves measures of complexity, accuracy and fluency (CAF; see Section 2.7.2). Planning has consistently been shown to affect these measures although the absence of alpha correction is an important limitation in this research (see Section 2.7.2.4). The second approach to measurement involves rating scales (see Section 2.7.3). This second approach has not provided consistent evidence of a pre-task planning effect. However, the rating scales that have been investigated so far have not been created to describe a specific population of test takers to a specific group of raters. Research findings indicate that when rating scale content does not represent the contextualised variations in spoken proficiency that raters regard as salient, the potential for test scores to uncover a planning impact is limited (see Section 2.7.3.2). Turner and Upshur's (1996) EBB method of rating scale development is an alternative approach that may successfully discriminate between performances after different levels of planning (see Section 2.7.3.3). To investigate speech planning, this study utilises three analytical approaches to language measurement; measures of CAF, an EBB rating scale and an analytic rating scale

(Iwashita et al., 2001). The analytic scale was selected to enhance the comparability between the current study and research that has used the same scale to investigate pre-task planning in language tests (Iwashita et al., 2001, Nitta and Nakatsuhara, 2014).

Research in task-based language teaching (TBLT) has shown that the impact of planning on task completion varies substantially between different task types. In short, the more challenging the language learner finds a language task, the larger the planning impact (see Section 2.5). Positive findings have generally been recorded for picture-based narrative tasks. Picture-based narrative tasks may be regarded as more challenging than non-picture based tasks if they involve obligatory content that test takers do not have adequate language resources to describe. However, the ability to generate and communicate content independently is an important skill for assessment. Therefore, this study investigates the effect of planning on two task types; picture-based-narratives and non-picture-based description tasks.

This study is designed to assess the impact of pre-task planning with learners who are limited in second language proficiency (see Section 1.1). The research literature presents mixed results for the relationship between planning and proficiency (see Section 2.6.2). One confounding factor in this is that consistent methods for reporting language proficiency, such as the reference level descriptors in the Common European Framework (Council of Europe, 2001) were not used in the studies. It is difficult to understand what terms like ‘low-intermediate’ (Genc, 2012, p. 72) and ‘limited proficiency’ (Sasayama and Izumi, 2012, p. 29) actually refer to without recourse to a common scale. The present study systematically investigates proficiency as a potential variable in the result of planning for a language test by reporting participants' L2 proficiency in terms of the Common European Framework reference level descriptors (Council of Europe, 2001) and comparing planning results between different levels.

Research in TBLT most frequently investigates the impact of providing language learners with ten minutes to plan their speech (see Section 2.5.2). This amount of planning time has generally resulted in positive impacts on CAF. However, the language testing literature generally investigates the impact of providing much shorter amounts of planning time (most typically one minute). This amount of planning time has not had the effect that has typically been observed after ten minutes planning. There is a clear gap in the literature in relation to planning time. At present it is unclear how increasing planning time in exam conditions influences test scores. This study therefore investigates the amount of planning time that most substantially impacts CAF and test scores.

The research questions to be answered in this study are:

1. Does variation in planning time operationalized as 30 seconds, one minute, five minutes and ten minutes impact the results of a language test when assessed with
    a) an EBB scale
    b) an analytic scale
    c) measures of complexity, accuracy, and fluency (CAF)?

Evidence of pre-task planning effects has primarily been reported in measures of CAF (e.g. Foster and Skehan, 1996). Rating scales have proved less effective in demonstrating an effect of planning on test scores (e.g. Iwashita et al., 2001). Wigglesworth (1997) found increases in CAF after planning but no corresponding effect on test scores. The comparison of CAF scores and rating scale scores after variation in planning time is an important focus of this study.

If the answer to research question 1 is affirmative,

> 1.1 Which amount of planning time (30 seconds, one minute, five minutes, ten minutes) most substantially impacts test scores and CAF results?

Studies in TBLT consistently report increases in CAF after a period of ten minutes (e.g. Foster and Skehan, 1996) and five minutes (e.g. Sasayama and Izumi, 2012). In contrast, studies with a language testing focus indicate that planning for one minute or 30 seconds (e.g. Iwashita et al, 2001) has little impact on CAF and test scores.

> 1.2 Does the impact of the four planning conditions on test scores vary between the analytic scale and the EBB scale?

Research findings consistently demonstrate that variation in pre-task planning time makes little difference to scores on the analytic scale (Elder et al., 2002, Elder and Iwashita, 2005, Iwashita et al. 2001, Nitta and Nakatsuhara, 2014), whereas research is yet to investigate the impact of planning on EBB scale scores.

1.3 Does the impact of the four planning conditions on test scores and CAF results vary between groups of test takers who have different levels of language proficiency?

Proficiency may be a key variable in the effect of variation in pre-task planning time (Mochizuki and Ortega, 2008, Kawauchi, 2005). However, this is difficult to establish given the absence of systematic methods in the literature to measure participant proficiency in the L2 (see Section 2.1).

2. Does the impact of the four planning conditions on test scores and CAF results vary between picture-based narrative tasks and non-picture-based description tasks?

Skehan (2009) proposes that the extent to which a task obliges test takers to use specific language forms is a key indication of task difficulty. Picture-based narratives have a constraining effect, which may pose problems when test takers lack the requisite language to complete the task. For this reason, the impact of planning may vary between the two task types.

If the answer to research question 2 is affirmative,

2.1 Which task type and planning condition has the largest impact on test scores and CAF results?

4 Pilot studies

4.1 Introduction

This chapter reports the data collection, analytical procedures, and results of two pilot studies. The chapter includes information about the development of two EBB rating scales (‘The scale is *empirically* derived, requires *binary* choices by raters, and defines the *boundaries* between score levels (EBB)’, Turner and Upshur, 1996, pp. 60-61), rater training on the EBB scale and the analytic scale (Iwashita et al., 2001), and score analysis. It provides information about the choice of complexity, accuracy and fluency (CAF) measures and the statistical procedures adopted in the analysis. The results are discussed and the implications of the findings for the main study are set out.

4.2 Pilot 1

A pilot study was designed to trial the data collection and analytical procedures. This process involved trials of two picture-based narrative tasks, the data recording software, the method of transcription, the CAF measures, the EBB scale development methodology, the analytic scale and the multi-faceted Rasch measurement (MFRM). Based upon the review of the literature, the following research questions were formulated:

4.2.1 Research questions for Pilot 1

1. Does a ten-minute planning condition lead raters to award higher scores on an EBB rating scale in relation to a one-minute planning condition?
2. Does a ten-minute planning condition lead raters to award higher scores on an analytic scale in relation to a one-minute planning condition?
3. Does a ten-minute planning condition lead to gains in measures of complexity, accuracy and fluency (CAF) on a language test in relation to a one-minute planning condition?

4.2.2 Methodology

This pilot study was undertaken to assess the impact of including different lengths of pre-task planning time (ten minutes and one minute) in a test of second language speaking. Two approaches to assessment were used: a CAF analysis and rater scores based on an EBB scale and an analytic scale. Data were collected and analysed at two stages. During the first stage, 17 test taker participants were recruited in order to produce test samples for use during the EBB scale development (see Section 4.2.2.4). However, the analytical approach adopted in the study involved tests of statistical significance and it was important for the results of these tests to be reliable to inform the design of the main study procedures. Using small sample sizes (i.e. less than 30 participants) results in measurement error, which prevents reliable interpretation of test scores (Van Voorhis and Morgan, 2007). It was necessary to increase the number of participants beyond the original 17 to complete statistical analysis of test scores. The number of test takers was therefore increased to 30 upon completion of the EBB scale. This necessitated a further increase in the number of raters. These procedures are discussed in detail in the following sections.

4.2.2.1 Participants

Test Takers. During the first stage of data collection, 17 participants took part in the study. During the second stage, an additional 13 participants took part in the study. In total, 30 participants took the test. Ages ranged from 18 to 24 (mean age 19.4, SD 1.6). All participants were enrolled in the English preparatory program in a university in western Turkey. At the time of data collection, participants had been assessed on an in-house proficiency exam containing reading, listening, writing, and speaking components and had been placed into courses designed for pre-intermediate, intermediate and upper intermediate levels of proficiency. These levels were designed to correspond to the levels A1+, A2, and B1 on the Common European Framework (Council of Europe, 2001). All participants signed an approved consent form informing them about the purpose of the study (see Appendix 2). This consent form was used in both piloting and the main study.

Raters EBB Scale 1. Raters were recruited from a pool of teachers working at the institution. In total, 13 EBB raters were involved in the study. The raters' ages ranged from 23 to 35 (mean = 28.3, SD = 3.97) and each had between two and seven years of teaching experience (mean = 4.85, SD = 1.77). Raters 1 to 7 contributed to the EBB scale design and graded the original 34 speech samples elicited from the original 17 test takers. These raters are referred to as EBB scale constructors. Raters 8 to 13 were recruited at a later stage of the project. These raters are referred to as standardised raters: they did not contribute to the EBB scale development but were standardised to the scale. During standardisation, Raters 8 to 13 were presented with the EBB rating scale, informed about the development procedures (see Section 4.2.2.4), and discussed the contents of the scale as a group. Following this, the raters listened to three speech samples representing the lowest, mid and highest levels in the database (based on the original analysis) and independently awarded a grade to each sample. Once the independent grading was complete, the raters compared their scores and discussed the basis for their decisions. Standardisation was complete once the group agreed upon suitable grades for the three samples.

The literature indicates that one key concern when implementing EBB scales is that reliable usage is dependent upon scale users being involved in the scale construction process (Turner and Upshur, 2002). In order to account for this, a method was required to compare the levels of reliability (defined as agreement between raters and consistency within raters) between the EBB scale constructors and the standardised raters. This was achieved by having different raters grade the same test samples (see Section 4.2.2.6). Raters 8 to 13 graded the additional 26 recordings and the original 34 recordings (the total number of grades awarded by each rater ranges from 12 to 17). This measure was taken in order to provide common scores that would enable comparisons between raters to be made in the MFRM (see Section 4.2.2.6) and to assess the impact of introducing standardised raters to the analysis (see Table 5).

Table 5 Distribution of participants' test samples between EBB scale 1 raters

Participants	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	R11	R12	R13
1-5	X	X	X	X	X	X	X	X			X		
6-10	X	X	X	X	X	X	X			X		X	
11-15	X	X	X	X	X	X	X		X				X
16-20	X	X	X	X	X	X	X	X	X	X	X	X	X
21-25								X	X	X	X	X	X
26-30								X	X	X	X	X	X

Raters Analytic Scale. Four raters were recruited to complete the assessment using the analytic rating scale (Iwashita et al., 2001). The raters were instructors of English at the university and regularly took part in speaking assessment. Their ages ranged from 35 to 50 (mean = 43.3, SD = 7.9) and their teaching experience ranged from two to 25 years (mean = 15.3, SD = 11.2). The raters first took part in a standardisation session involving three test samples representing the low, mid and high ability levels. The raters assigned scores to the original 34 samples. However, these raters could not commit to grading the remaining 26 samples during the second round of rating. The total count of ratings therefore varies between the raters. Rater 1 provided scores for each test sample (*n* = 60). Raters 2 and 3 each provided additional grades for 15 different samples (*n* = 49). Rater 4 did not participate in the second round of rating (*n* = 34).

4.2.2.2 Tasks

Two picture-based narrative tasks (accessed on 21 September, 2014 https://bwcdigital.wordpress.com/tag/wordless/) were selected to elicit speech samples (see Figures 5. Pilot 1 Task 1 and 6. Pilot 1 Task 2). Based on an analysis of the task content and discussions with teachers employed by the institution, these tasks were deemed suitable for the test takers because the material a) was not culturally specific; the use of culturally unfamiliar content increases task difficulty (Fulcher, 2003) and b) required lexis predicted to be familiar to test takers; the absence of relevant vocabulary in the test taker's repertoire to describe obligatory content increases task difficulty (Skehan, 2009). Test takers completed each task under different planning conditions: a one minute-planning condition and a ten-minute planning condition. Decisions relating to planning conditions and task type were made based on frequency in the literature (see Section 2.2). Tasks and planning conditions were counterbalanced between the test takers. Test takers were permitted to take notes during the planning stage and were informed that the notes would be removed before the test taker began speaking. This decision was made to prevent test takers reading directly from the notes. This measure was adopted in both piloting and the main study.

Figure 5. Pilot 1 Task 1.

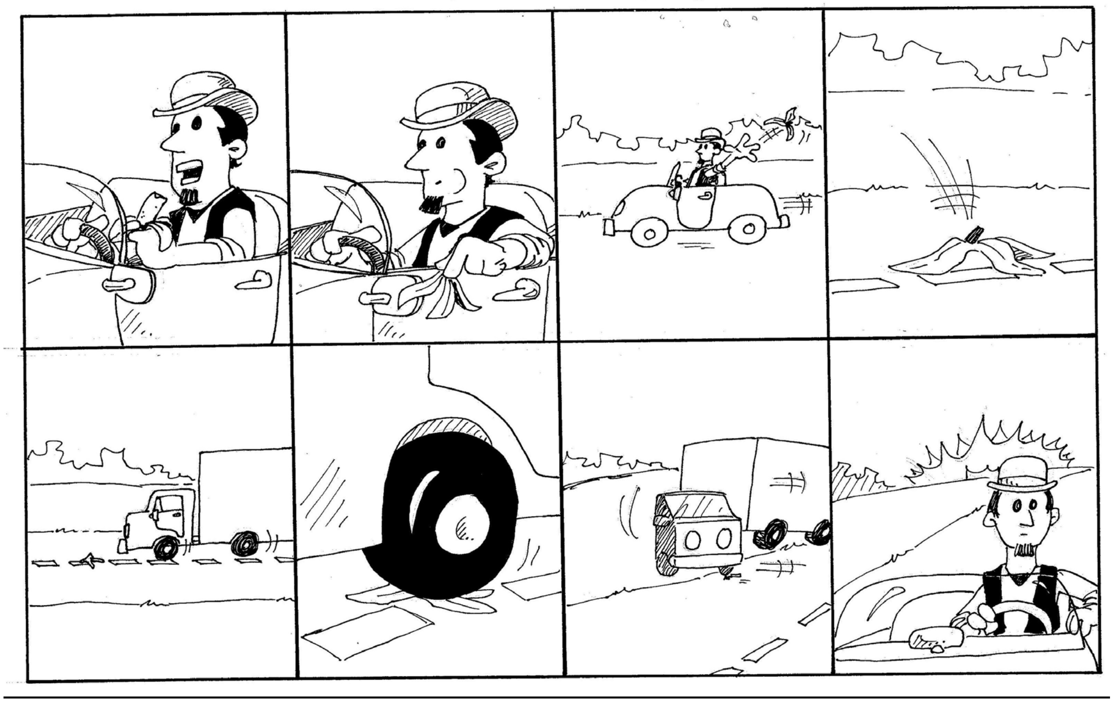
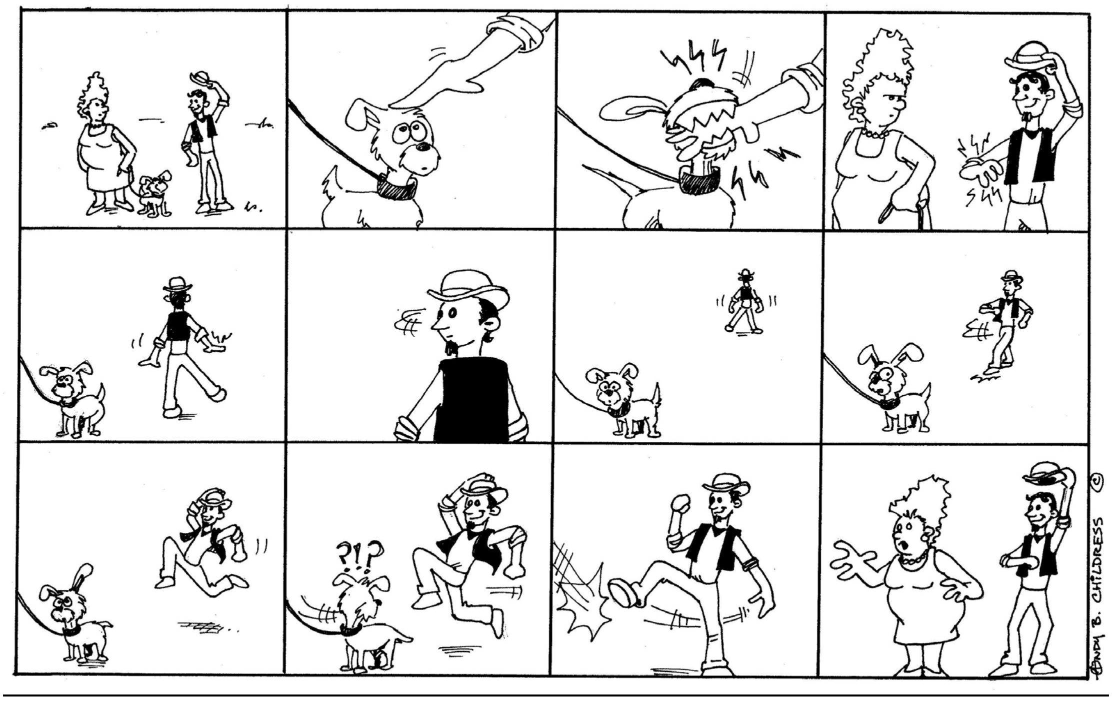


Figure 6. Pilot 1 Task 2.



4.2.2.3 Recording and transcription

The tests were recorded using Audacity (2.0.6, 29 September 2014, http://audacity.sourceforge.net) and saved as Audacity files and in MP3 format. Audacity files were transcribed for analysis. The transcription format was adapted from Fulcher and Davidson (2007), Jefferson (2004) and from examples described in Foster et al. (2000) as these were viewed to be the most appropriate notation methods for the current purpose. Reliability of transcripts was assessed with a second transcriber transcribing a proportion of the total. Following Brown et al. (2005, p. 63), 10 per cent of the data was transcribed and coded for all measures. Inter-coder reliability (TOTAL AGREEMENT/ n x 100) was 93.3 per cent. In order to run syntactic analysis, transcripts were analysed as AS-units: a 'single speaker's utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with either' (Foster et al., 2000, p. 365). Performance was assessed with measures of CAF (see Section 4.2.2.7). The CAF measures were selected to replicate those used in the existing planning literature (see Section 2.7.2) and thereby enhance comparison between the findings.

4.2.2.4 EBB scale development procedures

The scale was constructed following guidelines in Turner and Upshur (1996). Eight samples were selected holistically on the basis that they represented the range of abilities in the samples. The eight samples were then rank ordered by the group. The group decided that the eight samples represented five levels of language ability. A series of paired comparisons was then completed on the samples, which is described at length in the remainder of this section.

The group decided that the primary feature separating the higher half from the lower half of the samples was the amount of speech the test taker was able to produce. The consensus was that the lower half samples were unable to produce a sufficient amount of speech to communicate the events of the narrative. The group then began wording a yes/no question to describe this feature: *Does the speaker produce enough speech to describe the story?*

During the next stage, the group compared samples at levels 4 and 5. The group described higher levels of accuracy in grammar and vocabulary use, lack of repetition and hesitation and more control over pronunciation at level 5. The following question was then formulated: *Does the speaker demonstrate accurate use of grammar and vocabulary without repetition or hesitation while maintaining good pronunciation?*

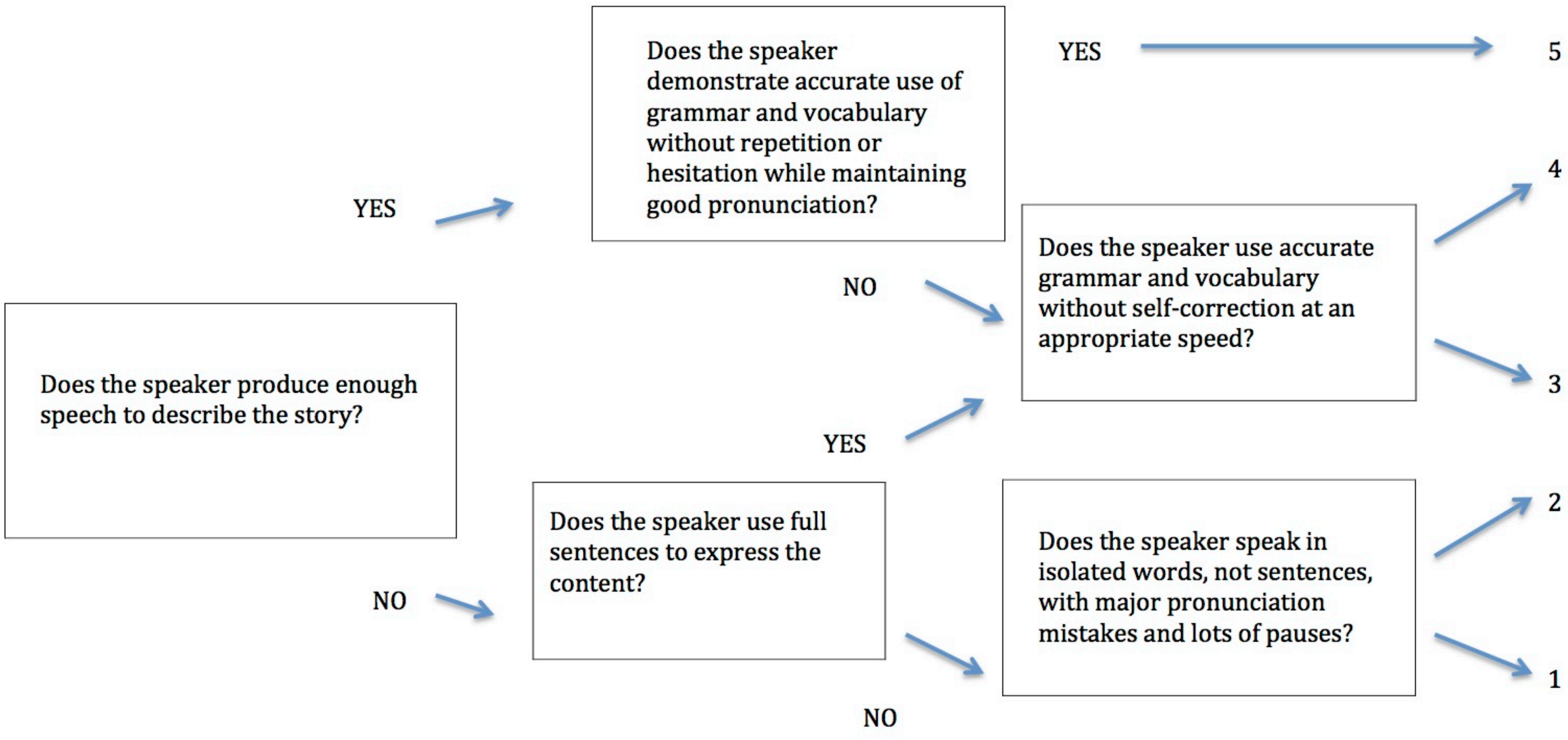
Comparison of levels 3 and 4 revealed that level 4 demonstrated stronger command over grammar and vocabulary, did not self-correct and spoke with a faster speech rate. The following question was then developed: *Does the speaker use accurate grammar and vocabulary without self-correction at an appropriate speed?*

Comparison of levels 3 and 2 revealed that at level 2 test takers were more likely to speak in isolated phrases than what the group termed ‘sentences’. The following question was therefore created: *Does the speaker use full sentences to express the content?*

The final comparison between levels 1 and 2 revealed that at level 1, test takers spoke in isolated words, paused frequently and had a low level of control over pronunciation. The following question was then formulated: *Does the speaker speak in isolated words, not sentences with major pronunciation mistakes and lots of pauses?*

It is clear that a proportion of the scale content does not correspond closely to current theoretical models of the spoken utterance (Foster et al., 2000). For instance, the notion that speakers speak in sentences has been widely rejected. Likewise, whether test takers at level 5 had indeed spoken without hesitation may be questionable. However, it is important to note that the descriptors were created through an analysis of spoken samples and should therefore be considered to be representative of the ways in which the group perceived the speech samples. As such, the descriptors do involve a degree of psychological reality for the raters.

Figure 7. EBB scale 1


Does the speaker produce enough speech to describe the story?
YES
Does the speaker demonstrate accurate use of grammar and vocabulary without repetition or hesitation while maintaining good pronunciation?
YES
5
NO
Does the speaker use accurate grammar and vocabulary without self-correction at an appropriate speed?
4
3
NO
Does the speaker use full sentences to express the content?
YES
NO
Does the speaker speak in isolated words, not sentences, with major pronunciation mistakes and lots of pauses?
2
1

4.2.2.5 Analytic scale procedures

An analytic scale developed by Iwashita et al. (2001) was selected (see Section 2.7.3). The scale was originally developed to assess the impact of variation in the cognitive demands of a narrative task. Five band levels of proficiency are described for complexity, accuracy and fluency that range from beginner to advanced level language user (see Appendix 1). The scale has been used in research on planning (Elder et al., 2002, Elder and Iwashita, 2005, Nitta and Nakatsuhara, 2014) and therefore is considered to represent a valid alternative to the EBB scale with potential to compare results between studies.

4.2.2.6 MFRM analysis

A score on a language test represents more than a test taker's language ability (McNamara, 1996). Test scores are the product of interactions between test elements (e.g. the ability of the test taker, the severity and consistency of the rater, the difficulty of the task). Test elements are collectively referred to as facets. In the current study, the facets under investigation are test taker ability, difficulty of the task that the test taker completed, severity of the rater that awarded a score to the performance, and the amount of planning time that was available to the test taker. Multi-faceted Rasch measurement (MFRM) constructs a probabilistic model of the interaction between test facets (Linacre, 2013). In the current study, the interaction between test facets is specified in the following MFRM model (adapted from Linacre, 2013):

$$\log \left(P_{nijsk} / P_{nijsk-1}\right) = B_n - D_i - R_j - -G_s - F_k$$

B_n = ability of test taker n
D_i = difficulty of task i
R_j = severity of rater j
G_s = time of planning s
F_k = difficulty of category k relative to k -1
$P_{nijlmsk}$ = probability of receiving rating k under these circumstances
$P_{nijlmsk-1}$ = probability of k-1

MFRM calibrates the test data to share one common interval scale, known as the logit scale, that rank orders the elements of the test facets (i.e. test takers, raters, tasks) and expresses the distance between these elements on the scale (i.e. in terms of ability, severity, and difficulty). The elements of the test facets are assigned logit measure values, which represent a probabilistic score on the logit scale. This means that in addition to establishing ability levels for test takers, severity levels for raters, and difficulty levels for tasks, the different planning conditions can be mapped onto the logit scale to identify the condition that leads to the highest scores.

MFRM can be conducted using the computer program *Facets* (3.71.4, 18 January 2014, www.winsteps.com). In order to run MFRM using *Facets*, sufficient connectivity between the facets in the model is required. This is typically achieved by having different raters assign scores to the same test takers so that the test data contains various observations of the same performance.

In addition to constructing the logit scale, *Facets* converts the logit measure values back into scores on the original rating scale (EBB and analytic). These scores are termed 'fair average' scores (Linacre, 2013, p. 272). The fair average is the score on the original rating scale adjusted for the measures (e.g. levels of severity, task difficulty) of the test elements that combined to produce the score. The fair average is the score that would be awarded if all elements of the facets had equal measure values (e.g. the raters were equally severe, the tasks were equally difficult).

Facets measures the extent to which the test data conforms to the probabilistic MFRM model and hence how much trust can be placed in the MFRM results (Linacre, 2013). These measures are termed fit statistics. In the language testing literature, the most commonly reported fit statistic is the infit mean-square statistic. Infit mean-square statistics report the extent to which the observed results match the predictions made by the MFRM model. Various ranges of infit mean-square acceptability have been suggested in the literature ranging from 0.5 to 1.5 (Lunz and Stahl, 1990) to .7/.8 to 1.3/1.2 (Linacre, 1993). Linacre (2013, p. 266) provides the following guidelines for interpreting fit statistics:

- >2.0 Distorts or degrades the measurement system.
- 1.5 - 2.0 Unproductive for construction of measurement, but not degrading.
- 0.5 - 1.5 Productive for measurement.
- <0.5 Less productive for measurement, but not degrading.

The infit mean-square statistics may indicate misfit of the data to the model or overfit of the data to the model. Values above 2.0 signify misfit and indicate that scores are unpredictable. Values below 0.5 indicate model overfit and show that observed results closely match predicted results. Test takers and raters that demonstrate substantial misfit distort the model (i.e. there is too much randomness to generate reliable information) and may have to be removed from the analysis. However, McNamara (1996) suggests that model overfit is not generally regarded as a problem in assessment involving rater judgement because predictability concerning test taker scores is a desired feature of the test. In contrast, significant test taker overfit in a multiple-choice test may be indicative of test taker guessing, or a poorly constructed test.

The *Facets* output provides a model standard error statistic for each measurement. The standard error statistic is expressed in logits and measures the precision of the predictions made by the MFRM. The standard error value describes the shortest distance on the logit scale before differences between measures can be thought of as important (Linacre, 2013). For example, with a standard error of .10, a difference of .05 logits between test takers on the logit scale may be due to measurement error rather than differences in test taker ability. Winke, Gass and Myford (2012) observe that high standard error values tend to occur when samples receive a low number of ratings. This is because the model may not have sufficient data to make precise predictions, which may limit the certainty of the results. This limitation may account for misfit in the results of the EBB scale analysis as standardised raters provided fewer scores than the EBB scale constructors (see Section 4.2.2.1).

In addition to the standard error, *Facets* provides information regarding the levels of separation within the test facets in the MFRM model. The separation index and strata statistics report the number of statistically distinct levels of test performance within a given population (Linacre, 2013). Separation index is the number of statistically distinct levels in a normally distributed sample when the ends of the normal distribution are assumed to be a result of measurement error. The strata value is the number of statistically distinct levels when the ends of the normal distribution represent real levels of performance (e.g. very high and very low scoring test takers, and very severe and very lenient raters). This study reports both the separation index and strata values. A measure of the reliability of test taker and rater separation is also provided in the *Facets* output. Unlike conventional reliability statistics, which report the extent to which facets (e.g. the severity of raters) are "reliably the same" (e.g. consistent agreement between raters), the *Facets* reliability statistics report how "reproducibly different the measures are" (Linacre, 2013, p. 314). The reliability statistic ranges from 0 to 1, with values of 1 indicating reliable separation. High reliability of difference between test takers (near 1.0) is desired because the test should be shown to reliably separate the test takers. Conversely, low reliability of difference between raters (near 0) is ideal because raters should not demonstrate consistently different levels of severity. *Facets* calculates a fixed chi-square test on the data to test whether the facets share the same measure after the amount of measurement error has been accounted for (Linacre, 2013). For example, the results of the chi-square test indicate whether the difference in difficulty between planning conditions is statistically significant. If statistical significance is reached, the differences in scores between planning conditions are not due to measurement error and planning has had an impact on test scores.

In the current pilot study, the MFRM of the EBB scale data involved four test facets: test taker ability, rater severity, task, and planning time. Raters 8 to 13 were marked (with a dash: -) during the data input to distinguish the standardised raters from the scale constructors in the *Facets* output. Because the standardised raters provided fewer grades than the EBB constructors (see Section 4.2.2.1), their error values are higher (Winke et al., 2012). In contrast to the MFRM of the EBB scale data, the MFRM of the analytic scale data involved five facets: test taker ability, rater severity, task, planning time and rating scale category (complexity 1-5, accuracy 1-5 and fluency 1-5). Following the initial MFRM of the analytic scale data, a series of independent MFRM analyses were completed on each category of the scale, complexity, accuracy and fluency to generate statistical information regarding the impact of the planning conditions on these categories.

4.2.2.7 CAF measures

Based upon the review of the literature (see Section 2.7.2), the following measures of complexity, accuracy and fluency were used in the study. This section presents each measure, a brief description of the measure and the purpose of its use in the study.

Complexity

- Guiraud's Index (G.INDEX). This measure provides an indication of lexical density. The traditional approach to measuring lexical density is type-token ratio. However, text length has been shown to influence type-token ratio significantly (Kuiken and Vedder, 2007). Guiraud's Index bypasses this limitation with a mathematical equation that mitigates the effect of text length. It is calculated by dividing the number of types by the square root of the number of tokens.
- Clauses per AS-unit (CAS). This measure describes the amount of subordination and coordination that occurs in the text. It is an indication of syntactical variety.

Accuracy

- Percentage of error free AS-units (EFAS). This measure describes the ratio between AS-units that contain errors and those that are error free. It is calculated by identifying the total number of error free AS-units and dividing this by the total number of AS-units.
- Mean number of errors per AS-unit (MNE). This measure is calculated by dividing the total number of errors by the number of AS-units. It provides a broader impression of grammaticality than is possible with error free AS-units.
- Percentage of verbs with correct agreement (AGR.). Total number of verbs with correct agreement divided by the total number of verbs supplied.
- Percentage of correct article use (ART.) Total number of correctly supplied articles divided by the total number of obligatory instances. In combination with the previous measure (AGR.), specific measures identify structures that cause inaccuracies in the speech.

Fluency

- Speech Rate (SPR). Total words divided by total time multiplied by 60 and expressed in seconds. Speech rate indicates the speed of the speech.
- Phonation-time ratio (PTR). This measure is expressed as a percentage of the amount of time the speaker spent pausing during the task. It is a measure of 'breakdown fluency' (Skehan, 2009) and indicates the extent to which on-line planning was necessary during the task. Pauses were calculated as a length of consecutive silence in the recording in excess of 0.25 seconds (Kormos, 2006). This standard was preferred to Foster and Skehan's (1996) criteria of one second because 0.25 is an established standard in SLA research (Kormos,

2006) that generates a more precise impression of the way planning impacts speech fluency.

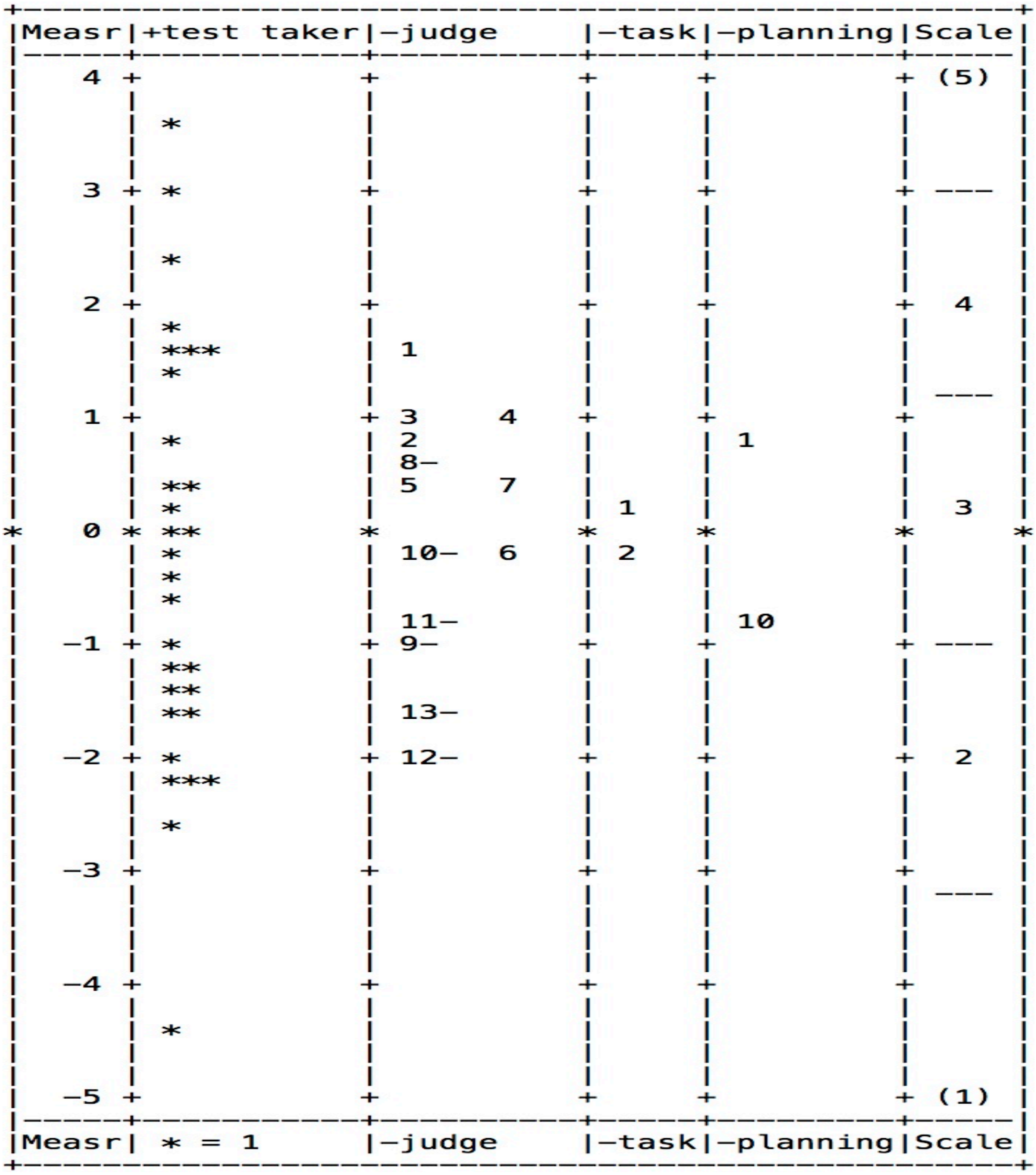
- Mean number of filled/unfilled hesitations per AS-unit (MNH). Hesitation is indicative of controlled processing and suggests that the speaker is having difficulty producing speech. This is in contrast to pausing, which occurs between syntactic boundaries and is a common feature of fluent speech (Field, 2011). Hesitation takes the form of filled and unfilled gaps in the speech between clauses
- Mean length of utterance between filled/unfilled pauses/hesitations in words (MLU). This is a measure of the number of words a speaker produces between pauses and hesitations. This is an important indicator of fluency. Length of utterance is an indication of the extent to which the processes of retrieval and encoding of the speech have become proceduralised (Field, 2011). As such it is a key signal of developing proficiency.

Shapiro-Wilks tests (*p* = .05) demonstrated that five of the ten measures were non-normally distributed (see Section 4.2.6). Paired samples t-tests were carried out on the normally distributed results using Bonferroni adjusted alpha levels of *p* = .01 (.05/5). Wilcoxon signed-rank tests were completed on the non-normally distributed results using Bonferroni adjusted alpha levels of *p* = .01 (.05/5). Using an adjusted alpha level ensures against type one error, which is important when running multiple statistical tests (see Section 2.7.2.4). However, using an adjusted alpha level also increases the chances of encountering type two error. Nonetheless, given the number of tests that were completed the adjusted alpha level was deemed appropriate.

4.2.3 Results: EBB scale

Figure 8 presents the results of the MFRM in the form of a Wright map. The map summarizes test scores and indicates differences in test taker ability, rater severity, task difficulty and the effect of planning time. The first column represents the measurement scale and expresses differences between the facets in terms of logits (see Section 4.2.2.6). The logit scale is an interval scale that ranks the separate elements of the facets (i.e. ability, severity, task and planning condition) and indicates the degree of variation within each facet (McNamara, 1996). The second column presents the range of ability in the test taking population. Test takers are represented with an asterisk and higher scoring test takers are located toward the upper end of the map. The test takers are spread between -4 and +4 on the logit scale, which suggests that there was a broad range of test taker abilities in the test taking population. The second column represents the range of rater severity. Raters are identified with numbers and more severe raters are located toward the higher end of the map. The raters are situated between -2 and +2 on the logit scale. The third column reports differences in task difficulty. Task 1 is located toward the top of the map indicating that this task recorded the lowest overall scores. The fourth column reports the difference between the planning conditions. The one-minute planning condition recorded the lowest scores and is situated above the ten-minute planning condition. The final column represents the five levels of ability on the EBB scale. Level 5 is presented in parenthesis indicating that the facets do not reach this level of ability, severity or difficulty. Summary statistics for each facet are discussed at length in the following sections.

Figure 8. Wright map EBB scale 1



```
+----------------------------------------------------+
|Measr|+test taker|-judge     |-task|-planning|Scale|
|-----+-----------+-----------+-----+---------+-----|
|   4 +           +           +     +         + (5) |
|     |           |           |     |         |     |
|     | *         |           |     |         |     |
|     |           |           |     |         |     |
|     |           |           |     |         |     |
|   3 + *         +           +     +         + --- |
|     |           |           |     |         |     |
|     |           |           |     |         |     |
|     | *         |           |     |         |     |
|     |           |           |     |         |     |
|   2 +           +           +     +         +  4  |
|     | *         |           |     |         |     |
|     | ***       | 1         |     |         |     |
|     | *         |           |     |         |     |
|     |           |           |     |         | --- |
|   1 +           + 3    4    +     +         +     |
|     | *         | 2         |     | 1       |     |
|     |           | 8-        |     |         |     |
|     | **        | 5    7    |     |         |     |
|     | *         |           | 1   |         |  3  |
*   0 * **        *           *     *         *     *
|     | *         | 10-  6    | 2   |         |     |
|     | *         |           |     |         |     |
|     | *         |           |     |         |     |
|     |           | 11-       |     | 10      |     |
|  -1 + *         + 9-        +     +         + --- |
|     | **        |           |     |         |     |
|     | **        |           |     |         |     |
|     | **        | 13-       |     |         |     |
|     |           |           |     |         |     |
|  -2 + *         + 12-       +     +         +  2  |
|     | ***       |           |     |         |     |
|     |           |           |     |         |     |
|     | *         |           |     |         |     |
|     |           |           |     |         |     |
|  -3 +           +           +     +         +     |
|     |           |           |     |         | --- |
|     |           |           |     |         |     |
|     |           |           |     |         |     |
|  -4 +           +           +     +         +     |
|     |           |           |     |         |     |
|     | *         |           |     |         |     |
|     |           |           |     |         |     |
|     |           |           |     |         |     |
|  -5 +           +           +     +         + (1) |
|-----+-----------+-----------+-----+---------+-----|
|Measr| * = 1     |-judge     |-task|-planning|Scale|
+----------------------------------------------------+
```

4.2.3.1 Facets in the MFRM model

Test takers. The MFRM statistics show a substantial degree of variation between the test takers' ability measures. This variation is demonstrated in the fair average values. The fair average is a contextualised average score on the original rating scale that accounts for all of the facets in the analysis that influence a test taker's score (i.e. rater severity, task difficulty, planning time). The range of test taker ability was from 1.20 to 4.88 by fair average on the EBB scale. This indicates a wide range of test taker ability in the test taking population. The separation index was 2.93 and test takers were separated in to 4.24 separate strata (see Section 4.2.2.6). This strata value corresponds to the five levels of ability described in the EBB scale and indicates that there were between four and five statistically distinct levels of test taker ability in the test taking population. The mean of the model standard error was .54 (standard deviation .21). Accordingly, grades reported in the analysis may be imprecise by half a level on average. The reliability was high at .90 indicating that the separation of the test takers was reliable. Reliable separation of test takers is necessary to assess the impact of the planning conditions on test scores: interpretation of the planning facet would be invalid if the separation of test takers was due to chance.

In the model, test taker fit statistics varied substantially from .10, showing considerable model overfit, to 1.92, showing model misfit. The degree of variance may be due to the fact that many of the samples received a minimal number of scores. This poses a problem for the interpretation of the test scores. When the total count of scores varies, it may be the case that misfitting infit values are due to rater behaviour that is inconsistent with the model (i.e. a particularly severe rater awards an unpredictably lenient grade that did not match the grades awarded by other raters). However, this does not seem to be the case here: a test taker who received 17 grades recorded the highest infit value of 1.92. Test taker 17 raw scores and rater information are provided in Table 6 (lenient raters have lower measures).

Table 6 Scores awarded to test taker 17 by nine raters

Rater	1 (1.61)	2 (.77)	3 (.91)	4 (.91)	5 (.49)	6 (-.11)	7 (.49)	8 (.63)	9 (-.98)
Scores	2/3	2/2	2/1	1/1	2/2	1/1	1/1	1/2	3

Measure values appear in parenthesis below rater ID.

Test taker 17 received relatively high scores from the most severe rater (Rater 1) and low scores from a lenient rater (Rater 6). As a result, the infit values for this test taker were high. However, as Bachman (2004, p. 147) explains, fit statistics above ‘2.0’ are cause for concern and may necessitate removal of data. As the level of misfit does not exceed 2.0, the test taker’s scores were retained.

Raters. Rater statistics are presented in Table 7. The range of severity measures, a measure of the relative severity and leniency in the rater population and expressed in logits (see Section 4.2.2.6), is -1.97 to 1.61 on the logit scale. The rater fair average is the mean of the rater’s scores on the rating scale. The mean is adjusted to account for differences between the ability levels of the test takers in that rater’s sample and the overall ability measures in the test taking population (see Section 4.2.2.6). For example, if a rater is assigned disproportionately high numbers of high scoring test takers, the fair average is corrected to reflect the variation of scores in the entire test taking population. The raters’ fair average values ranged from 2.04 to 3.83 on the EBB scale. This indicates a substantial degree of difference between the levels of rater severity in the model. The separation index was 3.09 and raters were separated into 4.45 strata. This means that there were approximately three levels of rater severity within the group of 13 raters according to the separation index and four levels of rater severity as measured by strata (see Section 4.2.2.6). The reliability statistic was .91 indicating that the differences in severity between the raters were reliable:

raters were consistent in their degree of severity and leniency when assigning scores to the tests.

Infit mean-square statistics indicate the levels of predictability in the MFRM of rater's scores. Values below 0.5 indicate model overfit and suggest that the rater is not using the full extent of the rating scale due to conservatism. Values above 1.5 indicate model misfit and suggest that the distribution of scores is erratic and inconsistent (see Section 4.2.2.6). The infit mean statistics ranged from .50 to 1.38. These values indicate that the raters assigned grades with an acceptable level of consistency. The mean standard error value was .31 and the range was from .26 to .42. The wide range of standard error values appears to be due to the difference between the number of grades the standardised raters awarded and the number the EBB scale constructors awarded. The EBB scale constructors (Raters 1 to 7) provided a higher number of grades (34) than the standardised raters (Raters 8 to 13 provided 12 to 17 grades). The EBB constructors recorded standard error statistics that were lower than the standardised raters'. The standard error range for the EBB scale constructors was from .26 to .29. In contrast, the standardised raters recorded standard error values that ranged from .36 to .42. Clearly, raters that did not contribute to the EBB scale construction and awarded fewer grades demonstrated higher standard error values (see Section 4.2.2.6). An implication of this finding is that in order to enhance rater consistency when using EBB scales, all raters should be involved in the scale construction.

Table 7 Report of rater severity: EBB scale 1

Rater	Fair Average	Severity Estimate	Error	Infit mean-square index
12-	3.83	-1.97	.39	.65
13-	3.57	-1.56	.38	.50
9-	3.26	-.98	.37	1.31
11-	3.21	-.89	.36	.58
10-	2.92	-.28	.42	.52
6	2.84	-.11	.26	1.08
5	2.57	.49	.26	.93
7	2.57	.49	.26	.74
8-	2.51	.63	.26	.88
2	2.44	.77	.27	1.12
3	2.37	.92	.27	1.18
4	2.37	.91	.27	1.38
1	2.04	1.61	.29	.75
Mean	2.81	.00	.31	.89
SD	.51	1.03	.06	.29
Reliability of difference in severity of raters .91				

Tasks. Task statistics are presented in Table 8. The tasks varied in terms of difficulty by .42 logits. *Facets* calculates a fixed chi-square test of the hypothesis that the tasks vary in difficulty (see Section 4.2.2.6). The result of the fixed chi-square test showed that the difference between the tasks was statistically significant at p = .01. However, the difference in difficulty between the tasks was marginal and the counterbalancing of the tasks between the participants means that the difference does not affect the conclusions that can be drawn from the results.

Table 8 Tasks 1 and 2 fair average, measure and infit statistics: EBB scale 1

Task	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
2	2.89	-.21	0.86	χ^2 = 6.7, p = .01
1	2.69	.21	1.03	

Planning. Planning statistics are presented in Table 9. The planning condition clearly impacted test scores. The difference was .71 by fair average and the result of the fixed chi-square test was significant at p = .01. The fair average score awarded under the one-minute planning condition was mid-level 2 on the EBB scale. This increased to low-level 3 under the ten-minute planning condition. The hypothesis that planning would impact the test scores on the EBB scale is therefore confirmed.

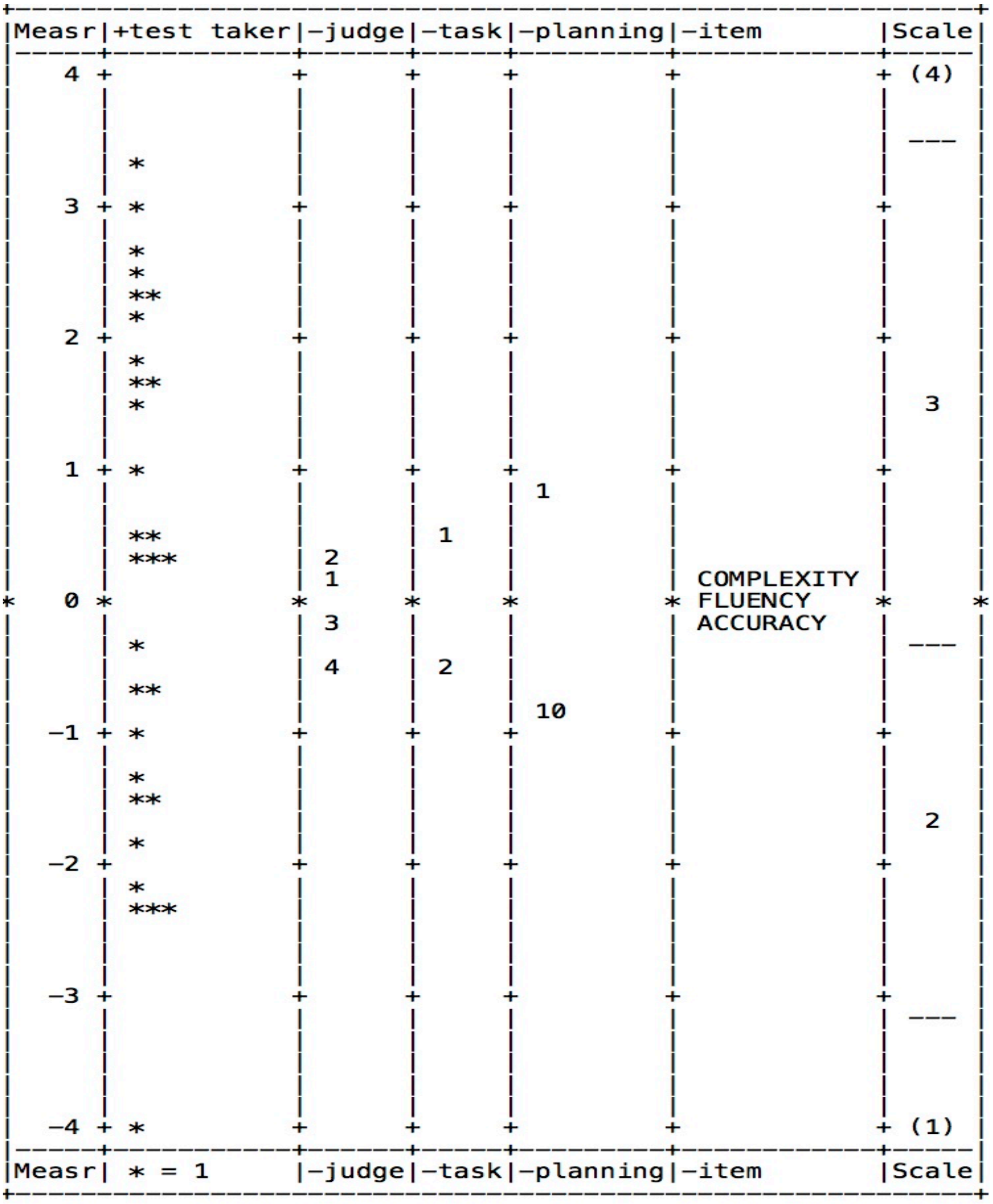
Table 9 Planning fair average, measure and infit statistics: EBB scale 1

Time	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min	3.15	-.77	0.97	χ^2 = 86.5, p = .00
1 min	2.44	.77	.91	

4.2.4 Results: analytic scale

Results of the MFRM (analytic) analysis are presented in Figure 9. The spread of test takers on the logit scale indicates a wide range of ability levels. Raters varied in their levels of severity between -1 and 1 logits. The map shows Task 2 was scored higher than Task 1 indicating that test takers may have found Task 2 less challenging. The length of planning time also impacted scores; the ten-minute condition resulted in scores that were higher than the one-minute condition. The items complexity, accuracy and fluency are mapped onto the logit scale in column six. Complexity was the most difficult category, followed by fluency and accuracy. The following section discusses the MFRM statistics at length.

Figure 9. Wright map analytic scale



```
+------------------------------------------------------------+
|Measr|+test taker|-judge|-task|-planning|-item        |Scale|
|-----+-----------+------+-----+---------+-------------+-----|
|   4 +           +      +     +         +             + (4) |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             | --- |
|     | *         |      |     |         |             |     |
|     |           |      |     |         |             |     |
|   3 + *         +      +     +         +             +     |
|     |           |      |     |         |             |     |
|     | *         |      |     |         |             |     |
|     | *         |      |     |         |             |     |
|     | **        |      |     |         |             |     |
|     | *         |      |     |         |             |     |
|   2 +           +      +     +         +             +     |
|     | *         |      |     |         |             |     |
|     | **        |      |     |         |             |     |
|     | *         |      |     |         |             |  3  |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|   1 + *         +      +     +         +             +     |
|     |           |      |     | 1       |             |     |
|     |           |      |     |         |             |     |
|     | **        |      | 1   |         |             |     |
|     | ***       | 2    |     |         |             |     |
|     |           | 1    |     |         | COMPLEXITY  |     |
*   0 *           *      *     *         * FLUENCY     *     *
|     |           | 3    |     |         | ACCURACY    |     |
|     | *         |      |     |         |             | --- |
|     |           | 4    | 2   |         |             |     |
|     | **        |      |     |         |             |     |
|     |           |      |     | 10      |             |     |
|  -1 + *         +      +     +         +             +     |
|     |           |      |     |         |             |     |
|     | *         |      |     |         |             |     |
|     | **        |      |     |         |             |     |
|     |           |      |     |         |             |  2  |
|     | *         |      |     |         |             |     |
|  -2 +           +      +     +         +             +     |
|     | *         |      |     |         |             |     |
|     | ***       |      |     |         |             |     |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|  -3 +           +      +     +         +             +     |
|     |           |      |     |         |             | --- |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|     |           |      |     |         |             |     |
|  -4 + *         +      +     +         +             + (1) |
|-----+-----------+------+-----+---------+-------------+-----|
|Measr| * = 1     |-judge|-task|-planning|-item        |Scale|
+------------------------------------------------------------+
```

4.2.4.1 Facets in the MFRM model

Test Takers. The MFRM statistics show that test taker ability measures varied. The range of fair average values was 1.28 to 3.45 on the analytic scale. The separation index was 3.91 and test takers were separated into 5.55 strata. The mean standard error was .46 (standard deviation .12) indicating that scores may have been imprecise by approximately half a logit on average. This value is high and indicates that the scores may have been imprecise. The mean value of infit mean-square index statistics was .98 (standard deviation .44). This represents a range of .51 to 2.20. The literature clearly indicates that values that exceed 2.0 are cause for concern (Bachman, 2004, Linacre, 2013). Test taker 3 (2.00) and test taker 8 (2.20) record unacceptable fit statistics according to this standard. In order to investigate the cause of the misfit, the test takers' raw scores are presented in Table 10.

Table 10 Scores awarded to test takers 3 and 8 by four raters

Test taker	Rater 1 (.17)	Rater 2 (.36)	Rater 3 (-.09)	Rater 4 (-.44)
3 (2.20)	4.3.3 / 3.3.3	2.2.2 / 2.1.1	3.3.3 / 3.3.2	4.2.3 / 4.3.4
8 (2.00)	3.3.4 / 3.2.2	4.3.4 / 2.2.2	4.4.3 / 3.3.2	4.4.4 / 3.3.3

Test taker 3 received high grades from Rater 1 who was more severe than Raters 3 and 4. In addition, Rater 4 awarded level 2 for accuracy to test taker 3, which appears to be inconsistent with the levels of lenience that are established for this rater in the model. Test taker 8 received high scores from the most severe rater (Rater 2) and relatively high scores from Rater 1. This level of inconsistency may explain the misfit of these scores to the MFRM model. However, in order to maintain equality between the number of test takers in the MFRM of the EBB data and the MFRM of the analytic data, these test takers were retained.

Raters. Table 10 presents the results of the MFRM of the rater data. The raters clearly varied in severity. The range of rater severity measures was from -.44 to .36 on the logit scale. The range of fair average values was from 2.57 to 2.80 logits. Standard errors ranged from .14 to .18 (mean = .16). The infit mean-square index values were within an acceptable range of 0.81 to 1.23, which indicates that the raters fit the model well. The separation index was 1.60 and raters were separated into 2.47 strata indicating that there were broadly two levels of rater severity. The reliability statistic was .72, which indicates reliable differences between the raters' distribution of scores. To compare the distribution of scores between the analytic scale raters (reliability .72) and the EBB scale raters (reliability .91), analytic scale raters were more likely to agree about suitable scores than the EBB scale raters: variation in rater severity was less likely to impact on test scores on the analytic scale.

Table 11 Report of rater severity: analytic scale

Rater	Fair Average	Severity Estimate	Error	Infit mean-square index
4	2.80	-.44	.18	.81
3	2.70	-.09	.16	.93
1	2.63	.17	.14	.84
2	2.57	.36	.15	1.23
Mean	2.67	.00	.16	.95
SD	.09	.30	.02	.17
Reliability of difference in severity of raters .72				

Tasks: Table 12 presents the results of the MFRM of the tasks. The results show that Task 2 was scored higher than Task 1 by .32 logits on the fair average scale. The result of the fixed chi-square test shows that this difference was significant at $p < .001$. The ordering of the tasks was similar in both the analytic and EBB analysis. The difference between the tasks indicates that different picture-based narrative tasks are required for the main study.

Table 12 Tasks 1 and 2 fair average, measure and infit statistics: analytic scale

Task	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
2	2.83	-.54	1.03	χ^2= 48.8, *p* = .00
1	2.51	.54	.89	

Planning. Table 13 presents the results of MFRM of the planning conditions. The difference in planning time clearly impacted the scores. The ten-minute condition led to a fair average value that was .51 higher than the one-minute value. The fixed chi-square test demonstrates that this difference was significant at $p < .001$. The hypothesis that planning would impact the scores on the analytic scale is therefore confirmed.

Table 13 Planning fair average, measure and infit statistics: analytic scale

Time	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min	2.90	-.87	.86	χ^2 = 126.4, *p* = .00
1 min	2.39	.87	1.05	

Complexity, Accuracy, Fluency. The score frequencies per category are presented in Table 14. The table shows that raters clearly avoided awarding level 5 to the test takers on all categories. There also appears to be a central tendency effect in operation as the third level on the scale was clearly the most frequently used (Myford and Wolfe, 2004).

Table 14 Frequencies of scores on analytic scale categories

Score	Accuracy	Complexity	Fluency
1	26	28	29
2	59	67	56
3	82	79	83
4	21	14	19
5	0	0	0

In addition to the overall MFRM of the analytic scale data, separate MFRM analyses were run to determine the planning impact on each category of the analytic scale. Table 15 presents the results. To begin with fluency, the ten-minute condition resulted in scores that were .56 logits higher on the fair average scale than the one-minute condition. The result of the fixed chi-square test demonstrates that this difference was significant at p < .001. On the accuracy category, the ten-minute planning condition recorded a fair average value that was .43 higher than the one-minute condition. The fixed chi-square test shows that this result was significant at p < .001. On the complexity category the ten-minute condition led to fair average values that were .55 higher than the one-minute condition on the fair average scale. The result of the fixed chi-square test was significant at p < .001. These results clearly demonstrate that extra planning time increased scores in each category of the analytic rating scale. The largest increases in scores occurred on the fluency and complexity categories.

Table 15 Complexity, accuracy, fluency fair average, measure and infit statistics

Category	Planning	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
Fluency	10 min	2.97	- 1.05	1.01	χ^2 = 54.7, *p* = .00
	1 min	2.41	1.05	.92	
Accuracy	10 min	2.94	-.86	.89	χ^2 = 37.7, *p* = .00
	1 min	2.51	.86	1.02	
Complexity	10 min	2.87	-1.09	.74	χ^2 = 54.0, *p* = .00
	1 min	2.32	1.09	1.14	

4.2.5 Ordering of the test takers in the scales

Table 16 compares the ordering of the test takers in the MFRM of the EBB scale data and the analytic scale data. The rank ordering of the test takers differs between the two scales. This indicates that raters may have assigned scores based on different features of the test taker speech depending on the scale that was used. This is not surprising as the scales contain different criteria. However, despite the differences between the scales, the addition of extra planning time increased scores on both scales.

Table 16 Ranking of the highest and lowest scoring test takers in the scales

EBB lowest five test takers	Analytic lowest five test takers	EBB highest five test takers	Analytic highest five test takers
4	4	29	19
10	17	16	29
13	5	19	12
11	10	30	23
17	14	8	30

4.2.6 CAF results

Shapiro-Wilks tests (*p* = .05) and an evaluation of their histograms demonstrated that the following variables were non-normally distributed: mean number of hesitations per AS unit, mean length of utterance, percentage of error free AS units, mean number of errors per AS unit and mean number of clauses per AS unit. The results of the remaining variables were normally distributed: speech rate, phonation time ratio, Guiraud's index, percentage of correctly supplied articles in obligatory contexts, percentage of correctly supplied verb forms in obligatory contexts. The statistical analysis of CAF results required two different tests of significance to account for the differences in distribution. Paired samples t-tests were completed on the normally distributed measures and Wilcoxon signed-rank tests were completed on the non-normally distributed measures. A Bonferroni adjusted alpha level of *p* = .01 (.05/5) was set for the t-tests and the Wilcoxon signed-rank tests.

Results of the paired samples t-tests are presented in Table 17. The results demonstrate a statistically significant impact of the planning variable on speech rate (*t*(29) = 5.75, *p* < .001) (SPR) and phonation time ratio (t(29) = 3.27, *p* = .003) (PTR). Speech rate results demonstrate an increase of 14.5 words per minute when planning time was extended to ten minutes. The results of the phonation time ratio show that planning served to increase the percentage of time test takers spent producing speech in relation to time spent in silence by approximately 6 per cent. The remaining measures did not reach the adjusted level of statistical significance. However, the results of Guiraud's Index did approach statistical significance (*p* = .024). Guiraud's Index results demonstrate that increased planning may have improved the variety of lexis that test takers were able to utilize during the task. However, this increase was relatively minimal at .41.

Table 17 T-tests: CAF measures

	Planning				
	1 minute	10 minutes	T	df	p
SPR	62.31 (21.83)	76.74 (20.47)	*5.746	29	.000
PTR	61.49 (15.91)	67.41 (13.23)	*3.266	29	.003
G.INDEX	4.17 (.95)	4.58 (.79)	2.375	29	.024
AGR	63.52 (25.82)	59.38 (26.14)	-.785	29	.439
ART	59.65 (28.26)	56.24 (27.82)	-.548	29	.588

Note. *= *p* < .01. Standard deviations appear in parenthesis below means.

Results of the Wilcoxon signed-rank tests are presented in Table 18. The table shows that the introduction of extra planning time did not have a statistically significant impact on the results of these measures. There is some indication that extra planning served to increase the mean length of utterance from 2.82 to 3.17. However, this result did not reach statistical significance (*p* = .026).

Table 18 Wilcoxon signed-rank tests: CAF measures

Measure	Planning	Mean	SD	Median	Z	p
MNH	1	1.90	.91	1.79	-1.103	.27
	10	1.75	.90	1.48		
MLU	1	2.82	1.12	2.76	-2.232	.026
	10	3.17	1.06	3.07		
E.FREE	1	23.97	21.37	20.00	-.821	.412
	10	20.12	18.29	15.65		
M.ERR	1	1.20	.54	1.18	-.895	.371
	10	1.34	.65	1.28		
C.AS	1	1.33	.28	1.39	-1.723	.085
	10	1.41	.25	1.25		

4.2.6.1 Summary

The results of the CAF analysis reveal that increasing planning time from one minute to ten minutes increased speech rate and phonation time ratio. Overall there was general improvement in two key aspects of speech fluency after planning. Planned speech was therefore quicker speech that involved fewer and shorter instances of silent periods during task completion.

4.2.7 Discussion

To summarize the findings of Pilot 1, the results of the MFRM of the EBB scale and the analytic scale data indicate increases in scores when planning time was extended to ten minutes. This result contradicts findings in the language testing literature on pre-task planning. Many researchers working in the field have argued that planning does not have a beneficial effect on second language speech when elicited under assessment conditions (Wigglesworth 1997, Elder et al., 2002, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006, Ellis, 2005). However, the results of this pilot study indicate that extra planning time led to statistically significant increases in scores by .51 fair average on the analytic scale and .71 fair average on the EBB scale.

In terms of CAF measures, the impact of pre-task planning seems to be limited to two fluency measures: speech rate and phonation time ratio. This is a surprising result given the overwhelming amount of evidence that has been produced in the literature of impacts on CAF results after increases in planning time (see Section 2.2).

The differences between the results of the current study and those reported in the literature may be due to the adjusted alpha level adopted in this study (p = .01). The literature review demonstrates that much of the research in TBLT reports gains in CAF after extra pre-task planning time using an unadjusted alpha value of p = .05. This alpha value is applied regardless of the number of statistical tests that were run. However when conducting multiple statistical tests it is important to adjust the significance level to ensure that results are not susceptible to type one error. The absence of this adjustment is a major flaw of much of the research into pre-task planning (see Section 2.7.2.4).

The results of this pilot study raise some important questions. How can the lack of correspondence between the consistent increases on rating scale results and the relatively limited impact on CAF results be explained? One possible interpretation of this is that pre-task planning may have supported areas that were measured by the rating scales but not the CAF measures. Additional measures of CAF that closely correspond to the rating criteria were required to examine this possibility in Pilot study 2.

Beginning with the fluency measures, the results of the MFRM analytic scale data showed that raters noticed improvements in fluency when planning time was increased from one minute to ten minutes. This increase is corroborated by the increases in speech rate and phonation time ratio that were observed in the CAF results. However the remaining measures of fluency such as hesitations and pauses were not affected by the increases in planning time. The frequency and duration of pauses and hesitations in a test taker's speech are important measures of speech fluency that have consistently been shown to improve after extra pre-task planning time (see Section 2.7.2.3). It is surprising that no such difference was recorded here. In this pilot study, it was hypothesised that precise measurement of the way that planning impacts upon pausing and hesitation would be possible using Kormos' standard of 0.25 (see Section 4.2.2.7). However, pauses and hesitations lasting 0.25 seconds and above were very common in the transcripts. Raters may not have regarded a period of silence of 0.25 seconds as sufficient time to constitute a pause or hesitation. The criteria adopted in Foster and Skehan's (1996) work for identifying pauses of one second may provide a more reliable indication of the way pre-task planning affects pausing and hesitation and this possibility was investigated in Pilot study 2.

One possible explanation for the increases in complexity and accuracy scores on the analytic scale in the absence of increases in complexity and accuracy measure results is the presence of a halo effect in the rater assessments (Weir, 2005). Halo effects are observed when one category of the scale, such as fluency, influences other categories, such as complexity and accuracy. The CAF results showed that only fluency was impacted by the introduction of extra planning time. However, the analytic scale results clearly showed increases in complexity and accuracy. This finding indicates that complexity and accuracy scores may have been influenced by the raters' appraisal of speech fluency.

Regarding the tasks that were used in this study, the MFRM analyses demonstrated that Task 2 was scored consistently higher than Task 1. This result may be due to differences between the tasks in terms of the number of images involved

(Skehan, 2009). Task 2 contains more images, events and participants than Task 1, which may have permitted the test takers to demonstrate a better range of their speaking ability (see Section 2.5.1.2). Although the order of the tasks and planning conditions was counterbalanced between the test takers, it is desirable to have tasks that pose equal levels of difficulty. In the main study, tasks that were more equal (in terms of the amount of content they contain) were utilized.

4.2.8 Implications for the main study

The results of this pilot study indicated that the following features of the research design required further investigation:

Additional CAF measures based on rater analysis were required to link scale results to CAF results. This required raters to identify features of speech that were salient to their assessment of speaking ability. In the second pilot study, rater identified speech features provided the basis for the design of appropriate CAF measures. Furthermore, Skehan and Foster's (1996) standard of one second for identifying pauses and hesitations was adopted.

The current pilot study established that the ten-minute planning condition increased scores over the one-minute condition on a speaking test using two kinds of rating scales. However, practicality constraints dictate that ten minutes planning time may be unsuitable in a test environment (Elder and Wigglesworth, 2006). A further concern is that Weir et al. (2006) demonstrate increases in test scores after planning for just one minute (see Section 2.5.2). This indicates that test takers may not require ten minutes time for pre-task planning in order to increase their test scores. In the second pilot study, the amount of planning time was reduced and the impact of a five-minute planning condition and a 30-second planning condition on test scores was investigated.

Another important area of investigation is the interaction between task type and planning time. By using picture-based narrative tasks the test developer standardises the content of the speech and this impacts the degree of cognitive load that is placed on the test taker (Kormos and Denes, 2004). When completing a picture-based task, the conceptualisation stage of speech production (Levelt, 1989) receives significant scaffolding, which may permit the speaker to allocate attentional resources to the retrieval of lexis and encoding of the message. On the other hand, standardising task content through images requires the test taker to describe obligatory content for which they may not have adequate language knowledge (Skehan, 2009). This may create the impression of disfluency as the test taker attempts to negotiate the task demands. When test takers generate the content of their speech independently, they are free to make their own decisions. The test taker may therefore produce language that they are confident with and avoid structures that are beyond their linguistic resources (Skehan, 2009). From this perspective, picture-based tasks may involve more cognitive demand than tasks that do not involve image-based input (see Section 2.5.1.2).

4.3 Pilot 2

Pilot 1 demonstrated that increasing pre-task planning time from one minute to ten minutes impacted the results of a speaking test involving two picture-based narrative tasks in terms of a fluency analysis, and rater scores. However, the literature review indicates that the impact of planning on test scores may vary according to the task type that test takers complete and the amount of time spent planning (see Sections 2.5.1 and 2.5.2). This section reports the results of a second pilot study designed to assess the impact of a 30-second and a five-minute pre-task planning condition on test takers' performances on two non-picture-based description tasks. Test performance was assessed with a second, task specific EBB scale, rater-generated measures of CAF and the Iwashita et al. (2001) analytic scale.

4.3.1 Research questions for Pilot 2

1. Does the amount of planning time (30 seconds and five minutes) included in a description task impact the results of an EBB rating scale?
2. Does the amount of planning time (30 seconds and five minutes) included in a description task impact the results of an analytic rating scale?
3. Does the amount of planning time (30 seconds and five minutes) included in a description task impact complexity, accuracy and fluency (CAF) results?

4.3.2 Methodology

4.3.2.1 Participants

Test Takers. Thirty English language learner participants took part in the study. The participants had been studying in the English language preparatory programme of the university (the same university as Pilot 1) for one year. Participants were studying in summer school classes designed to prepare them for the university admission exam, which assesses the candidates' ability to follow English-medium instruction at the undergraduate level (deemed 'B2' level on the CEFR (Council of Europe, 2001) by the university administration). Ages ranged between 18 and 25 (mean = 19.6, SD = 1.99). All participants were informed that they were taking part in a research project designed to assess the impact of variation in testing format and signed letters of consent (see Appendix 2).

Raters: EBB scale 2. Seven English language instructors from the university took part in the scale creation. Five were native speakers of English and two were native speakers of Turkish. Teaching experience ranged from three to 22 years (mean = 11.2, SD = 7.53). All of the EBB rater participants regularly acted as examiners in the institutional speaking exam.

Raters: Analytic scale. Seven instructors of English were trained in the use of the analytic scale (Iwashita et al., 2001). Two were native speakers of English and the remaining five were native speakers of Turkish. The range of their teaching experience was from five to 25 years (mean = 11.6, SD = 5.94). All of the analytic scale raters regularly worked as examiners in the institutional speaking exam. While it was desirable to work with groups of raters that shared similar L1 backgrounds and teaching experience in both the EBB and the analytic scale components of this study, the study was reliant upon volunteer participants and teachers were generally reluctant to take part during a busy period of the academic year. As a result, balancing the raters between L1 backgrounds and teaching experience was not possible. The implications of this are discussed at length in Section 4.3.6.

4.3.2.2 Tasks

Two description tasks were used in the study. The tasks had been used in the university as a component of the speaking section in the university admission English test (see Section 1.1) but were subsequently retired and were being used in mock examinations at the time of this study. Test takers had no experience with these specific tasks but were familiar with the exam structure and task type. The study took place during a mock examinations week in which participants had the opportunity to experience a trial run of the university admission English test. Participants were first required to take part in a warm up session in which they answered a series of questions about themselves. This session was not recorded and did not feature in the analysis. Following the warm up, participants completed two long turn, monologue tasks. The instructions of the monologue tasks read as below:

a) Describe something interesting you have recently heard in the news.

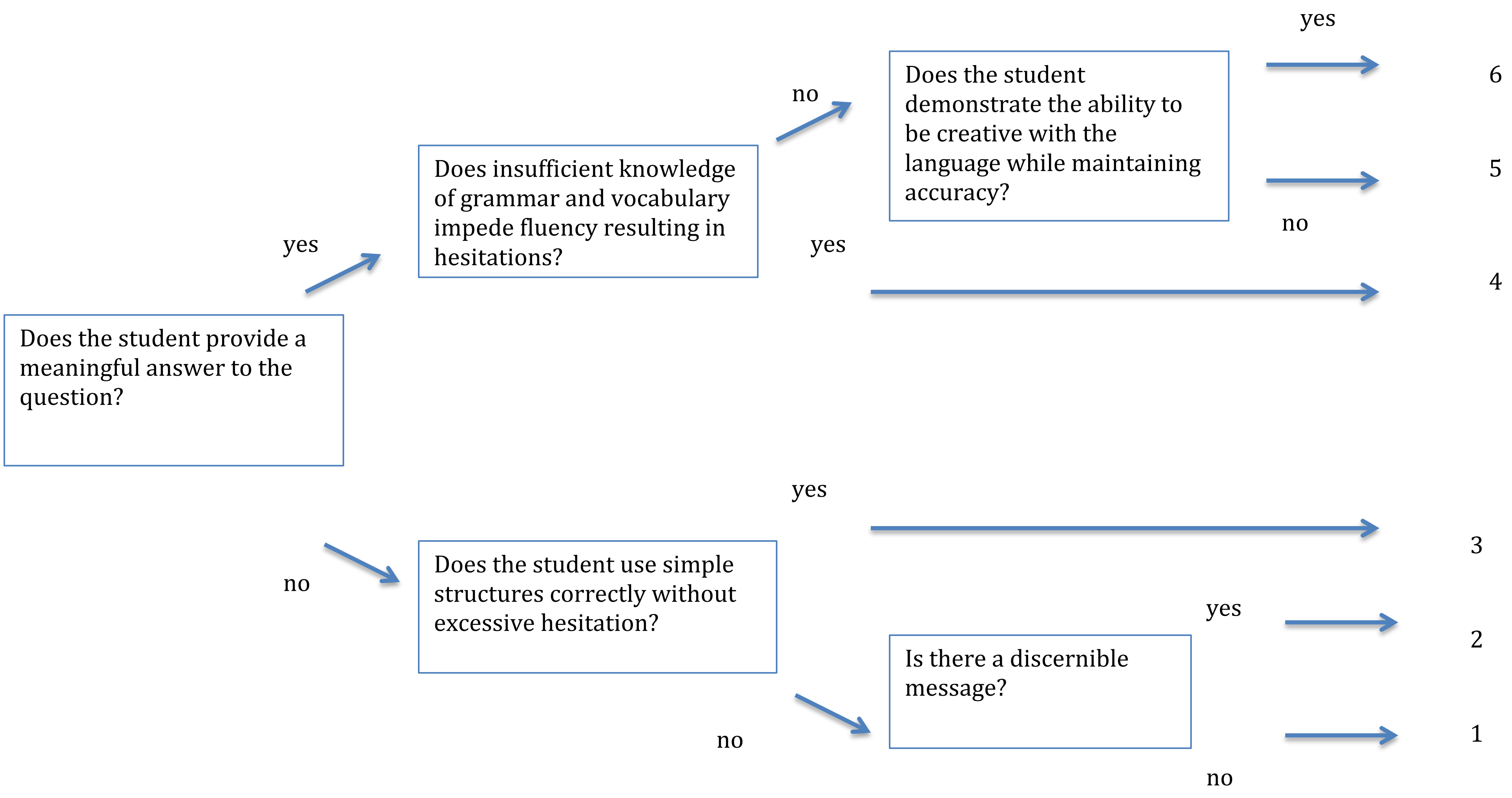
b) Describe an experience that changed your life.

The pre-task planning conditions were 30 seconds, and five minutes. The task order and planning conditions were counterbalanced between the participants. The test takers were permitted to take notes during the planning stage of the task but were informed that their notes would be removed before they began to speak. This decision was made in order to prevent test takers reading directly from their notes. Speech was recoded and transcribed according to guidelines discussed in Pilot 1 (see Section 4.2.2.3).

4.3.2.3 EBB scale procedures

The EBB scale construction process followed guidelines set out in Turner and Upshur (1996). The researcher identified a range of abilities in the database and selected samples that were representative of the low, mid and high levels of proficiency. Ten test samples were holistically selected as representative of the range of proficiency in the database. The samples were equally divided between the two tasks. However, the focus for selection was between strong and weak performances rather than the 30-second and five-minute planning conditions. The study design did not presume any difference between test samples under different planning conditions. During the scale creation process, the raters listened to all ten samples and were asked to take notes on the test-takers' language ability. Raters were instructed to be as specific as possible in their description of the performance. The notes were later collated and used to design measures of CAF. The samples were then rank ordered through paired comparison by the raters. Six levels of proficiency were identified. Raters were then asked to reach a consensus regarding the features that separated the top three test samples from the bottom three samples. This was then formulated as a yes or no question: *Does the student provide a meaningful answer to the question?* During the next stage, the samples that represented levels five and six were compared with the sample that represented level four in order to formulate a yes/no question that would separate levels five and six from level four. This question was: *Does insufficient knowledge of grammar and vocabulary impede fluency resulting in hesitations?* This process was repeated through a series of paired comparisons on the ranked samples until all of the levels were identified and the scale was completed (see Figure 10. EBB scale 2).

Figure 10. EBB scale 2



Does the student provide a meaningful answer to the question?

Does insufficient knowledge of grammar and vocabulary impede fluency resulting in hesitations?

Does the student demonstrate the ability to be creative with the language while maintaining accuracy?

Does the student use simple structures correctly without excessive hesitation?

Is there a discernible message?

4.3.2.4 Rating process: EBB scale

The samples were divided between the seven raters in order to ensure that each sample was rated twice. This process involved multiple matches between raters so that comparisons of rater severity could be made in the MFRM (see Section 4.2.2.6). An evaluation of the most efficient way to link the raters showed that every rater needed to grade 20 samples. The grades were analysed using MFRM with four facets under analysis: test-taker, rater, task and planning (see Section 4.2.2.6).

4.3.2.5 Rating process: analytic scale

The raters first took part in a short standardisation session of 30 minutes involving one test sample. A mid-level sample was selected for standardisation based upon the results of the EBB MFRM. While a more thorough standardisation session involving more samples would have been desirable, time constraints and timetable clashes meant that standardisation had to be completed within a limited timeframe. During standardisation, the raters discussed the fluency, accuracy and complexity of the sample with reference to the scale content and agreed upon a suitable grade.

Following the standardisation session, the raters were each assigned 20 samples. The raters were matched multiple times in order to run MFRM. Results were analysed using MFRM with five facets entered into the analysis: test taker, rater, task, planning and category (complexity, accuracy and fluency). Initial analysis showed that two raters had infit mean-square statistics that did not fit the model and were subsequently removed from the analysis (see Section 4.2.2.6). This may have been an unfortunate consequence of the short standardisation session that essentially detracted from the strength of the model. However, despite the removal of the two raters, there was sufficient connectivity between the facets in the model to run the MFRM (i.e. all test performances received at least one grade).

4.3.2.6 CAF measures

CAF measures were obtained through discussion with the EBB scale development group (see Section 4.3.2.3). The group was requested to make detailed notes about the features of speech they regarded as salient to their decisions about test taker proficiency. These notes were collected to identify appropriate CAF measures in the literature and to develop new CAF measures to match the raters' notes.

The grammatical accuracy measures used in this study correspond very closely to the features of speech accuracy that raters identified as salient. Raters identified errors in the test takers' production of the following forms: articles, prepositions, pronouns, modals, tenses, conditionals, word forms and verbs. These features were investigated directly by calculating the number of obligatory contexts in each transcript and establishing the percentage that were correctly supplied. In terms of speech complexity, raters identified range and depth of vocabulary, relative clauses and discourse markers as important features of the speech. For fluency, the raters identified speech speed, number and duration of pauses and hesitations and the duration of the performance (how much time the test taker actually took to complete the task). It was necessary to consult the literature in order to generate suitable approaches to measure these features of the speech. The TBLT pre-task planning literature (see Section 2.2) was explored to establish how researchers had operationalized these features of speech. The following measures were identified:

Complexity

- Lexical density assessed through Guiraud's Index (G.INDEX)
- Lexical sophistication assessed through VocabProfile (K1/K2/AWL/NONE)
- Clauses per AS-unit (C.AS)
- Use of discourse markers (*)

Accuracy

- Percentage of correctly supplied articles in obligatory contexts (ART)
- Percentage of correctly supplied prepositions in obligatory contexts (PREP)
- Percentage of correctly supplied modals (*)
- Percentage of correctly supplied pronouns in obligatory contexts (PRO)
- Percentage of correctly used tense (TENSE)
- Percentage of correctly supplied verbs in obligatory contexts (not omitted/ correct semantic usage/ including do and be and infinitive) (VERBS)
- Percentage of correctly supplied conditionals (*)
- Number of incorrect word forms per AS-unit (*)
- Percentage of errors that are self corrected (SELF)
- Mean number of errors per AS unit (ERRORS)

Fluency

- Mean number of hesitations per AS-unit (MNH)
- Phonation Time Ratio (PTR)
- Percentage of hesitations that are filled (F.HES)
- Percentage of pauses that are filled (F.PA)

- Mean Length of Utterance (MLU)
- Total Speaking Time (TST)
- Speech Rate (SPR)

(*) indicates insufficient data to run the analysis.

Pauses and hesitations were defined as a period of silence in excess of one second (Foster and Skehan, 1996). This standard was applied after accounting for the high frequency of pauses and hesitations in excess of 0.25 seconds during transcription for Pilot 1, which seemed to exaggerate the number of disfluencies in the spoken samples (see Section 4.2.7). All instances of pauses and hesitations were identified through analysis of the waveform provided in the *Audacity* program (2.0.6, 29 September 2014, http://audacity.sourceforge.net). The VocabProfile program (Cobb, n.d) [accessed 1 October 2015 from http://www.lextutor.ca/vp/] was used to calculate lexical sophistication. The results of the lexical sophistication analysis provide an indication of the percentage of the text that is made up of the first and second most common 1000 words and the academic word list (AWL). Words that belong to the second list and the AWL list are less frequent and the ability to use such vocabulary in speech is indicative of a more developed lexicon (see Section 2.7.2.1). The remaining results were calculated manually through analysis of the transcripts.

Four of the features identified by the raters were not used in the analysis. These were number of discourse markers, the percentage of correctly supplied modal verbs, the percentage of correctly supplied conditional clauses, and the number of incorrect word forms per AS unit. Evaluation of the transcripts demonstrated that many participants did not produce modal verbs, discourse markers or conditional clauses. In addition there was no use of incorrect word forms in the majority of the transcripts. As a result these measures were not used in the CAF analysis.

Two additional measures that were not identified by the raters were selected for the analysis. A measure of global accuracy was required to identify learner errors that were not acknowledged by the EBB raters. This measure was mean number of errors per AS unit (ERRORS). The second measure was the number of idea units produced in each sample (IDEAS). Idea units are defined as 'short phrases and clauses connected with *and*, *or*, *but* or *that*, or not joined by conjunctions at all but simply spoken next to each other, with possibly a short pause between them' (Luoma, 2004, p. 12). Ellis and Barkhuizen (2005, p. 154) write that the number of idea units is an indication of the level of 'propositional completeness' involved in a text. The number of idea units may have some bearing on the EBB criteria: '*Does the student provide a meaningful answer to the question?*' and '*Is there a discernable message?*'. In order for a test taker to produce a meaningful answer to the question, a minimum number of ideas must be communicated. Luoma (2004, p. 12) writes that idea units may not contain a verb and are commonly:

- 'about two seconds or about seven words long'
- 'spoken with a coherent information contour'
- 'often limited on both sides by pauses or hesitation markers'

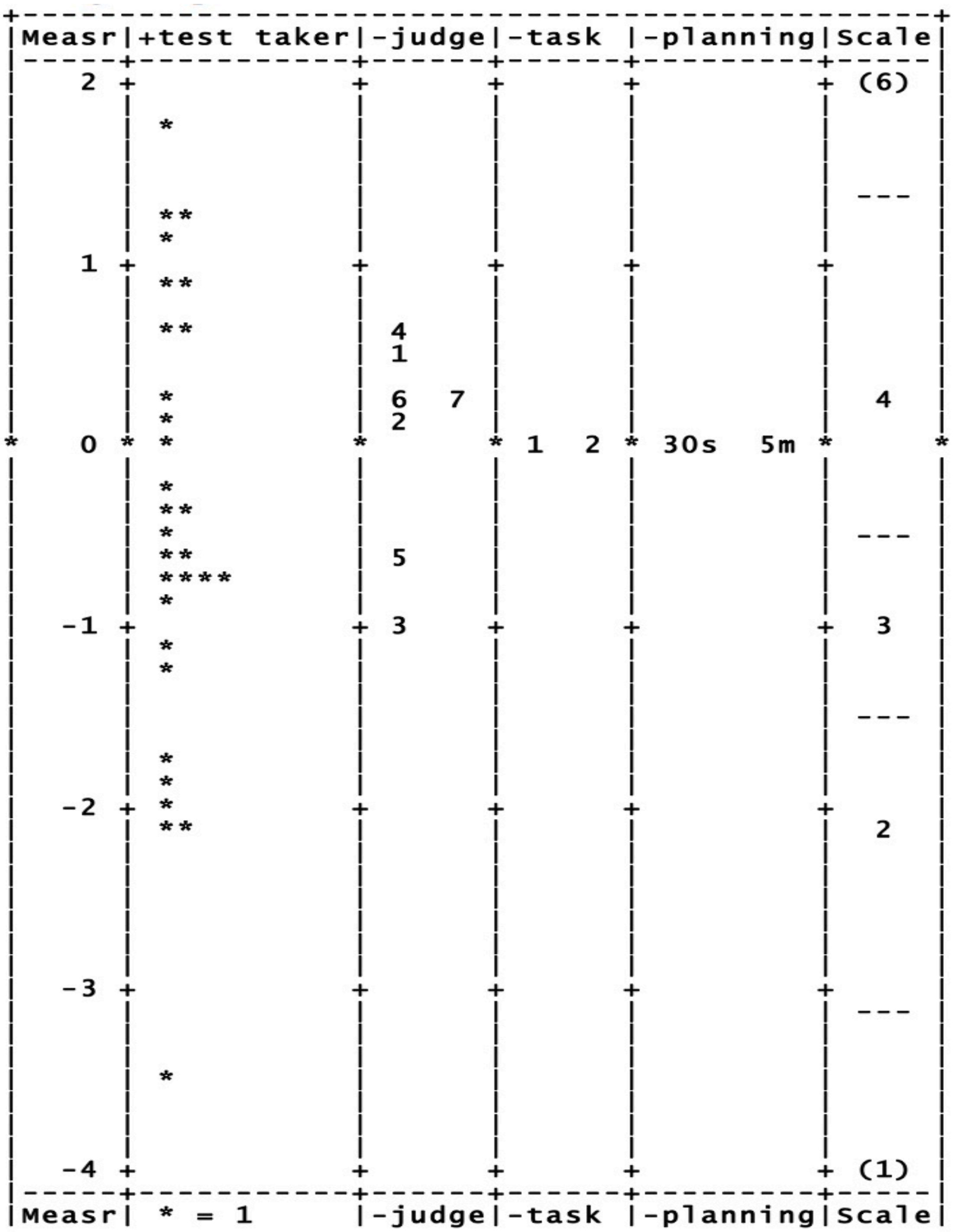
In addition, Frost, Elder and Wigglesworth (2011, p. 356) write that idea units may be composed of:

- ‘coordinated verb phrases (*we could improve bus service, or build better subway*)’
- ‘coordinated nouns and noun phrases connected to a common verb phrase (*it was less polluted air, quiet place, more tourism*)’
- ‘coordinated independent adjectives connected to a common verb phrase (*the city would be less noisy, less polluted*)’

4.3.3 Results: EBB scale 2

Figure 11 presents the output of the MFRM (EBB scale) in the form of a Wright map. The Wright map contains six columns that contain the logit scale, the spread of test taker abilities, the spread of rater (judge) severity, task difficulty, planning condition difficulty, and the EBB scale.

Figure 11. Wright map EBB scale 2



```
+-----------------------------------------------+
|Measr|+test taker|-judge|-task |-planning|Scale|
|-----+-----------+------+------+---------+-----|
|   2 +           +      +      +         + (6) |
|     |           |      |      |         |     |
|     | *         |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         | --- |
|     | **        |      |      |         |     |
|     | *         |      |      |         |     |
|   1 +           +      +      +         +     |
|     | **        |      |      |         |     |
|     |           |      |      |         |     |
|     | **        | 4    |      |         |     |
|     |           | 1    |      |         |     |
|     |           |      |      |         |     |
|     | *         | 6  7 |      |         |  4  |
|     | *         | 2    |      |         |     |
*   0 * *         *      * 1  2 * 30s  5m *     *
|     |           |      |      |         |     |
|     | *         |      |      |         |     |
|     | **        |      |      |         |     |
|     | *         |      |      |         | --- |
|     | **        | 5    |      |         |     |
|     | ****      |      |      |         |     |
|     | *         |      |      |         |     |
|  -1 +           + 3    +      +         +  3  |
|     | *         |      |      |         |     |
|     | *         |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         | --- |
|     |           |      |      |         |     |
|     | *         |      |      |         |     |
|     | *         |      |      |         |     |
|  -2 + *         +      +      +         +     |
|     | **        |      |      |         |  2  |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|  -3 +           +      +      +         +     |
|     |           |      |      |         | --- |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|     | *         |      |      |         |     |
|     |           |      |      |         |     |
|     |           |      |      |         |     |
|  -4 +           +      +      +         + (1) |
|-----+-----------+------+------+---------+-----|
|Measr| * = 1     |-judge|-task |-planning|Scale|
+-----------------------------------------------+
```

4.3.3.1 Facets in the MFRM model

Test Takers. The higher a test taker is placed on the map, the greater the score. The majority of the test-takers appear in the upper half of the map indicating that test taker ability was generally high. Test taker ability measures ranged from -3.44 to 1.72 on the logit scale. These values represent a range of ability from 1.39 to 4.66 logits on the fair average scale. The mean of standard errors was .58 in logits. This value is rather high, which indicates imprecision in the measures (see Section 4.2.2.6). The separation index was 1.76 and the test takers were separated into 2.68 strata. These values represent a rather limited range of ability constituting a high proficiency group, a mid-level proficiency group, and a low proficiency group. Reliability of the separation of test takers was .76 indicating that the separation was satisfactory, though perhaps due to the high standard error value, not as reliable as observed in Pilot 1. Fit statistics were acceptable with the exception of test takers 6 and 16. Test takers 6 and 16 record values above 2.00, at 2.60 and 2.03 respectively. Information relating to these test takers' raw scores and rater severity measures is presented in Table 19.

Table 19 Scores awarded to test taker 6 and 16 by seven raters

Test	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
taker	.51	.09	-.99	.64	-.67	.21	.22
6	1	1		4	1	1/2	2
16		2	4	4	4	1	4

Severity measure values appear below rater ID.

Test taker 6 received the lowest scores possible from a lenient rater (Rater 5) and a relatively high score from the most severe rater (Rater 4). Test taker 16 received predictably high grades from lenient raters. However, the most severe rater uncharacteristically awarded the same grade. These scores did not fit the predictions made in the model. Furthermore, the raters with moderate severity values (Raters 6 and 7) disagreed considerably on this test taker's ability.

The CAF results demonstrate that test takers 6 and 16 recorded below average values for all measures with the exception of subject verb agreement, and total speaking time (test taker 6), and tense, percentage of filled hesitations, and speech rate (test taker 16). The cause of the inconsistency in scores may have been due to variation between the raters' appraisal of these features. For example, some raters may have regarded correct use of tense as a central component of a 'meaningful answer' (see EBB scale 2) and therefore assigned high grades to test taker 16. On the basis of this interpretation and in the interest of having the same number of test samples in the MFRM of the EBB scale data and the analytic scale data, the misfitting test takers were retained.

Raters. The Wright map ranks raters (the third column) by placing lenient raters toward the bottom of the map. Inspection of the map demonstrates that the raters displayed different levels of severity. Table 20 reports the rater statistics. The range of rater severity measures was from -.99 to .64. At its most extreme, the difference between rater fair average values was 1.20 logits. Rater separation was 1.81 and the strata value was 2.74. The reliability index was .77, which indicates that the variation between the raters' levels of severity was reliable. Rater fit statistics ranged from .73 to 1.35, showing that the raters were internally consistent in their distribution of scores (see Section 4.2.2.6). However, Rater 4 demonstrated a tendency toward model misfit (1.35) indicating a slight degree of inconsistency, though not enough to warrant removal.

Table 20 Report of rater severity: EBB scale 2

Rater	Fair Average	Severity Estimate	Error	Infit mean-square index
3	4.12	-.99	.29	.87
5	3.96	-.67	.28	1.25
2	3.42	.09	.25	.73
6	3.33	.21	.27	.85
7	3.32	.22	.27	1.04
1	3.05	.51	.27	.83
4	2.92	.64	.27	1.35
Mean	3.45	.00	.27	.99
SD	.41	.56	.01	.22
Reliability of difference in severity of raters .77				

Tasks. The fourth column compares the two tasks. According to the Wright map, performances on the tasks received very similar grades. MFRM statistical results of the task data are shown in Table 21. According to the table, Task 2 was scored slightly higher than Task 1 by .09 logits on the fair average scale. However, the fixed chi-square test demonstrates that this result was not statistically significant. This analysis indicates that the tasks posed similar levels of difficulty to the test takers.

Table 21 Tasks 1 and 2 fair average, measure and infit statistics: EBB scale 2

Task	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
1	3.46	0.05	0.90	$\chi^2 = .3, p = .60$
2	3.55	-0.05	1.07	

Planning. The fifth column compares the planning conditions. The map shows that test taker scores did not vary considerably between the conditions. Table 22 presents the MFRM planning condition statistics. The fair average statistics demonstrate that scores were slightly higher under the five-minute planning condition than the 30-second planning condition. However, the difference was small (.05 logits on the fair average scale) and did not reach statistical significance. This difference in scores does not indicate a major impact of pre-task planning on the EBB scale scores. Both the longer and shorter planning fair average values were mid-level 3 on the EBB scale.

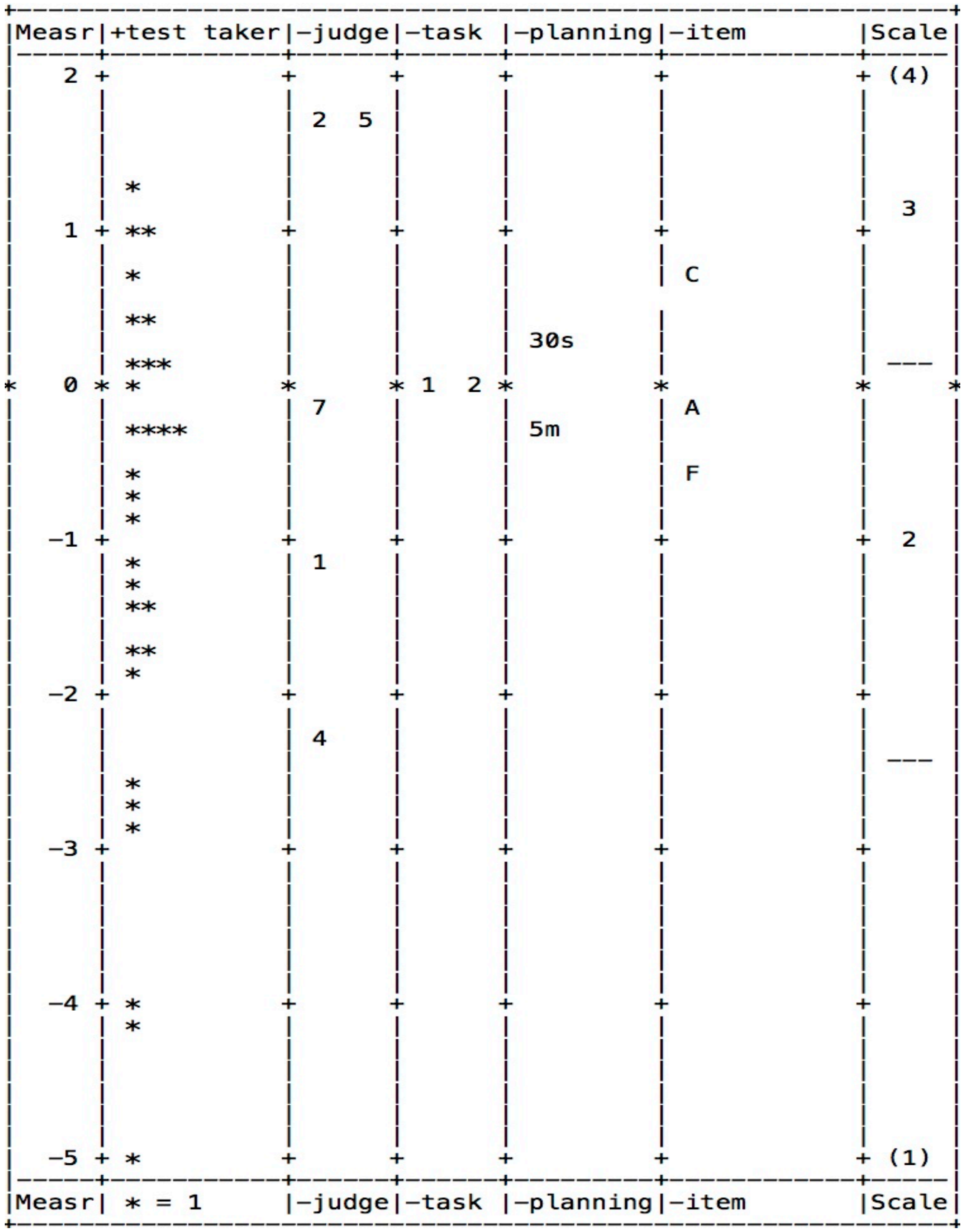
Table 22 Planning fair average, measure and infit statistics: EBB scale 2

Planning	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
5 minutes	3.53	- .03	.98	$\chi^2 = .1, p = .80$
30 seconds	3.48	.03	.98	

4.3.4 Results: analytic scale

Figure 12 presents the output of the MFRM in the form of a Wright map. The map summarizes rating results and indicates differences in test taker ability, rater severity, task difficulty and planning time. The item column shows differences in difficulty between the fluency (F), accuracy (A) and complexity (C) categories. The following section reports the results of the MFRM in detail.

Figure 12. Wright map analytic scale



```
+-----------------------------------------------------------+
|Measr|+test taker|-judge|-task |-planning|-item      |Scale|
|-----+-----------+------+------+---------+-----------+-----|
|   2 +           +      +      +         +           + (4) |
|     |           |      |      |         |           |     |
|     |           | 2  5 |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     | *         |      |      |         |           |     |
|     |           |      |      |         |           |  3  |
|   1 + **        +      +      +         +           +     |
|     |           |      |      |         |           |     |
|     | *         |      |      |         | C         |     |
|     |           |      |      |         |           |     |
|     | **        |      |      |         |           |     |
|     |           |      |      | 30s     |           |     |
|     | ***       |      |      |         |           | --- |
*   0 * *         *      * 1  2 *         *           *     *
|     |           | 7    |      |         | A         |     |
|     | ****      |      |      | 5m      |           |     |
|     |           |      |      |         |           |     |
|     | *         |      |      |         | F         |     |
|     | *         |      |      |         |           |     |
|     | *         |      |      |         |           |     |
|  -1 +           +      +      +         +           +  2  |
|     | *         | 1    |      |         |           |     |
|     | *         |      |      |         |           |     |
|     | **        |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     | **        |      |      |         |           |     |
|     | *         |      |      |         |           |     |
|  -2 +           +      +      +         +           +     |
|     |           |      |      |         |           |     |
|     |           | 4    |      |         |           |     |
|     |           |      |      |         |           | --- |
|     | *         |      |      |         |           |     |
|     | *         |      |      |         |           |     |
|     | *         |      |      |         |           |     |
|  -3 +           +      +      +         +           +     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|  -4 + *         +      +      +         +           +     |
|     | *         |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|     |           |      |      |         |           |     |
|  -5 + *         +      +      +         +           + (1) |
|-----+-----------+------+------+---------+-----------+-----|
|Measr| * = 1     |-judge|-task |-planning|-item      |Scale|
+-----------------------------------------------------------+
```

4.3.4.1 Facets in the MFRM model

Test Takers. The majority of the test takers cluster between -2 and 1 logits on the Wright map. This corresponds to levels 2 and 3 on the analytic scale. The majority of test takers therefore attained scores toward the mid-lower end of the scale. Test taker ability measures ranged from -6.66 to 1.24 on the logit scale indicating that there was a range of proficiency in the sample. The range of fair average values was from 1.01 to 3.03 logits. The mean standard error was .62. This value is high and suggests that the results may lack precision (see Section 4.2.2.6). High standard error values typically occur when the MFRM is calculated using a low number of observations, e.g. because the size of the population is small or test samples receive a low number of ratings (Winke et al., 2012). The mean standard error value can be attributed to the removal of the misfitting raters, which decreased the number of observations made in the model. However, this removal was necessary because misfitting raters have a distorting effect on MFRM that generates misleading results. High standard error values are the cost of data fit to the MFRM model. The separation index was 2.37 and test takers were separated into 3.50 strata. The reliability of the separation was .85, demonstrating that the separation of the test takers into different levels of proficiency was reliable. Test taker fit statistics were within an acceptable range of .7 to 1.3 except for test takers 11 (3.01) and 9 (2.11). The value for test taker 11 was particularly extreme and required further investigation. Table 23 presents test taker 11 and 9 raw scores and the rater severity measures.

Table 23 Scores awarded to test taker 11 and 9 by five raters

Test taker	Rater 1	Rater 2	Rater 7	Rater 4	Rater 5
	-1.11	1.75	-.07	-2.27	1.69
11	1/1/1	3/3/2	2/2/2		2/3/3
9				2/2/1 1/1/1	

Severity measures appear below rater ID

Test taker 11 received the lowest possible grade from Rater 1 and relatively high grades from the most severe raters (2 and 7). Test taker 9 received uncharacteristically low grades from a relatively lenient rater (Rater 4). This behaviour did not match the majority of observations in the model and caused misfit. Analysis of CAF results (see Section 4.3.5) demonstrated that test taker 11 made fewer errors per AS unit than the population average and showed higher than average levels of accuracy in tense, and prepositions. Accuracy of articles and pronouns were also higher than average under the five-minute planning condition. The misfit may be due to the extent to which levels of speech accuracy were central to raters' decisions about test taker proficiency. In a similar way, test taker 9 registered values that were below average in mean length of utterance, accuracy of articles and prepositions and self-correction. Rater 4 may have regarded these features of speech as particularly important, which may explain the low grades.

Test taker misfit would normally not cause such a large problem for MFRM but the earlier removal of two raters (3 and 6) weakened the power of the analysis and increased the sensitivity of the fit measures. In order to maintain the potential for comparison between the EBB and analytic scale results, the misfitting test takers were retained.

Raters. The Wright map shows that levels of rater severity varied considerably. Table 24 presents the MFRM rater statistics. Differences between raters ranged from 1.46 to 3.11 logits on the fair average scale. This represents a substantial difference between the most lenient and the most severe raters. Clearly, the rater represents an important variable in the results of this examination. This is further evidenced by the reliability statistics. The reliability index was .98. Reliability measures that are close to 1 indicate that there is a great deal of variety in rater severity and that this has impacted upon the results (Winke et al., 2012). The separation index was 6.81 and the strata value was 9.41. Infit statistics reveal that the raters were within an acceptable range of .7 to 1.3. The raters can therefore be said to fit the model. They were not erratic or overly predictable in their assessment of the speech but clearly differed in severity.

Table 24 Report of rater severity: analytic scale

Rater	Fair Average	Severity Estimate	Error	Infit mean-square index
4	3.11	-2.27	.22	.95
1	2.56	-1.11	.20	.95
7	2.11	-.07	.22	.74
5	1.48	1.69	.24	1.08
2	1.46	1.75	.25	1.04
Mean	2.14	.00	.23	.98
SD	.64	1.57	.02	.13
Reliability of difference in severity of raters .98				

Tasks. The tasks elicited performances that received very similar grades. Table 25 presents the MFRM statistics of the analysis of Tasks 1 (*Describe something interesting you have recently heard in the news*) and 2 (*Describe an experience that changed your life*). Task 1 (fair average 2.10) was scored slightly higher than Task 2 (fair average 2.06). This is a reversal of the EBB results in which Task 2 recorded scores that were higher than Task 1 by .09 fair average. However, the result of the fixed chi-square demonstrates that the difference was not statistically significant.

Table 25 Tasks 1 and 2 fair average, measure and infit statistics: analytic scale

Task	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
1	2.10	-0.05	0.90	$\chi^2 = .2, p = .65$
2	2.06	0.05	1.07	

Planning. The Wright map indicates that there was an increase in scores between the 30-second and five-minute planning conditions. Table 26 presents the MFRM planning statistics. The five-minute planning condition recorded a value that was .19 logits higher than the 30-second planning condition value on the fair average scale. The result of the chi-square test showed that this result was statistically significant at p = .01. However, the increase in scores that came from extra planning time was minimal and did not cause the fair average value to increase substantially.

Table 26 Planning fair average, measure and infit statistics: analytic scale

Planning	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
5 minutes	2.18	- .25	.87	$\chi^2 = 6.0, p = .01$
30 seconds	1.99	.25	1.10	

Complexity, Accuracy, Fluency. In the Wright map, the categories on the analytic scale are ordered in terms of difficulty with the more difficult categories toward the top of the map. The order of the categories indicates that complexity was the most difficult category on the scale and that the fluency scores were highest. The frequency of scores on each category is presented in Table 27. There were no instances of level 5 being awarded to any test taker and the majority of the scores tended to occur at the low end of the scale.

Table 27 Overall score frequency on the categories of the analytic scale

Score	Accuracy	Complexity	Fluency
1	26	42	42
2	36	31	30
3	25	17	18
4	10	7	7
5	0	0	0

Separate MFRM analyses were run on each category of the analytic rating scale to identify the component of the scale (complexity, accuracy, fluency) that was impacted most by the introduction of extra planning time. Table 28 presents the results. The table indicates that the five-minute condition resulted in higher fair average measures than the 30-second condition on the complexity (.26), accuracy (.22) and fluency (.28) categories. However, the chi-square tests revealed that these increases did not reach statistical significance.

Table 28 Complexity, accuracy, fluency fair average, measure and infit statistics

Category	Planning	Fair Average	Measure	Infit mean-square index	Fixed (all same) chi-square
Fluency	5 min	2.05	- .35	.86	$\chi^2 = 3.5, p = .06$
	30 sec	1.77	.35	1.14	
Accuracy	5 min	2.30	-.33	.70	$\chi^2 = 2.9, p = .09$
	30 sec	2.08	.33	1.28	
Complexity	5 min	2.02	-.34	.86	$\chi^2 = 3.3, p = .07$
	30 sec	1.76	.34	1.15	

4.3.5 CAF results

Shapiro-Wilks tests (*p* = .05) and an evaluation of their histograms indicated that the scores of the following variables were normally distributed: Guiraud's Index, speech rate, the percentage of pauses that are filled, total speaking time, mean number of hesitations, and mean number of errors per AS unit. The results of these measures were therefore analysed with paired samples t-tests using a Bonferroni adjusted alpha level of *p* = .008 (.05/6). The remaining variables were found to be non-normally distributed: lexical sophistication, articles, prepositions, verbs in obligatory contexts, pronouns, self-correction, tense, clauses per AS unit, phonation time ratio, mean length of utterance, percentage of hesitations that are filled, and number of idea units. These measures warranted a non-parametric test to judge the impact of planning. Wilcoxon signed-rank tests were completed on the non-parametric data using a Bonferroni adjusted alpha level of *p* = .004 (.05/12).

Table 29 reports the results of the paired samples t-tests. The only result that reached statistical significance was total speaking time. The mean scores for total speaking time were 84.50 seconds under the 30-second planning condition and 98.47 seconds under the five-minute condition (*t*(29)=-2.931, *p* = .003). These results indicate that the extra planning time impacted significantly on the amount of time test takers took to complete the task. Completing extra pre-task planning increased task time by a period of approximately 14 seconds.

Table 29 Results of the paired samples t-tests

	Planning			
	30 seconds	5 minutes	T	df
SPR	71.03 (17.67)	71.28 (16.83)	-.122	29
TST	84.50 (22.70)	98.47 (26.50)	*-2.931	29
F.PA	63.90 (19.90)	62.10 (19.90)	.567	29
MNH	2.54 (1.26)	2.79 (1.25)	-.912	29
G.INDEX	4.31 (.74)	4.45 (.67)	.968	29
ERRORS	1.32 (.59)	1.35 (.54)	.264	29

Note. *= *p* < .008. Standard deviations appear in parenthesis below means.

Table 30 presents the results of the Wilcoxon signed-ranks tests. Results showed that planning led to statistically significant increases in phonation time ratio (z = 3.518, p < .001). The median value under the 30-second planning condition was 62.25 and under the five-minute condition was 73.00. This result indicates that planned speech involved fewer empty pauses and hesitations (i.e. the test takers spent more of the task time producing speech). The remaining results did not reach statistical significance.

Table 30 Results of the Wilcoxon signed-rank tests

Measure	Planning	Mean	SD	Median	Z	p
IDEA	30 sec	8.00	2.59	7.00	-1.963	.050
	5 min	9.45	3.68	9.00		
PTR	30 sec	61.88	20.15	62.25	*-3.518	.000
	5 min	77.92	27.10	73.00		
F.HE	30 sec	64.94	19.98	63.97	-2.54	.011
	5 min	71.73	20.44	71.66		
MLU	30 sec	2.83	1.04	2.57	-0.041	.967
	5 min	2.82	.88	2.83		
C.AS	30 sec	1.36	.24	1.32	-1.061	.309
	5 min	1.31	.21	1.30		
K1	30 sec	68.66	9.18	67.40	-.278	.781
	5 min	67.98	9.97	69.00		
K2	30 sec	3.69	3.22	3.35	-.011	.991
	5 min	3.45	2.05	2.96		
AWL	30 sec	.82	1.33	0.00	-.784	.433
	5 min	1.14	1.84	0.00		
NONE	30 sec	26.80	9.40	26.6	-.267	.789
	5 min	27.40	8.50	26.9		
ART	30 sec	47.22	35.09	50.00	-.457	.648
	5 min	46.00	29.67	43.65		
PREP	30 sec	67.70	30.06	75.00	-1.387	.166
	5 min	74.06	28.69	80.00		
PRO	30 sec	92.63	14.84	100.00	-.784	.433
	5 min	90.49	12.84	100.00		
SELF	30 sec	13.48	16.93	.00	-1.531	.126
	5 min	18.97	20.52	15.50		
TENSE	30 sec	82.46	15.03	83.30	-1.486	.137
	5 min	86.87	14.04	88.75		
VERBS	30 sec	86.52	14.24	88.90	-.923	.356
	5 min	85.58	8.59	85.15		

* *p* =.004

4.3.6 Discussion

The following discussion summarizes the aims, methodology and results of Pilot 2. The study assessed the impact of varying pre-task planning time (from one minute/ ten minutes to 30 seconds/ five minutes) and task type (from picture based-narrative tasks to non-picture-based description tasks) in a test of second language speaking ability. The results can be broadly summarized as such; the five-minute planning condition had no statistically significant impact on the results of MFRM involving a task specific EBB scale. However, the extra planning time did impact the results of an analytic scale comprising descriptors of complexity, accuracy and fluency in relation to the 30-second planning condition, although the increase was minor at .19 in fair average. Rater generated measures of complexity, accuracy and fluency (CAF) recorded gains in measures of fluency: total speaking time and phonation time ratio. To extrapolate from these results then, it might be tentatively suggested that under the five-minute planning condition, test takers spoke for a longer period of time and their speech involved less silent pauses and hesitations, which can be linked to minor increases in scores on the analytic scale.

Table 31 compares the results of the EBB scale 2 and the analytic scale. Whereas the results of the analytic scale demonstrate statistically significant increases after extra pre-task planning time, the results of the EBB scale show no statistical difference between the planning conditions. This result may be due to the greater number of levels on the analytic scale, which may foster finer distinction between the test takers: this interpretation is supported by the test taker strata, which indicate that raters were more likely to notice differences between the spoken samples when using the analytic scale (the number of strata was 3.50) than the EBB scale (the number of strata was 2.68). However, the extra levels of the analytic scale also seem to have generated more inconsistency between the raters. Whereas raters were separated into 2.74 strata on the EBB scale, the analytic scale generated 9.41 separate rater strata. Comparing the rater separation, strata and point biserial correlation index statistics, it becomes clear that the EBB scale generated substantially more agreement between raters. Differences in rater severity were much more likely to impact the results of the analytic scale than the EBB scale.

Table 31 Comparison of the EBB and analytic scale results

	EBB Scale	Analytic Scale
Candidate Discrimination (strata)	2.68	3.50
Rater Separation (difference in logits between harshest and most lenient rater)	.65	4.02
Rater Strata	2.74	9.41
Rater Reliability (Point Biserial Correlation Index)	.72	.63
Mean Infit Values (SD)	.99 (.22)	.98 (.13)
Planning	.05 (*p* = .80)	.19 (*p* = .01)

The levels of variation within the analytic scale results may be due to the fact that raters could only participate in a very limited standardisation session in which to familiarize themselves with the scale content and the standardisation sample. A longer standardisation session may have allowed raters to tune their levels of severity so that they were more in line with the group. This would have led to better strata, reliability and fit statistics. A second possibility is that rater demographics may constitute a variable in the levels of severity observed. Winke et al. (2012) found that native speakers and non-native speakers of the language being tested tend to vary in their levels of severity when rating performance. Given that the balance between native speaker and non-native speakers was uneven between the rating groups (see Section 4.3.2.1), the possibility that the rater demographic impacted the results is feasible.

In comparing the results with those reported in Pilot 1, it is important to note that the tasks differed between the two studies. The tasks in Pilot 2 required different cognitive operations (e.g. generating content) from test takers than those employed in Pilot 1 (e.g. discussing obligatory content). However, the difference between the picture-based and non-picture-based task types cannot be sufficiently evaluated based on the results of this study as the planning variable was also altered (from 1 minute and ten minutes to 30 seconds and five minutes). The comparison between the two task types and four planning conditions was made in the main study by including task type and planning time as within subject variables in the research design.

4.3.7 Implications for the main study

To summarize the results, increasing planning time from 30 seconds to five minutes did not impact the results of a speaking test consisting of two description tasks to the same extent that was observed in the results of Pilot 1. These results indicated that the following features of the research design required further investigation.

- Task and time: Test takers appear to respond less to extra planning time when completing non-picture-based description tasks in relation to picture-based narrative tasks. However, attempts to measure the impact of the task variable are confounded by the variation in planning time between Pilots 1 and 2. The main study therefore investigated the impact of the four planning conditions across the two task types.
- Standardisation: Short standardisation sessions may lead to substantial disagreement between raters. Standardisation therefore consisted of a minimum of three samples in the main study.

4.4 Summary of the pilot studies

The pilot studies investigated the effects of providing various amounts of pre-task planning time to test takers as they completed picture-based narrative tasks and non-picture-based description tasks. The results revealed interesting interactions between planning time and task type. The results of Pilot 1 showed statistically significant increases in test scores under the ten-minute planning condition in relation to the one-minute condition when test takers completed picture-based narrative tasks. In contrast, in Pilot 2 the impact of extra planning time on analytic scale scores was minor and there was no statistically significant impact on EBB scores. However, the impact of pre-task planning on speech fluency was relatively consistent between the pilot studies, phonation-time ratio increased after extra planning time in both studies, whereas increases in speech rate were observed in Pilot 1 and increases in total speaking time were observed in Pilot 2. Given the levels of variation between the pilot study results, the main study investigated the following test facets:

- Planning time: Pilot 1 compared scores that were awarded under one-minute and ten-minute planning conditions. Pilot 2 reduced the planning time to 30 seconds and five minutes. The amount of planning time available to test takers may be an important variable in the results of speech assessment. Statistical procedures that permit comparison of CAF results and rating scale results between the four different planning conditions were required to investigate this variable.
- Task type: The tasks used in Pilot 1 were picture-based and required the test taker to deliver a short narrative. In Pilot 2, test takers were required to

develop the content of their speech more independently without the support of the images. Statistical procedures that permit comparison of CAF results and rating scale results between the different task types were required.

- Proficiency: Independent measures of proficiency were not acquired for test taker participants in the pilot studies. Language ability may be an important variable that determines the impact of planning on test scores (see Section 2.6.2). The main study investigated the interaction between second language proficiency and pre-task planning time.
- A method for comparing the interaction between planning time, task type and second language proficiency was required to examine the CAF results and test scores.

5 Methodology

5.1 Introduction

The purpose of the main study was to identify the test conditions that resulted in the largest impact of the pre-task planning variable on a test of second language speaking proficiency with Turkish learners of L2 English. The literature review and results of two pilot studies demonstrate that the effect of including pre-task planning time in a test of second language speaking ability varies according to a series of factors: the amount of planning time provided, the task type, the levels of the test takers' language proficiency and the approach to measurement. This chapter describes the research design and method of analysis adopted to investigate the interaction between these factors.

5.2 Research questions

1. Does variation in planning time operationalized as 30 seconds, one minute, five minutes and ten minutes impact the results of a language test when assessed with

    a) an EBB scale

    b) an analytic scale

    c) measures of complexity, accuracy, and fluency (CAF)?

If the answer to research question 1 is affirmative,

1.1 Which amount of planning time (30 seconds, one minute, five minutes, ten minutes) most substantially impacts test scores and CAF results?

1.2 Does the impact of the four planning conditions on test scores vary between the analytic scale and the EBB scale?

1.3 Does the impact of the four planning conditions on test scores and CAF results vary between groups of test takers who have different levels of language proficiency?

2. Does the impact of the four planning conditions on test scores and CAF results vary between picture-based narrative tasks and non-picture-based description tasks?

If the answer to research question 2 is affirmative,

2.1 Which task type and planning condition has the largest impact on test scores and CAF results?

5.3 Participants

Two groups of participants took part in the study. The first group was the test taker group. This group was compiled of students enrolled in the English preparatory school (see Section 4.3.2.1) who were studying to reach a level of English proficiency deemed suitable to begin English-medium undergraduate studies (deemed 'B2' level on the CEFR, Council of Europe, 2001). The second group was the English instructor group. The instructors were responsible for the delivery of course content and assessment in the English preparatory school. The test takers and instructors were recruited on a voluntary basis and did not participate in the pilot studies. At the time of the study, there were over 1,500 students enrolled in the English preparatory school and 140 teachers. The aim was to recruit participants that were representative of the wider population of students and instructors. However, the study relied upon volunteers with relatively high levels of motivation to participate, which may reduce the generalizability of the findings (see Section 8.4).

5.3.1 Test takers

Test takers were recruited from the school of foreign languages in a university in Turkey. The total number of test taker participants in the study was 47. Ages ranged from 18 to 22 (mean = 18.9, SD = 1.01) and there was an even divide between male and female participants. At the time of the study, all participants were studying in the English preparatory programme in the university in which Pilots 1 and 2 were conducted and were receiving six hours daily English tuition. The participants' level of exposure to English was restricted to the educational environment and any English they encountered through popular culture. Test takers did not have the regular opportunity to use English outside the language classroom. The study took place during a mock examinations week, in which test takers were able to experience an informal, trial run of the university admission English test. Test takers were informed that they would complete four tasks, which were similar to those they would encounter in the admission test (see Section 5.4.2). All test takers signed an approved consent form informing them about the purpose of the study (see Appendix 2).

5.3.2 Raters: EBB scale 3

Seven English language instructors took part in the scale construction and grading. All participants worked in the university's English preparatory programme. Five were native speakers of English and two were native speakers of Turkish.

Teaching experience ranged from one to 20 years (mean = 10, SD = 5.9). All participants regularly acted as examiners in institutional speaking assessment for both formative and summative purposes. The raters were recruited to participate in the study based on their availability and willingness to participate.

5.3.3 Raters: analytic scale

Ten rater participants were involved in the analytic scale grading. All participants were English instructors in the university's English preparatory programme and regularly acted as examiners during institutional speaking assessment. Teaching experience ranged from five to 15 years (mean = 10.1, SD = 3.1). Five of the raters were native speakers of English and five were native speakers of Turkish. With the exception of Rater 1 (R1) who awarded scores with both the EBB scale and the analytic scale to provide connectivity in the MFRM matrix (see Section 4.2.2.6), the EBB raters and analytic raters were entirely different.

5.4 Procedures

5.4.1 Oxford Quick Placement Test

In order to obtain an independent measure of language proficiency, the Oxford quick placement test (QPT; UCLES, 2001) was administered to the test takers before the speaking test. The QPT is a multiple choice, paper-based test comprising a series of labelling and cloze activities that test knowledge of lexis and grammatical forms. The test was deemed appropriate as an independent measure of language proficiency as scores are reported in terms of the Common European Framework reference levels;

a common way of reporting language proficiency (Council of Europe, 2001). Seven participants did not arrive in time to sit the QPT and went directly to the speaking test. The total number of completed QPT tests was 40. The QPT results demonstrate that the average score in the sample was 26.9 (A2) and the standard deviation was 4.28. The range of scores was from 17 (A1) to 36 (B1). A break down of scores is presented in Table 32.

Table 32 QPT results

Points	CEFR Level	Number of Students
0-17	A1	3
18-29	A2	25
30-39	B1	12

5.4.2 Speaking tasks and planning conditions

Four planning conditions were investigated in the research. These planning conditions were 30 seconds, one minute, five minutes and ten minutes. During the planning stage, participants were provided with a pen and paper and instructed that they could take notes on the task but that their notes would be removed before beginning the task. This measure was taken to prevent students reading directly from their notes, which would provide an unrealistic impression of spoken ability.

The speaking test consisted of four speaking tasks: two non-picture-based description tasks that had been used in the university admission test but had since been retired and two picture-based narrative descriptive tasks taken from Inoue (2013). The tasks were selected because they did not presume any background knowledge of academic language and involved topics that are common in everyday interaction. The tasks elicit a range of language structures, and test the ability to organise a narrative, describe a scene and speculate about character motivations. These task features are important for the university admission test as the ability to successfully describe common topics in the L2 is presumed to transfer to more academic topics during undergraduate study and indicates that the test taker is ready to begin English-medium education (see Section 1.2). The non-picture-based description tasks are presented and discussed in Pilot 2 but are presented again below for convenience.

Task 1: Tell me about something interesting you have recently heard in the news.

Task 2: Tell me about an event that has changed your life.

The two narrative tasks (see Figures 13 and 14) consist of six images depicting a pair of children playing a practical joke on a caregiver. The tasks feature identical numbers of characters and events and were therefore hypothesised to pose a relatively similar level of challenge to the test takers. This was important as the results of Pilot 1 showed that differences in the number of characters and events can cause the tasks to vary in terms of difficulty (see Sections 4.2.3.1 and 4.2.4.1). However, it is important to note that Inoue (2013) does not regard the tasks as equivalent: in her research Japanese learners of English did not perform equally well on the two tasks in terms of rater assessment or in terms of CAF measurement. Based on her interview data, Inoue (2013, p. 188) argues that Task 3 posed more of a cognitive challenge to her test takers than Task 4 because it contained culturally unfamiliar content (i.e. the 'balloon seller', 'washing-related objects'). Although the participants in the current study come from a different cultural background from the context investigated in Inoue (2013), the datedness of the washing objects may have the effect that some task content is unfamiliar to the participants. Performance on Task 3 may therefore be more susceptible to variation in planning time, as unfamiliarity with task content increases task challenge and the more challenging the task the more benefits can be derived from planning (see Section 2.5.1.4). Nonetheless, it is evident that the tasks share more commonalities in terms of number of characters and events than those used in Pilot 1 and should pose a relatively similar challenge to test takers. In the analysis, the *Balloon Task* is referred to as Task 3 and the *Baby Task* is referred to as Task 4. Order of tasks and pre-task planning time was counterbalanced between the test takers (see Table 33) to compensate for any influence of task order and planning time order and to cancel out the differences in the tasks identified in Inoue (2013).

Table 33 Order of tasks and planning conditions

Test Taker	Task & Planning			
1,9,19,33,41	1=1 min.	2=10 min.	3=30 sec.	4=5 min.
2,10,20	3 =5 min.	4=30 sec.	2=1 min.	1=10 min.
3,11, 21	2=10 min.	1=1 min.	4=30 sec.	3= 5 min.
4, 12, 22	4=5 min.	3=30 sec.	1=10 min.	2=1 min.
5, 23	3=1 min.	4=10 min.	1=30 sec.	2=5 min.
6, 24	4=1 min.	3=10 min.	2=30 sec.	1=5 min.
7, 28	1=5 min.	2=30 sec.	3=10 min.	4=1 min.
8, 26	2=5 min.	1=30 sec.	4=10 min.	3=1 min.
13, 29, 45, 37	3=30 sec.	4=5 min.	1=1 min.	2=10 min.
14	4=30 sec.	3=5 min.	2=10 min.	1=1 min.
15	1=10 min.	2=1 min.	3=30 sec.	4=5 min.
16	3=10 min.	4=1 min.	1=5 min.	2=30 sec.
17	4=10 min.	3=1 min.	2=5 min.	1=30 sec.
18	1=30 sec.	2=5 min.	4=1 min.	3=10 min.
25	4=1 min.	3=10 min.	1=30 sec.	2=5 min.
27	1=30 sec.	2=5 min.	3=1 min.	4=10 min.
30, 46, 38	3=5 min.	4=30 sec.	1=10 min.	2=1 min.
31, 39, 47	4=30 sec.	3=5 min.	2=1 min.	1=10 min.
32, 40	4=5 min.	3=30 sec.	2=10 min.	1=1 min.
34, 42	1=10 min.	2= 1 min.	3=5 min.	4=30 sec.
35, 43	2=1 min.	1=10 min.	4=30 sec.	3=5 min.
36, 44	2=10 min.	1=1 min.	4=5 min.	3=30 sec.

Table 33a Number of samples per task and planning condition

Task	30 sec.	1 min.	5 min.	10 min.
1	8	17	5	17
2	5	17	8	17
3	17	6	17	7
4	17	7	17	6

Figure 13. Balloon task (Task 3)

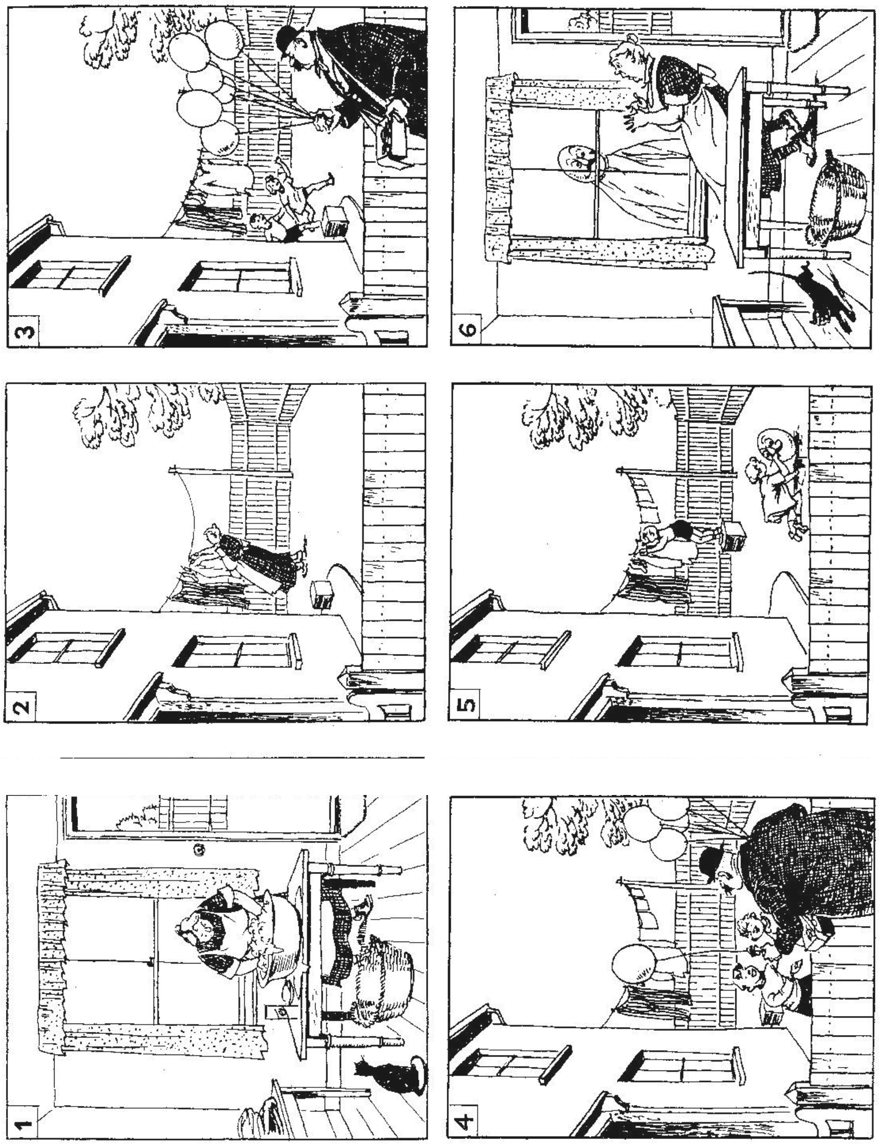
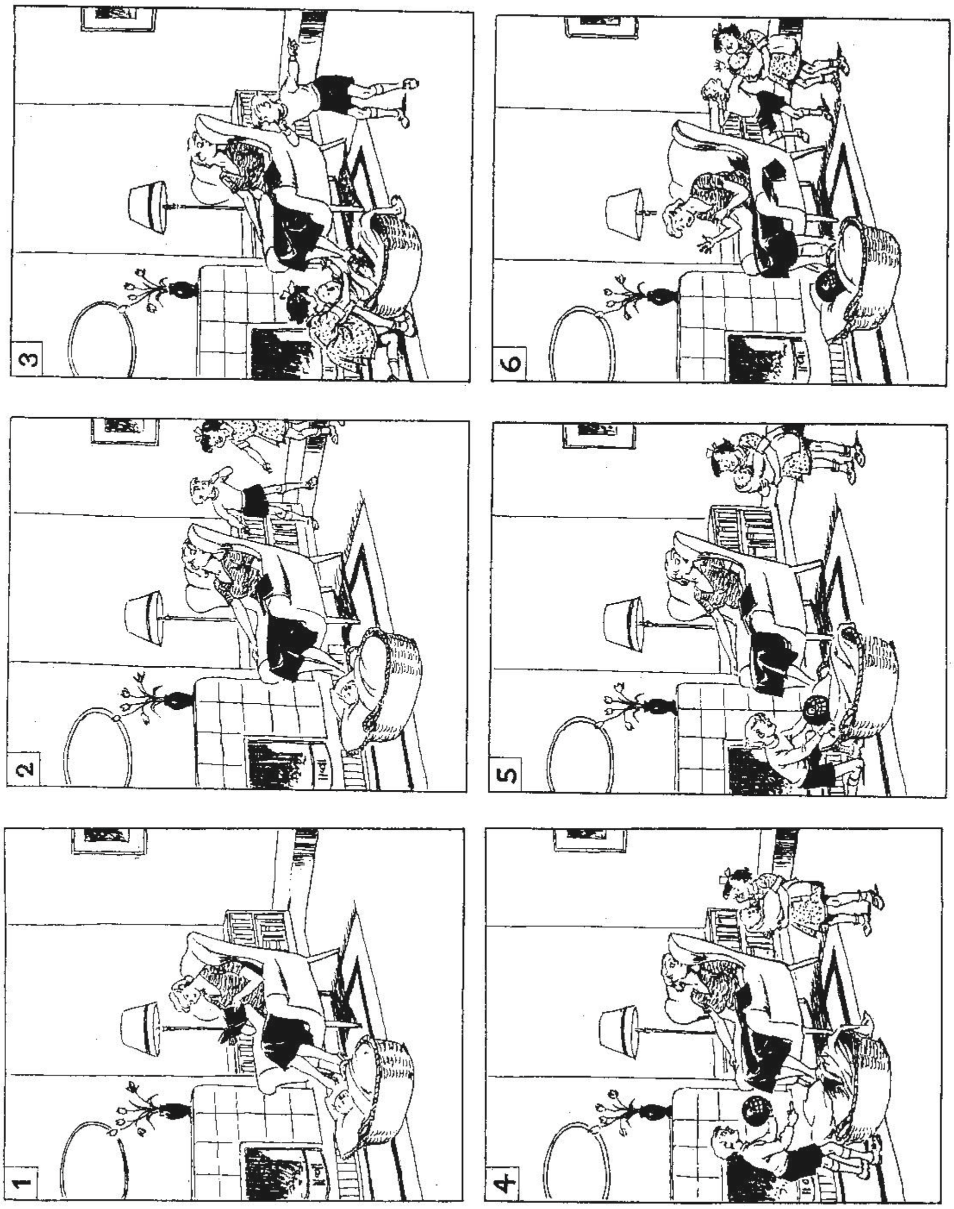


Figure 14. Baby task (Task 4)



5.4.3 Recording and transcription

Test takers completed the four speaking tasks in one session. The tests were recorded using the Audacity program (2.0.6, 29 September 2014, http://audacity.sourceforge.net) and saved as both Audacity files and MP3 files. The Audacity files were used to make transcriptions of the speech. The Audacity program presents the sound wave of the spoken sample in visual format. This facilitates the identification and measurement of disfluencies such as pauses and hesitations. Transcription was adapted from guidelines discussed in Fulcher and Davidson (2007) and Jefferson (2004) and from examples described in Foster et al. (2000). These methods of transcription provide guidelines for the recording and measuring of pauses, for unintelligible speech, and for relativisation, coordination and subordination (for an example of transcript sample see Section 7.2). It was important to be consistent when identifying such features of the speech in order for the CAF measures to produce reliable results. The MP3 files were used for scale construction, rater training and distribution of the tests to the raters.

5.4.4 EBB scale construction process

EBB scales require raters to make a series of binary decisions to separate the boundaries between performance levels (Turner and Upshur, 1996). Scale content is empirically derived from rater analysis of test performance. The scale is therefore rater oriented and designed for use within a specific assessment context (see Section 2.7.3.3). This level of specificity is intended to increase the relevance of the scale content to the users of the scale (the raters) and to reflect the range of abilities in the test taking population.

Seven English language instructors developed the EBB scale. This process involved various stages. The researcher listened to the recorded samples and holistically identified three samples per task: a high-ability level sample, a mid-ability level sample and a low-ability level sample. This ensured that a range of ability levels was represented for each task. Twelve samples were identified to generate the EBB rating scale. At the first stage of the scale construction process, the raters rank ordered the samples through paired comparisons. During this stage, raters were requested to take notes on features of the speech that they regarded as salient to their comparisons. The raters discussed their rank orders with recourse to their notes and agreed upon the rank ordering of five of the samples. The raters were unable to agree upon the rank ordering of the remaining seven samples. The scale therefore contains five levels of spoken proficiency. In the next stage, the raters discussed their criteria for making the five distinctions and formulated yes/no questions to identify boundaries between the ability levels.

The first comparison that the raters made was between levels 3, 4, 5 and levels 1, 2. The raters reached a consensus that the distinguishing feature of the top three levels was the test takers' ability to complete the task (the test taker was able to relate the events of the picture narrative or personal experience/news event) without causing undue effort for the rater. In other words, the raters were not required to make guesses about the test taker's communicative intention and had a clear understanding of what happened during the events being related.

The second comparison was between levels 1 and 2. The raters agreed that at level 2, test takers had completed the task but with substantial disfluencies involving long pauses and hesitations, repetitions and reformulations. In contrast at level 1, test performance was characterised by the test takers' inability to complete the task due to insufficient lexical knowledge. At level 1, test takers frequently abandoned utterances because they did not seem able to find words to express what they wanted to say.

The third comparison was between levels 3 and 4, 5. Raters observed a wider range of lexis and grammatical structures at levels 4 and 5 that allowed the test taker to express more nuanced descriptions of the events. For instance, in the following extract of a sample, the raters were impressed with the test taker's vocabulary (*republic, laic system*) and the use of the relative clause (*we can't do whatever we want*).

> er we haven't er real republic and | m we can't do :: whatever we want mm |
> we want to have a laic system in Turkey

The raters agreed that performance at levels 4 and 5 was characterised by 'effective use' of grammar and vocabulary.

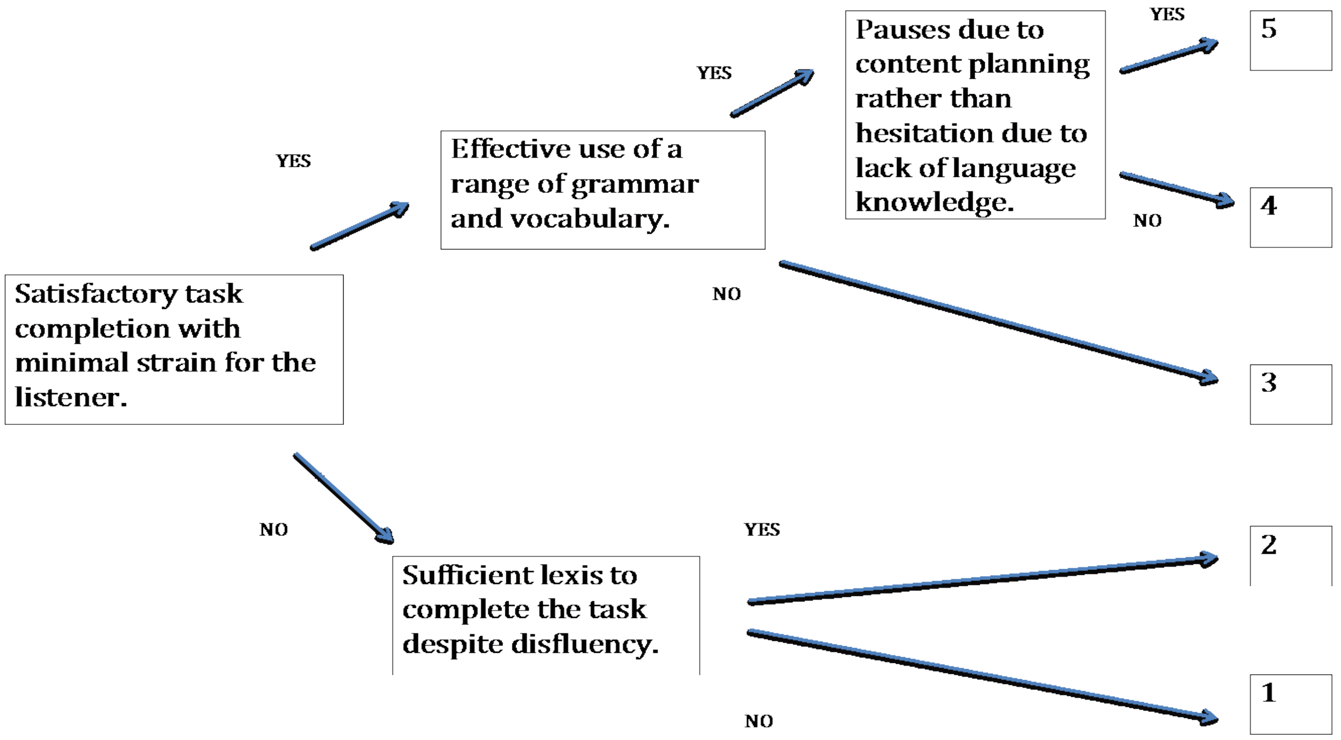
The final comparison was between levels 4 and 5. The raters agreed that test takers at levels 4 and 5 included gaps in their speech of very similar length. However, the consensus was that test takers at level 5 were planning content during these gaps, whereas at level 4 test takers needed to stop talking to search for lexis or complete a grammatical chunk. This is essentially a distinction between pausing for content and hesitating for language retrieval or assembly (Field, 2011). The distinction is exemplified in the following examples:

A. (1.5) they went to the (2) kitchen or somewhere (Level 4)

B. nowadays er my country has a problem | (2) this is like a war (Level 5)

In example A, the test taker hesitated to search for and consider a suitable word to match the image in Task 3. In contrast, in example B there were two pauses, the first occurred after the adverb and the second occurred between clauses: both at syntactic boundaries. The raters agreed that the pauses in example B did not disrupt the clause and helped the test taker think about the content of his next utterance, whereas the hesitation in example A was made because the test taker was unable to find the correct word to complete the utterance. The question used to distinguish between levels 4 and 5 asked about the apparent reason for disfluencies in the test takers' speech. The scale is presented in Figure 15. EBB Scale 3.

Figure 15. EBB scale 3


Satisfactory task completion with minimal strain for the listener.
YES
Effective use of a range of grammar and vocabulary.
YES
Pauses due to content planning rather than hesitation due to lack of language knowledge.
YES
5
NO
4
NO
3
NO
Sufficient lexis to complete the task despite disfluency.
YES
2
NO
1

5.4.5 EBB scale rating process

A standardisation session was held in which raters discussed and graded 13 samples together. Following the standardisation session, the raters were supplied with 20 test samples in MP3 format and asked to use the EBB scale to award grades. In order to run the multi-faceted Rasch measurement (MFRM), there should be overlap between the raters in the model (see Section 4.2.2.6). This requires different raters to assign scores to the same samples. Given the small number of raters and the large pool of samples, multiple matches between the raters were not possible. Therefore, Rater 1 awarded grades to each test sample (*n* = 188) to ensure that each sample was graded once. This yielded overlap between the raters (see Table 34; each X indicates that the rater assigned a score to at least one test taker in this row). Though it would have been desirable, each sample did not receive two ratings. However, each sample received at least one score, which provides a sufficient amount of connectivity between the facets to run MFRM using *Facets* (Linacre, 2013).

Table 34 Overlap between EBB scale 3 raters

Test Takers	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6	Rater 7
1-5	X	X	X	X	X	X	X
6-10	X	X	X	X	X	X	X
11-15	X	X	X		X		
16-20	X	X	X	X	X	X	X
21-25	X	X	X	X	X	X	X
26-30	X	X	X	X	X	X	X
31-35	X	X	X	X	X	X	X
36-40	X			X			
41-45	X			X			
45-47	X	X	X	X	X	X	X

5.4.6 Analytic scale standardisation and rating process

The raters using the analytic scale took part in a standardisation session in which three test samples were discussed and assessed by the group. The researcher identified three samples for standardisation to represent the high, mid and low ability levels of the database. Although it would have been desirable to include more samples in the standardisation session, time constraints meant that this was not possible. Following the standardisation session, the raters were assigned 20 samples to grade independently. Following the procedure adopted in the EBB rating process, Rater 1 provided grades for each sample (*n* = 188). This measure meant that there were sufficient matches between the raters to run the MFRM. The overlap between the raters is presented in Table 35 (each X indicates that the rater assigned a score to at least one test taker in this row).

Table 35 Overlap between analytic raters

Test takers	R 1	R 2	R 3	R 4	R 5	R 6	R 7	R 8	R 9	R 10
1-5	X	X	X	X	X	X	X	X	X	X
6-10	X	X	X				X	X		X
11-15	X	X	X	X	X	X	X	X	X	X
16-20	X		X	X	X	X		X		X
21-25	X		X	X	X	X	X	X	X	
26-30	X	X	X	X	X	X	X	X	X	
31-35	X	X		X	X	X	X		X	
36-40	X	X					X		X	
41-45	X					X	X		X	
45-47	X								X	

R refers to rater

5.4.7 Score analysis

The results of the EBB scale and analytic scale were analysed using *Facets* (3.71.4, 18 January 2014, www.winsteps.com), an application of MFRM (see Section 4.2.2.6). The facets under investigation in this study are specified in the following model, which was adapted from Linacre's (2013) model involving fewer facets:

$$\log (P_{nijlmsk} / P_{nijlmsk-1}) = B_n\text{-}D_i\text{-}R_j\text{-}C_l\text{-}A_m\text{-}G_s\text{-}F_k$$

B_n = ability of test taker n
D_i = difficulty of task i
R_j = severity of rater j
C_l = difficulty of scale l
A_m = proficiency of group m
G_s = time of planning s
F_k = difficulty of category k relative to k -1
$P_{nijlmsk}$ = probability of receiving rating k under these circumstances
$P_{nijlmsk-1}$ = probability of k-1

This model states that a test score results from the interaction between the test taker's ability, the difficulty of the task, the severity of the rater, the difficulty of the rating scale, the proficiency group that the test taker belongs to, and the amount of planning time the test taker was given. The inclusion of a common rater (R 1) allowed for the results of the EBB and analytic scales to share a common measure on the logit scale. The method of combining scores from different rating scales in MFRM is demonstrated in Turner and Upshur (2002) who use the method to assess the equivalence of three rating scales. In the current study, the common logit scale provides an overall measure of how pre-task planning impacted test scores across both scales and sets of raters.

The *Facets* output was evaluated to identify any cases of data misfit that might distort the model. This involved an evaluation of the infit mean-square index values provided in the *Facets* output (see Section 4.2.2.6). When infit mean-square index values exceed 2.0 this indicates that the data does not fit the model and poses a threat to the reliability of the measurement (Linacre, 2013). This may necessitate some form of data removal. Linacre (2010) recommends an iterative process for data removal in which the most distortive elements are removed until the data fit the model. *Facets* provides information about model misfit by identifying the number of unexpected responses that do not fit model: for example a very high score from a particularly severe rater. Data removal is described in detail in Section 6.2.

Logit measure values (referred to as measure values) indicating the effectiveness of each planning condition are supplied by the *Facets* output (see Section 4.2.2.6). *Facets* also calculates a chi-square test that indicates whether the overall difference in scores awarded under the four planning conditions is statistically significant (Linacre, 2013). However, the chi-square test does not identify precisely where the statistical significance is located (e.g. between the ten-minute condition and every other condition or just between the ten-minute condition and 30-second condition). In order to resolve this problem, Welch's (1951) t-tests were calculated on the logit values for each planning condition. This test was run in order to locate significant differences between the individual pre-task planning conditions. The Welch's t-test was selected as this test is capable of accounting for the unequal variances in the sample (Welch, 1951) that were likely to arise as a result of variation in the total number of scores assigned under each planning condition (30 = 163, 5 = 169, 10 = 194, 1 = 200).

Separate analyses were run using *Facets* to estimate the impact of the planning variable on each task (1, 2, 3, 4), the two rating scales (EBB and analytic), and the two proficiency levels (A1/2 and B1 on the CEFR, Council of Europe, 2001). This was achieved by altering the model to disregard certain facets. For example, in order to calculate the impact of the planning variable on the A level proficiency group (A1 n = 3, A2 n = 25), all instances of B level proficiency (B1 n = 12) were disregarded from the model. This allowed for measurement of the impact of the four planning conditions on the A level scores in which the B level information did not contribute to the result (e.g. by impacting fit statistics or severity measures).

5.4.8 CAF measures

A comprehensive description of the complexity, accuracy and fluency (CAF) measures and procedures involved in the development is provided in Section 4.3.2.3. The CAF measures used in the main study are listed below.

Complexity

- Guiraud's Index (G.INDEX)
- Lexical sophistication (K1/K2/AWL/NONE)
- Syntactic complexity (C.AS)
- Idea units (IDEAS)

Accuracy

- Articles (ART)
- Prepositions (PREP)

- Tense (TENSE)
- Verbs (VERBS)
- Self-correction (SELF)
- Pronouns (PRO)
- Mean number of errors per AS unit (M.N.E)

Fluency

- Mean number of hesitations per AS-unit (MNH)
- Phonation Time Ratio (PTR)
- Filled Hesitations (F.HES)
- Filled Pauses (F.PA)
- Mean Length of Utterance (MLU)
- Total Speaking Time (TST)
- Speech Rate (SPR)

5.4.9 CAF statistical analytical procedures

In the analysis, pre-task planning time, task number and proficiency group represent the independent variables. The 20 CAF measures represent the dependent variables. To assess the extent to which the three independent variables affected the results of the 20 dependent variables required a MANOVA test. In cases where statistical significance was reached in the MANOVA, for instance for the planning variable, ANOVA tests were required to identify how planning impacted the dependent variables, with an appropriately adjusted alpha level to account for the multiple tests (see Section 2.7.2.4).

5.5 Summary

This section has presented the methodology and the analytical approach of the main study (see Table 36). In the main study, four tasks were used to elicit speech from 47 test takers after varying amounts of pre-task planning time. The test takers were separated in to two language proficiency groups (A1/A2 and B1 on the CEFR, Council of Europe, 2001). Trained raters used an EBB scale and an analytic scale to assess the speech samples. The results were analysed using multi-faceted Rasch measurement (MFRM). In addition, an analysis of the complexity, accuracy and fluency (CAF) of the samples was completed from the transcripts. The results were analysed using statistical procedures, MANOVA and ANOVA.

Table 36 Summary of the research methodology

Research Question	Participants	Raters	Tasks and format	Measure(s) used	Analysis
1	47	17	Task 1 / 30 sec., 1 min. 5 min. 10 min. Task 2 / 30 sec., 1 min. 5 min. 10 min. Task 3 / 30 sec., 1 min. 5 min. 10 min. Task 4 / 30 sec., 1 min. 5 min. 10 min.	1. Logit measures 2. CAF	1. MFRM 2. MANOVA
1.1	47	17	Task 1 / 30 sec., 1 min. 5 min. 10 min. Task 2 / 30 sec., 1 min. 5 min. 10 min. Task 3 / 30 sec., 1 min. 5 min. 10 min. Task 4 / 30 sec., 1 min. 5 min. 10 min.	1. Logit measures 2. CAF	1. MFRM, Welch's t-test, Cohen's D 2. MANOVA, Cohen's D
1.2	47	7 EBB 10 Analytic	Task 1 / 30 sec., 1 min. 5 min. 10 min. Task 2 / 30 sec., 1 min. 5 min. 10 min. Task 3 / 30 sec., 1 min. 5 min. 10 min. Task 4 / 30 sec., 1 min. 5 min. 10 min.	Logit measures & fair average values	MFRM x 2 (each scale), Welch's t-test & Cohen's D
1.3	28 A level 12 B level	17	Task 1 / 30 sec., 1 min. 5 min. 10 min. Task 2 / 30 sec., 1 min. 5 min. 10 min. Task 3 / 30 sec., 1 min. 5 min. 10 min. Task 4 / 30 sec., 1 min. 5 min. 10 min.	1. Logit measures 2. CAF	1. MFRM x 2 (each level), Welch's t-test, Cohen's D 2. MANOVA, Cohen's D.
2	47	17	2 Picture-based tasks / 30 sec., 1 min. 5 min. 10 min. 2 Non-picture-based tasks / 30 sec., 1 min. 5 min. 10 min.	1. Logit measures 2. CAF	1. MFRM x 4 (each task), Welch's t-test, Cohen's D 2. ANOVA, Cohen's D
2.1	47	17	Task 1 / 30 sec., 1 min. 5 min. 10 min. Task 2 / 30 sec., 1 min. 5 min. 10 min. Task 3 / 30 sec., 1 min. 5 min. 10 min. Task 4 / 30 sec., 1 min. 5 min. 10 min.	1. Logit measures 2. CAF	1. MFRM x 4 (each task), Welch's t-test, Cohen's D 2. ANOVA, Cohen's D

6 Results

6.1 Introduction

This chapter reports the results of the main study. In order to account for the different analytical approaches to the assessment of the test takers (see Section 5.4.5, Section 5.4.6, and Section 5.4.8), the chapter is separated into two main sections: rating scales and complexity, accuracy and fluency (CAF). The multi-faceted Rasch measurement (MFRM) (see Section 5.4.7) of the test facets (test takers, raters, task, proficiency group, pre-task planning condition, rating scale) is first presented. Following this, the chapter examines the impact of the pre-task planning conditions on the analytic scale scores and the EBB scale scores. The effect of the four pre-task planning conditions on test scores is then presented according to test taker proficiency, and task type in the following subsections. In the second section, the results of the CAF analysis are presented. This section describes the interaction between planning, proficiency and task type in the CAF results.

6.2 Rating scales

The rating scale results were analysed using *Facets* (3.71.4, 18 January 2014, www.winsteps.com) with six facets entered into the analysis; test taker ability, rater severity, task difficulty, proficiency group (based upon the result of the Quick Placement Test, see Section 5.4.1), planning time, and rating scale. Raters 1-10 assigned scores using the analytic scale in which scores from three categories, complexity, accuracy and fluency were accumulated to produce an overall score from

15. Raters 1 and 11-16 used the EBB scale (see Figure 15), which contains five levels of ability. While it would have been desirable for equal numbers of raters to work with each scale, availability constraints meant that this was not possible. However, the raters awarded a sufficient number of scores to run MFRM using *Facets* (see Section 5.4.6).

An initial MFRM analysis demonstrated that three test takers (test taker 33: infit mean-square 2.19, test taker 36: infit mean-square 3.72, and test taker 20: infit mean-square 2.63) recorded a level of model misfit in which infit mean-square values exceeded 2.0 and so would have a distorting effect on measurement (see Section 5.4.7). *Facets* provides a list of unexpected responses (scores that do not fit the model) in the data set. The unexpected responses were checked to identify the causes of misfit. The misfitting test takers' raw scores were assessed for any potentially distorting observations and once they had been located, were removed. For instance, scores awarded to test taker 36 on the analytic scale are presented in Table 37. The table includes information about the raters, the rater severity measures, which report levels of leniency and severity relative to the rater population (see Section 4.2.3.1), scores and planning time. Rater severity in logits appears in parenthesis next to rater ID. Negative values represent leniency. The amount of planning time appears in parenthesis next to the score.

Table 37 Scores awarded to test taker 33 by six raters

Rater	R1 (-.10)	R4 (.53)	R5 (-.41)	R6 (.83)	R7 (-.39)	R9 (.11)
Grades	9 (10 min.) 10 (5 min.) 11 (1 min.) 3 (30 sec.)	8 (5 min.)	8 (30 sec.)	6 (5 min.)	6 (30 sec.)	6 (10 min.)

R refers to rater

As described in the Methodology (see Sections 5.3.2 and 5.3.3), Rater 1 assigned scores to each sample (under each pre-task planning condition: 30 sec., 1 min., 5 min., 10 min.). In addition to three relatively high scores, Rater 1 also awarded a score of 3, which contributed to the misfit. Inspection of the data showed that this was not a data entry error and this score was thus removed. Furthermore, Raters 6 and 7 varied in severity by 1.22 in logits but both provided a similarly low score of 6. The score that Rater 7 provided was uncharacteristically severe and was therefore also removed. In total 13 elements that contributed to the distorting infit mean-square values were removed. The removals reduced the total number of observations from 740 to 727. At this stage, an analysis was completed and the data fit the model (i.e. the infit mean-square values did not exceed 2.0, see Section 4.2.3.1). The results of the analysis are presented in the Wright map (see Figure 16).

Column one of the Wright map calibrates the six facets onto the common logit scale. This calibration allows for observations to be made regarding the overall impact of the six facets on the test scores (see Section 5.4.7). The second column presents the spread of test taker ability. The map ranks the test takers by positioning those receiving high scores toward the top. The map demonstrates that test taker ability measures were distributed between -2 and 2 logits approximately. This corresponds to a range of scores between 5 and 12 on the analytic scale (S.1) and 2 and 4 on the EBB scale (S.2).

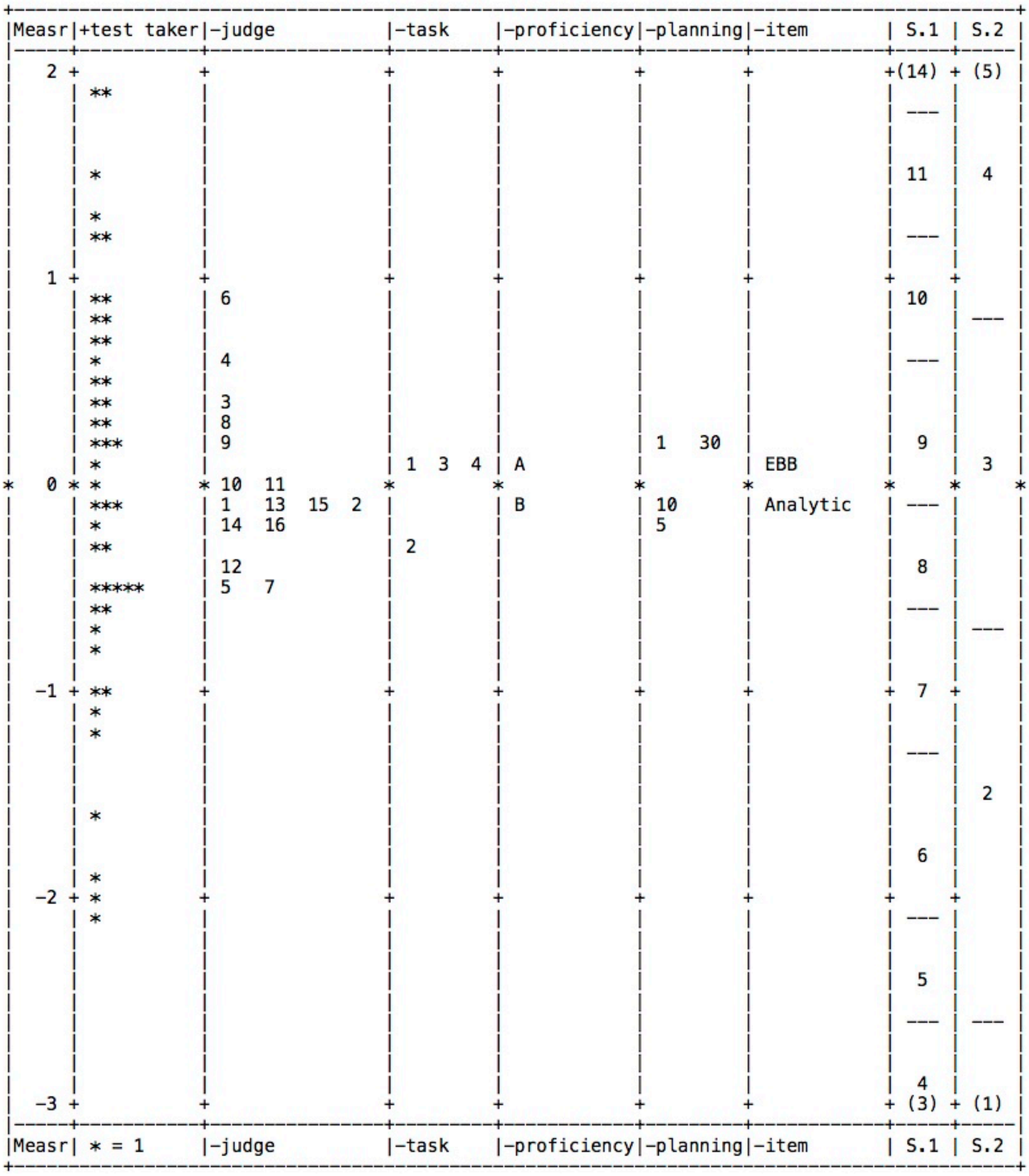
The third column, the rater column, orders raters by their relative severity, locating severe raters toward the top of the map and more lenient raters toward the bottom. According to the map, the raters demonstrate various levels of severity as indicated by their distribution. However, it is clear that the EBB scale raters (1 and 11-16) tend to cluster around zero logits to a greater extent than the analytic scale raters (1-10). This suggests that the EBB scale fostered more agreement between the raters than the analytic scale. This result may be a product of the greater number of levels on the analytic scale, which accumulates five levels of complexity, accuracy and fluency to calculate an overall score from 15 (the EBB scale contains five levels). It is also important to acknowledge that the EBB raters constructed the scale, whereas the analytic raters were standardised to an existing scale. The process of scale construction may have contributed to the high levels of agreement in the EBB scores. This is a strong basis for the use of EBB scales: they are easy to make and lead to higher levels of reliability than are possible through standardization of raters to existing scales (Turner and Upshur, 1996; see Section 2.7.3.2).

The remaining columns rank the facets in terms of difficulty by situating the most difficult element of each facet toward the top of the map. Column four, task, demonstrates that the test takers did not receive equal scores on all four tasks. For example, on Task 2 test takers were awarded substantially higher scores than on the other three tasks.

The fifth column, proficiency, indicates that the B level test takers achieved higher scores than the A level test takers. However, the difference between these groups was minimal at .16 of a logit (see Table 40). The pre-task planning facet is presented in column six. This column demonstrates that the highest scores were awarded under the five-minute planning condition followed by the ten-minute condition. Scores were very similar under the 30-second and one-minute planning conditions.

The next column presents the two scales. In this column, the analytic scale is located below the EBB scale. This indicates that test takers achieved higher scores on the analytic scale than the EBB scale. This result may be due to the higher number of levels on the analytic scale, which allows for closer distinctions between the levels of performance to be made. These findings are discussed in detail with reference to the MFRM statistics in the following section.

Figure 16. Wright map analytic scale and EBB scale 3



```
+--------------------------------------------------------------------------------------------+
|Measr|+test taker|-judge          |-task    |-proficiency|-planning|-item       | S.1 | S.2 |
|-----+-----------+----------------+---------+------------+---------+------------+-----+-----|
|   2 +           +                +         +            +         +            +(14) + (5) |
|     | **        |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            | --- |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |     |     |
|     | *         |                |         |            |         |            | 11  |  4  |
|     |           |                |         |            |         |            |     |     |
|     | *         |                |         |            |         |            |     |     |
|     | **        |                |         |            |         |            | --- |     |
|     |           |                |         |            |         |            |     |     |
|   1 +           +                +         +            +         +            +     +     |
|     | **        | 6              |         |            |         |            | 10  |     |
|     | **        |                |         |            |         |            |     | --- |
|     | **        |                |         |            |         |            |     |     |
|     | *         | 4              |         |            |         |            | --- |     |
|     | **        |                |         |            |         |            |     |     |
|     | **        | 3              |         |            |         |            |     |     |
|     | **        | 8              |         |            |         |            |     |     |
|     | ***       | 9              |         |            | 1   30  |            |  9  |     |
|     | *         |                | 1  3  4 | A          |         | EBB        |     |  3  |
*   0 * *         * 10  11         *         *            *         *            *     *     *
|     | ***       | 1   13  15  2  |         | B          | 10      | Analytic   | --- |     |
|     | *         | 14  16         |         |            | 5       |            |     |     |
|     | **        |                | 2       |            |         |            |     |     |
|     |           | 12             |         |            |         |            |  8  |     |
|     | *****     | 5   7          |         |            |         |            |     |     |
|     | **        |                |         |            |         |            | --- |     |
|     | *         |                |         |            |         |            |     | --- |
|     | *         |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |     |     |
|  -1 + **        +                +         +            +         +            +  7  +     |
|     | *         |                |         |            |         |            |     |     |
|     | *         |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            | --- |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |     |  2  |
|     | *         |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |  6  |     |
|     | *         |                |         |            |         |            |     |     |
|  -2 + *         +                +         +            +         +            +     +     |
|     | *         |                |         |            |         |            | --- |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |  5  |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            | --- | --- |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |     |     |
|     |           |                |         |            |         |            |  4  |     |
|  -3 +           +                +         +            +         +            + (3) + (1) |
|-----+-----------+----------------+---------+------------+---------+------------+-----+-----|
|Measr| * = 1     |-judge          |-task    |-proficiency|-planning|-item       | S.1 | S.2 |
+--------------------------------------------------------------------------------------------+
```

*Planning: 30 seconds, 1 minute, 5 minutes, 10 minutes

6.2.1 Facets in the MFRM model

Test Takers. Test taker ability measures ranged from -2.12 to 1.93 on the logit scale. These figures indicate that the range of test taker ability in the sample was greater than the range of rater severity, task difficulty, planning time difficulty, and scale effect. The separation index was 3.72 and the strata were 5.30. Using the strata, there were approximately five statistically distinct levels of performance in the test taking population. The mean of standard errors was .24 indicating a degree of imprecision in the measurement by .24 of a logit (see Section 4.2.3.1). Test taker infit mean-square statistics ranged from .39 to 1.99. While the fit statistics do indicate a degree of misfit, the measures do not exceed 2.00 (Linacre, 2013). The reliability of the separation was high at .93. Reliability figures that are close to 1.00 indicate that the test taker separation was measured reliably with minimal measurement error (see Section 4.2.3.1).

Raters. The rater statistics are presented in Table 38. The first statistic that is reported is the severity measure. The severity measures ranged from -.51 to .90 indicating that raters demonstrated various degrees of severity. Of the 16 raters involved in the test, it is evident that the five most severe raters awarded scores using the analytic scale (R9, R8, R3, R4, and R6). Furthermore, the two most lenient raters also used the analytic scale (R7 and R5). Relative to the EBB scale, the raters that used the analytic scale were less consistent and recorded severity levels that ranged from -.51 to 90. Raters tended to agree more when using the EBB scale; the raters that used the EBB scale (R1 and R 11-16) demonstrated levels of severity that ranged from -.42 to 0 on the logit scale.

The separation index was 1.44 and the strata were 2.25. These figures demonstrate that the model identified approximately two levels of rater severity in the data: a lenient level and a severe level. The reliability of rater separation was .67, which indicates that the raters demonstrated consistently different levels of severity when awarding scores. Concerning the infit mean-square statistics, as stated in

Section 4.2.3.1, rater infit mean-square values that exceed '2.00' indicate that the scores are unpredictable and distort the measurement (Linacre, 2013, p. 266). According to this standard, rater fit statistics were acceptable and did not distort the measurement. However, raters 11 (1.73) and 5 (1.73) did record levels of misfit that approached 2.00. This indicates that these raters tended to be inconsistent when awarding scores. However, the fit statistics did not exceed 2.00 signifying that the raters were not excessively inconsistent and could be retained.

Table 38 Report of rater severity

Rater	Severity Estimate	Error	Infit mean-square index
7	-.51	.17	.99
5	-.46	.17	1.73
12	-.42	.22	.99
16	-.20	.37	.70
14	-.17	.23	1.05
2	-.15	.20	1.43
15	-.14	.35	.67
13	-.13	.22	.58
1	-.13	.05	.89
10	-.01	.16	1.04
11	.00	.22	1.73
9	.18	.17	.89
8	.34	.15	1.39
3	.36	.16	1.01
4	.56	.17	1.44
6	.90	.17	1.21
Mean	.00	.20	1.11
SD	.37	.07	.34
Reliability of difference in severity of raters .67			

Tasks. Table 39 presents the MFRM task statistics. The table shows that the tasks varied in terms of difficulty. Furthermore, the results of the chi-square test (see Section 5.4.7) demonstrated that these differences were statistically significant at $p < .001$. It is evident that on average, the highest scores were recorded when test takers completed Task 2 ('*Tell me about an event that has changed your life*'). In contrast, test takers recorded the lowest scores on Tasks 1 (‘*Tell me about something interesting you have recently heard in the news*’) and 3 (*Balloon Task*). Task 4 (*Baby Task*) was situated between the difficult tasks and the simple task on the logit scale.

Table 39 Measure and infit statistics: tasks

Task	Measure	Infit mean-square index	Fixed (all same) chi-square
2	- .35	1.14	χ^2 = 34.6, *p* < .001
4	.07	.81	
1	.14	1.14	
3	.14	1.02	

Proficiency. Table 40 presents the results of the proficiency analysis. The measure statistics demonstrated that the B level proficiency group achieved higher scores than the A level group. This is not a surprising result given the difference in scores on the QPT (see Section 5.4.1). The chi-square test showed that this result was statistically significant (*p* = .05) at *p* = .04. However, the difference between the groups was minor at .16 of a logit and indicates that variation in test taker proficiency as demonstrated by the QPT did not amount to a substantial difference in terms of speaking test scores.

Table 40 Measure and infit statistics: proficiency

Proficiency	Measure	Infit mean-square index	Fixed (all same) chi-square
B	- .08	1.08	χ^2 = 4.3, *p* = .04
A	.08	1.00	

Planning. The planning statistics are presented in Table 41. The measure statistics demonstrate that the five-minute planning condition resulted in the highest scores (-.20), followed by the ten-minute condition (-.12). The difference between having 30 seconds and one minute for pre-task planning did not affect the logit measures; the measure value for each condition was .16. However, differences in individual test taker's scores between the 30-second and one-minute planning conditions are discussed in Section 6.2.2.1. Results of the chi-square test indicated that the overall difference between the planning conditions was statistically significant at $p < .001$.

These figures suggest that the optimal period of pre-task planning (in terms of increasing scores) was five minutes rather than ten minutes. However, the measure figures show that variation between five minutes and ten minutes pre-task planning time made little practical difference to the test scores: the difference was minimal at .08 of a logit. The distance in logits between the five-minute and one-minute/ 30-second planning conditions was larger at .36. This indicates that pre-task planning made a difference to test scores but that the difference was negligible.

Table 41 Measure and infit statistics: planning

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .20	1.22	$\chi^2 = 21.2, p < .001$
10 min.	-.12	1.02	
30 sec.	.16	.78	
1 min.	.16	1.09	

A series of Welch's t-tests was calculated on the data in order to identify statistically significant differences between the pre-task planning conditions. The analysis used a Bonferroni adjusted alpha level of p = .008 (.05/6) to account for the multiple statistical tests that were completed. In addition, the effect size was calculated using Cohen's *d* values (Cohen, 1988).

The results of the Welch's t-tests demonstrated statistically significant differences between scores under the following planning conditions: five minutes and one minute (t(364)= 3.64, p < .001) with a small to moderate effect size of .38, five minutes and 30 seconds (t(329)= 3.64, p < .001) with a small to moderate effect size of .40, ten minutes and one minute (t(391) = 2.83, p = .005) with a small effect size of .29, and ten minutes and 30 seconds (t(352)= 2.83, p = .005) with a small effect size of .30. Statistical significance was not reached for the difference between the scores awarded under the five-minute and ten-minute conditions (t(359)= 0.81, p = .42). Overall, the results indicated that scores were higher when the speaking test included extra pre-task planning time (five minutes and ten minutes). However, the differences in scores between the five-minute and ten-minute planning conditions did not reach statistical significance.

Scale. Table 42 presents the rating scale statistics. The severity measures demonstrate that test takers received higher scores on the analytic scale than the EBB scale. However, the difference was minimal at .10 of a logit. Furthermore, the chi-square test demonstrated that this result did not reach statistical significance.

Table 42 Measure and infit statistics: scales

Scale	Measure	Infit mean-square index	Fixed (all same) chi-square
Analytic	- .05	1.09	$\chi^2 = 1.9, p = .17$
EBB	.05	.89	

6.2.2 Further analyses

The literature review and the results of the pilot studies (see Sections 4.2 and 4.3) indicate that the pre-task planning effect may be stronger for certain task types, test takers, and with different rating scales. Separate MFRM analyses were therefore conducted on the scores of the two scales, the two proficiency groups, and the four tasks (see Section 5.4.7). The results of these separate analyses are presented in the following sections.

6.2.2.1 Planning and scale

Table 43 reports the impact of the planning variable on the analytic scale scores. The results of this analysis are relatively consistent with those observed in the overall MFRM (see Section 6.2.1). The five-minute planning condition resulted in the highest scores of 8.83 fair average, followed by the ten-minute condition at 8.65. However, in contrast to the overall analysis, there was a clear difference in scores between the 30-second and one-minute planning conditions. The table demonstrates an increase of .20 in logits on the fair average scale when test takers planned their speech for 30 seconds in comparison to 1 minute. On the analytic scale, the one-minute planning condition therefore resulted in the lowest scores. This finding suggests that completing the task relatively spontaneously, after 30 seconds, was more beneficial than completing the tasks after a limited time to plan, one minute. The result of the chi-square test was significant at $p < .001$.

Table 43 Analytic scale results: planning

Planning	Measure	Fair Average	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .21	8.83	1.14	$\chi^2 = 17.3, p < .001$
10 min.	-.10	8.65	1.04	
30 sec.	.10	8.31	.82	
1 min.	.21	8.11	.99	

Welch's t-tests showed that the differences between scores under the following planning conditions reached statistical significance: five minutes and one minute (t(199)= 3.71, $p < .001$) with a moderate effect size of .53, five minutes and 30 seconds (t(185)= 2.74, $p = .007$) with a small to moderate effect size of .40, and ten minutes and one minute (t(195)= 2.74, $p = .007$) with a small to moderate effect size of .39. The difference between the ten-minute and 30-second planning conditions did not reach statistical significance (t(181)= 1.7678, $p = .079$). Furthermore, statistical significance was not reached for the differences between scores awarded under the five-minute and ten-minutes planning conditions (t(187)= 0.9723, $p = 0.33$) or the 30-second and one-minute planning conditions (t(192)= 0.9723, $p = 0.33$). These results indicate that variation in planning time did impact on the test scores on the analytic scale but this impact was limited. The five-minute condition resulted in the highest scores and the largest effect size was between the five-minute and one-minute conditions. However, the increase in scores observed between the five-minute condition and the ten-minute condition was not statistically significant.

Table 44 reports the results of the EBB scale analysis. The table shows that the five-minute planning condition resulted in the highest scores of 3.08 fair average on the EBB scale, this was followed by the ten-minute condition at 3.06. The difference between the scores awarded under the five-minute and ten-minute conditions was minimal. In terms of the shorter planning conditions, the one-minute planning condition resulted in scores that were .15 higher on the fair average scale than the 30-second condition. This is a reversal of the result observed on the analytic scale. At its most extreme, the difference between the planning conditions was .66 of a logit, which corresponds to a difference of .31 on the EBB scale. The result of the chi-square test was statistically significant result *p* = .01.

Table 44 EBB scale results: planning

Planning	Measure	Fair Average	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .26	3.08	1.24	χ^2 = 10.5, *p* = .01
10 min.	-.23	3.06	.69	
1 min.	.08	2.92	1.17	
30 sec.	.40	2.77	.75	

The Welch's t-tests demonstrated statistically significant differences between the scores under the following planning conditions: five minutes and 30 seconds (*t*(142)= 2.75, *p* = .007) with a moderate effect size of .46, and ten minutes and 30 seconds (*t*(157) = 2.78, *p* = .006) with a moderate effect size of .43. Differences between the remaining results did not reach significance: five minutes and ten minutes (*t*(157)= 0.13, *p* = .90), five minutes and one minute (*t*(154)= 1.5, *p* = .14), ten minutes and one minute (*t*(193)= 1.46, *p* = .15), and one minute and 30 seconds (*t*(154)= 1.41, *p* = .16).

Comparison of the results of the two scales indicates that variation in pre-task planning time had more of an impact on the analytic scale than the EBB scale. Statistically significant differences between scores awarded under different planning conditions were more frequently observed on the analytic scale than the EBB scale.

So far, the results have been reported in terms of overall fair average values. However, it is also necessary to examine the way that variation in planning time affected the individual test takers' scores. In order to achieve this, this study adapted a method applied by Inoue (2013). Using the logit measurement figures generated by *Facets*, the levels that would be assigned to each candidate under the four different planning conditions on each scale can be determined by calculating:

$$B_n\text{-}G_s\text{-}R_i$$

B_n is the ability of candidate *n*, G_s is the difficulty of planning condition *s*, and R_i is the severity of rater *i* (Linacre, 2013). This calculation generates a logit value that can be mapped on to the levels of ability on the analytic scale and the EBB scale.

Table 45 displays the levels on the analytic scale and the corresponding logit values. The logit values represent a transition point which, when exceeded signify a step up on the analytic scale (Inoue, 2013). For example, the probability of receiving a level 4 on the scale increases when the test taker is assigned a logit value that is higher than -2.72 (Linacre, 2013). A logit value that is lower than -2.72 suggests that the probability of the test taker being awarded a level 3 is higher than that of the test taker being awarded a level 4.

Table 45 Analytic levels and corresponding logit values

Analytic Level	Logit Value
3	
4	(-2.72)
5	(-2.53)
6	(-2.31)
7	(-1.18)
8	(-.48)
9	(-.19)
10	(.64)
11	(1.28)
12	(1.51)
13	(2.61)
14	(3.27)

The rater severity measures were set to 0 to ensure the measurement of the effect of the four pre-task planning conditions was not affected by variation in rater severity (Inoue, 2013). The test takers' scores were then evaluated using the formula discussed above. For example, in the MFRM of the analytic scale data, test taker 46 had an ability value of .73. The scores predicted by the model for test taker 46 given a rater with a severity level of 0 were:

5 minutes: .73 – (-.21) – 0 = .94

10 minutes: .73 – (-.10) – 0 = .83

30 seconds: .73 – (.10) – 0 = .63

1 minute: .73 – (.21) – 0 = .52

According to these results, test taker 46 would receive a score of 10 on the analytic scale under the five-minute and ten-minute planning conditions, but a score of 9 under the 30-second and one-minute planning conditions. Therefore, variation in the amount of planning time had an important impact on this test taker's scores. For this test taker, extra planning time (five minutes and ten minutes) led to an increase of one level on the analytic scale.

The calculation was carried out on each of the four test samples produced by every test taker. Table 46 reports the number of test takers that would receive different grades after variation in pre-task planning time. The table shows that 26 of the test takers would receive different scores after variation in planning time if they were scored by a rater with a severity level of 0. In some cases the test taker would be placed into three different levels after planning for different lengths of time. For example, test taker 22 would receive a score of 9 under the five-minute condition, a score of 8 under the ten-minute condition, and a score of 7 under the 30-second, and one-minute conditions. Test taker 47 would receive a score of 9 under the five-minute condition, a score of 8 under the ten-minute and 30-second conditions, and a score of 7 under the one-minute condition. However, in most cases the increases in scores would be restricted to one level on the scale.

Table 46 Variation in scores after planning on the analytic scale

Test Taker	5 minutes	10 minutes	30 seconds	1 minute
4	8	8	7	7
6	10	10	9	9
7	11	10	10	10
9	10	10	9	9
10	8	7	7	7
14	9	9	9	8
15	10	10	9	9
16	10	10	10	9
18	8	8	7	7
19	10	9	9	9
21	9	9	9	8
22	9	8	7	7
23	7	7	7	6
24	9	9	9	8
25	11	10	10	10
26	7	7	6	6
27	8	8	7	7
29	7	7	6	6
33	9	9	8	8
34	10	10	10	9
36	10	9	9	9
39	9	8	8	8
41	10	9	9	9
44	9	8	8	7
46	10	10	9	9
47	9	8	8	7

Table 47 presents the levels on the EBB scale and the minimum logit value that test takers need to receive in order to be placed within these levels. For example, when test takers are awarded a logit value that is higher than 1.33, the probability of receiving level 4 on the EBB scale exceeds the probability of receiving level 3 (Linacre, 2013). If a test taker receives a logit value that is lower than -2.99, the probability of the test taker being assigned to level 1 exceeds the probability of the test taker being assigned to level 2.

Table 47 EBB levels and corresponding logit values

EBB Level	Logit Value
Level 1	
Level 2	-2.99
Level 3	-.98
Level 4	1.33
Level 5	2.64

To demonstrate the equation, the calculations for test taker 46 are presented below. Test taker 46 had an ability value of 1.24 and the four equations were:

5 minutes: 1.24 – (-.26) - 0 = 1.50

10 minutes: 1.24 – (-.23) - 0 = 1.47

1 minute: 1.24 – (.08) – 0 = 1.16

30 seconds: 1.24 - (.40) – 0 = .84

These results indicate that test taker 46 would be assigned to level 4 under the five-minute, and ten-minute conditions, but would be assigned to level 3 under the one-minute, and 30-second conditions. Thus planning had a meaningful impact on this test taker's scores. Table 48 displays the number of test takers that would also be placed into different levels on the EBB scale after variation in planning time. Twelve of the test takers would receive different scores after planning for different lengths of time. However, in contrast to the analytic results, on the EBB scale the increase in scores after variation in planning time would only be one level.

Table 48 Variation in scores after planning on the EBB scale

Test Taker	5 minutes	10 minutes	1 minute	30 seconds
6	4	4	3	3
7	5	5	4	4
10	3	3	3	2
16	4	4	4	3
24	3	3	3	2
29	3	3	2	2
34	4	4	4	3
38	2	2	1	1
41	4	4	3	3
45	3	3	3	2
46	4	4	3	3
47	3	3	3	2

6.2.2.2 Analytic scale: complexity, accuracy, and fluency

MFRM was completed on the three categories of the analytic scale to measure the impact of the planning variable on complexity, accuracy and fluency scores. The first category that is discussed is complexity. The results are presented in Table 49. The five-minute planning condition resulted in the highest complexity scores at 2.80 fair average. This value is a minor increase over the ten-minute planning condition, which resulted in fair average scores of 2.77. The difference between these conditions in terms of fair averages was therefore minimal. Regarding the one-minute and 30-second conditions, the one-minute condition resulted in the lowest scores on the complexity category. The 30-second condition resulted in scores that were .09 higher in fair average than the one-minute condition. The chi-square test demonstrated that the difference in complexity scores awarded under the four planning conditions was statistically significant at $p = .01$.

Table 49 Results after different planning conditions: complexity

Planning	Measure	Fair Average	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .37	2.80	1.03	χ^2 = 12.1, *p* = .01
10 min.	-.26	2.77	.94	
30 sec.	.17	2.65	1.02	
1 min.	.47	2.56	.97	

Welch's t-tests (*p* = .008) were completed to identify statistically significant differences between complexity scores. The only result that reached statistical significance was the difference between the five-minute and one-minute planning conditions (*t*(202)= 3.13, *p* = .002) with a moderate effect size of .44. This indicates that the five-minute planning condition led to a fair average value that was .24 higher than the least beneficial planning condition, one minute. This increase is minor and does not translate into any meaningful impact of planning on the complexity scores. The results of the ten-minute planning and one-minute planning conditions approached statistical significance but did not meet the adjusted alpha (*t*(194)= 2.6462, *p* = .009). The remaining results did not reach statistical significance: five minutes and ten minutes (t(191)= 0.4 *p* = .69), five minutes and 30 seconds (t(188)= 1.96, *p* = .05), ten minutes and 30 seconds (*t*(185) = 1.52, *p* = .13), one minute and 30 seconds (*t*(194)= 1.09, *p* = .28).

Table 50 presents the accuracy category statistics. The ordering of the pre-task planning conditions matched the overall analytic results and the complexity results. The five-minute planning condition resulted in the highest scores, followed by the ten-minute and 30-second conditions. The one-minute planning condition resulted in the lowest accuracy scores. However, the difference between the fair average values was negligible. The five-minute planning condition caused an increase of .19 in fair average over the one-minute condition. Furthermore, the chi-square test demonstrated that the differences between the planning conditions did not reach statistical significance. Variation in pre-task planning time did therefore not make a statistically significant difference to the accuracy scores.

Table 50 Results after different planning conditions: accuracy

Planning	Measure	Fair Average	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .21	2.91	1.08	$\chi^2 = 6.4, p = .09$
10 min.	-.14	2.88	1.03	
30 sec.	-.03	2.85	.84	
1 min.	.37	2.72	.96	

Table 51 presents the fluency category statistics. The general ordering of the pre-task planning conditions was consistent with the overall analytic results; the five-minute condition resulted in the highest scores, followed by the ten-minute and 30-second planning conditions. Once again, the one-minute planning condition resulted in the lowest scores on the fluency category. However, it is clear that planning made more of a difference to the fluency scores than the complexity and accuracy scores. When test takers completed the tasks under the five-minute or ten-minute planning conditions, fair average values passed into the following level on the analytic scale (from high level 2 to low level 3). The planning variable therefore appears to have made a meaningful difference to fluency scores. The result of the chi-square test was significant at $p < .001$.

Table 51 Results after different planning conditions: fluency

Planning	Measure	Fair Average	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .50	3.15	1.22	$\chi^2 = 19.6, p < .001$
10 min.	-.30	3.09	.91	
30 sec.	.40	2.87	.95	
1 min.	.40	2.87	.80	

The Welch's t-tests demonstrated that statistical significance was reached for the difference between scores under the five-minute and one-minute conditions ($t(188)= 3.44$ $p < .001$) with a moderate effect size of .48, the five-minute and 30-second conditions ($t(202)= 3.54$, $p < .001$) with a moderate effect size of .51, and the ten-minute and 30-second conditions ($t(196)= 2.68$, $p = .008$) with a small to moderate effect size of .39. Furthermore, statistical significance was approached for difference between scores under the ten-minute and one-minute conditions ($t(185)=$ 2.61, $p = .01$). Fluency scores were clearly sensitive to differences in pre-task planning time and the five-minute planning condition appears to result in the highest fluency scores. However the Welch's t-test showed that the difference between the five-minute and ten-minute results did not reach statistical significance ($t(190)= 0.76$, $p = .45$). There is little evidence of an optimal planning condition in the fluency scores. In the same way, it is not possible to state categorically that a specific planning condition was most detrimental to speech fluency as both the 30-second and one-minute conditions generated the same measure value.

6.2.2.3 Planning and proficiency

Table 52 reports the results of the MFRM of the effect of variation in planning on proficiency group A scores (A1 and A2 on the CEFR, Council of Europe, 2001). The results show that the ten-minute planning condition resulted in the highest scores at -.16 on the logit scale. This finding is inconsistent with the results observed in the overall analysis in which the five-minute condition resulted in the highest scores. However, the difference between the ten-minute and five-minute planning logit values was minimal at .03. The difference between the one-minute and 30-second conditions was .18 of a logit, indicating that the one-minute planning condition increased the A level test takers' scores in relation to the 30-second planning condition. The result of the chi-square test was statistically significant (χ^2 = 10.3, *p* = .02).

Table 52 Results after different planning conditions: proficiency group A

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min.	-.16	.90	χ^2 = 10.3, *p* = .02
5 min.	-.13	1.09	
1 min.	.05	1.24	
30 sec.	.23	.73	

The Welch's t-tests demonstrated statistically significant differences between the scores awarded under the ten-minute and 30-second conditions (*t*(193)= 2.76, *p* = .007) with a small to moderate effect size of .40. Increasing planning time from 30 seconds to ten minutes therefore caused proficiency group A scores to increase by .39 on the logit scale. The remaining tests did not reach statistical significance: five minutes and ten minutes (*t*(189)= 0.21, *p* = .83), five minutes and one minute (*t*(193)=

1.34, *p* = .18), five minutes and 30 seconds (*t*(181)= 2.55, *p* = .01), ten minutes and one minute (*t*(210)= 1.56, *p* = .12), and one minute and 30 seconds (*t*(199)= 1.34, *p* = .18). The A level proficiency group derived the most benefit from the opportunity to plan when the planning time was ten minutes. Moreover, the increases under the ten-minute planning condition only reached statistical significance when the comparison was with the 30-second planning condition, the shortest period of planning time.

To further explore the impact of test taker proficiency on planning effects, MFRM was completed on the scores awarded to the three test takers that scored A1 on the QPT. Table 51a reports the results. However, it is important to bear in mind that, given the low number of test takers in this group, the results are indicative and should be interpreted with caution. The ten-minute planning condition led to the highest scores at -1.28 on the logit scale (negative values indicate higher scores on the logit scale). The lowest scores were awarded under the 30-second condition, which resulted in a value of 1.47 on the logit scale. The result of the chi-square test was statistically significant at p = .00.

Table 52a Results after different planning conditions: proficiency group A 1

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min.	-1.28	.60	χ^2 = 15.0, *p* = .00
5 min.	-1.12	1.70	
1 min.	.94	.58	
30 sec.	1.47	1.02	

The Welch’s t-tests demonstrated that scores awarded under the following planning conditions were statistically significant: ten minutes and one minute (*t*(26) =

10.8925, p < .001) with an effect size of 3.81, ten minute and 30 seconds (t(10) = 11.9429, p < .001) with an effect size of 4.43, five minutes and one minute (t(8) = 6.8530, p < .001) with an effect size of 3.25, five minutes and 30 seconds (t(10) = 8.1159, p < .001) with an effect size of 3.88, and one minute and 30 seconds (t(25) = 2.6518, p = .01) with an effect size of 0.94. The differences between the ten-minutes and five-minute conditions did not reach statistical significance (t(10) = 0.4975, p = .6296).

Table 53 reports the MFRM of the impact of the planning variable on the B level proficiency group scores. Logit values increase from .13 logits under the one-minute planning condition to -.18 logits under the five-minute planning condition. However, it is clear that the planning variable did not impact the scores to the same extent as the A level proficiency group. Furthermore, the chi-square test demonstrated that these results did not reach statistical significance. These results indicate that proficiency was a key variable in the study. Specifically, the findings suggest that low proficiency test takers benefitted from extra planning time (see Section 8.5.1).

Table 53 Results after different planning conditions: proficiency group B

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .18	1.33	$\chi^2 = 3.1, p = .37$
30 sec.	.01	.92	
10 min.	.03	.97	
1 min.	.13	.84	

6.2.2.4 Planning and task

Separate analyses were conducted on each task by specifying that the MFRM model should disregard all tasks except the one under investigation in the analysis (see Section 5.4.7). For instance, to assess the impact of planning on Task 1 results, Tasks 2, 3, and 4 data were discounted from the analysis. Table 54 presents the results of the planning variable on Task 1: *Tell me about something you have recently heard in the news*. The ordering of the planning conditions by the measure statistics indicated that the ten-minute planning condition resulted in the highest scores. The measure statistics demonstrated that the ten-minute planning condition resulted in substantial increases over the five-minute condition (-.27), 30-second condition (.11), and one-minute condition (.79). The one-minute planning condition resulted in scores that were considerably lower than the remaining conditions on the logit scale. At its most extreme, the difference between the measure values was 1.42 in logits (between scores under the ten-minute and one-minute conditions). Results of the chi square test indicate that the impact of planning was statistically significant at $p < .001$.

Table 54 Results after different planning conditions: Task 1

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min.	- .63	.80	$\chi^2 = 3, p < .001$
5 min.	-.27	1.40	
30 sec.	.11	.89	
1 min.	.79	1.10	

The Welch’s t-tests demonstrated that the following differences between the scores were statistically significant: ten minutes and one minute (t(146)= 7.72, p <

.001) with a large effect size of 1.28, ten minutes and 30 seconds (t(59)= 3.10, *p* = .003) with a moderate effect size of .66, five minutes and one minute (*t*(25)= 3.7, *p* < .001) with a large effect size of .99, and one minute and 30 seconds (t(60)= 2.85, *p* = .006) with a moderate effect size of .59. The ten-minute planning condition resulted in scores that were substantially higher than those recorded under the one-minute planning condition, and the 30-second planning condition. Increases were also observed under the five-minute and 30-second planning conditions in relation to the one-minute planning condition. The differences reported between the ten-minute and five-minute planning conditions (*t*(25)= 1.28, *p* = .21), and the five-minute and 30-second planning conditions (*t*(35)= 1.19, *p* = .24) did not reach statistical significance.

The impact of the planning variable on Task 2 results is reported in Table 55. Task 2 required test takers to complete the following task: *Tell me about an event that has changed your life*. It is clear that the ordering of the planning conditions varied from that observed on Task 1. On Task 2, the highest scores were awarded under the five-minute planning condition, followed closely by the 30-second condition. The ten-minute condition did lead to higher scores but this was only in relation to the least beneficial, one-minute planning condition. The chi-square test demonstrates that the impact of planning was statistically significant at *p* < .001.

Table 55 Results after different planning conditions: Task 2

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
5 min.	- .34	1.85	$\chi^2 = 3, p < .001$
30 sec.	-.32	.57	
10 min.	-.25	.86	
1 min.	.91	.87	

The Welch's t-test results demonstrated that the differences between the following scores were statistically significant: five minutes and one minute (*t*(73)= 5.48, *p* < .001) with a large effect size of 1.16, ten minutes and one minute (*t*(129)= 5.86, *p* < .001) with a large effect size of 1.03, and 30 seconds and one minute (*t*(24)= 3.93, *p* < .001) with a large effect size of 1.08. The remaining tests did not reach statistical significance: five minutes and 30 seconds (*t*(29)= .06, *p* = .95), ten minutes and five minutes (*t*(73)= 0.4, *p* = .69) ten minutes and 30 seconds (*t*(24)= 0.22, *p* = .83). There is little evidence of an optimal planning condition in these results. However, there is a striking indication that the one-minute condition was the least beneficial planning condition. The measure value under the one-minute condition was considerably lower than the remaining conditions.

The impact of the planning variable on Task 3 (Balloon task; see Appendix 3) results is presented in Table 56. The ten-minute condition appears to be the optimal planning condition in terms of increasing scores with a measure value that was higher than the five-minute condition by .18 and the 30-second condition by 1.23. The ten-minute planning condition recorded a measure value that was 2.07 logits higher than the one-minute planning condition. This represents a major increase over the one-minute planning scores and indicates that the potential for pre-task planning to impact on test scores was particularly strong on Task 3.

Table 56 Results after different planning conditions: Task 3

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min.	- .87	1.33	$\chi^2 = 3, p < .001$
5 min.	-.69	1.17	
30 sec.	.36	.73	
1 min.	1.20	1.01	

The Welch's t-test results showed that the differences between the following scores were statistically significant: ten minutes and one minute (*t*(45)= 5.41, *p* < .001) with a large effect size of 1.55, ten minutes and 30 seconds (*t*(52)= 4.14, *p* < .001) with a large effect size of .95, five minutes and one minute (*t*(32)= 5.79, *p* < .001) with a large effect size of 1.45, and five minutes and 30 seconds (*t*(115)= 4.79, *p* < .001) with a large effect size of .88. The remaining results did not reach statistical significance: ten minutes and five minutes (*t*(50)= 0.62, *p* = .54), and one minute and 30 seconds (*t*(34)= 2.54, *p* = .02). The results indicate that planning for five minutes and ten minutes resulted in higher scores than were observed under the 30-second and one-minute planning conditions. This increase was particularly discernable between the ten-minute planning condition and the one-minute planning condition where the difference in logit values was 2.07.

Table 57 presents the results of the planning impact on Task 4 (Baby task; see Appendix 4). The ten-minute planning condition presented a clear advantage over the least beneficial planning condition, one minute, where the difference between the measure values was high at 3.76 in logits. The ten-minute condition also resulted in higher scores than the five-minute condition, a difference of 2.47 logits, and 30-

second condition, a difference of 2.17 logits. The results of the chi-square test indicated that the planning impact was statistically significant at $p < .001$.

Table 57 Results after different planning conditions: Task 4

Planning	Measure	Infit mean-square index	Fixed (all same) chi-square
10 min.	- 2.10	1.02	$\chi^2 = 3, p < .001$
30 sec.	.07	1.11	
5 min.	.37	.98	
1 min.	1.66	.70	

The results of the Welch's t-tests demonstrated that the following differences between scores were statistically significant: the ten-minute and five-minute conditions ($t(54)= 7.01, p < .001$) with a large effect size of 1.64, ten-minute and one-minute conditions (t(57)= 9.49, p < .001) with a large effect size of 2.49, ten-minute and 30-second conditions ($t(52)= 6.26, p < .001$) with a large effect size of 1.45, five-minute and one-minute conditions ($t(61) = 3.84, p < .001$) with a large effect size of .88, and one-minute and 30-second conditions ($t(59)= 4.82, p < .001$) with a large effect size of 1.08. The difference between the five-minute and 30-second conditions did not reach statistical significance ($t(107)= 1.09, p = .28$). These results clearly indicate that ten minutes was the optimal length of pre-task planning time on Task 4. Ten minutes resulted in sizeable increases in scores over all other planning conditions, most particularly over the one-minute planning condition, which resulted in remarkably low scores in relation to the ten-minute planning condition.

In order to compare the impact of pre-task planning between the four tasks, Table 58 reports the most extreme distance in logit measure values between the planning conditions and associated effect sizes. The table shows that the size of the planning impact varied between the tasks. Test takers were most likely to benefit from increased pre-task planning time when completing the picture-based narrative tasks. This is especially true for Task 4 where the ten-minute planning condition made a substantial difference to test taker scores in terms of logit measure values and effect sizes. Planning had less of an impact on the description tasks. The size of the impact was smallest on Task 2. Recall that on average, test takers received the highest scores on Task 2 and may have found this task relatively simple in comparison to the other three tasks. Ultimately, this degree of task simplicity may have reduced the potential for planning to impact performance. This is discussed at length in the Discussion (see Section 7.6).

Table 58 Largest differences between the planning conditions by task

	Difference in Logits	Cohen's *d*
Task 1	Ten minutes > 1.42 one minute	1.28
Task 2	Five minutes > 1.25 one minute	1.16
Task 3	Ten minutes > 2.07 one minute	1.55
Task 4	Ten minutes > 3.76 one minute	2.49

6.2.2.5 Summary of the findings

To summarise the results, when all facets were included in the analysis MFRM demonstrated that overall the highest scores were associated with the five-minute planning condition. However, Welch's t-tests demonstrated that the difference between the five-minute and ten-minute measure values was not statistically significant. The MFRM did not identify the planning condition that most substantially impacted on test scores. Statistically significant differences were observed between the following measure values: five minutes and one minute (a small to moderate effect size of .38), five minutes and 30 seconds (a moderate effect size of .40), ten minutes and one minute (a small effect size of .29), and ten minutes and 30 seconds (a small effect size of .30). On the logit scale, the largest difference between the planning conditions was .36 of a logit. This indicates that variation in pre-task planning time did not make a great difference to test scores. However, when the impact of variation in planning time on the individual test facets was investigated, the potential for pre-task planning to affect test scores became more apparent.

On the analytic scale, statistical significance was reached for the differences between the following measure values: the five-minute and one-minute, five-minute and 30-second, and ten-minute and one-minute planning conditions with small to moderate effect sizes (see section 6.2.2.1). Variation in pre-task planning time impacted the scores on the fluency category where statistical significance was reached for the difference in the following measure values: the five-minute and one-minute conditions, the five-minute and 30-second conditions, and ten-minute and 30-second conditions with small to moderate effect sizes (see Section 6.2.2.2). There was a statistically significant difference between the measure values of the ten-minute and 30-second conditions on the complexity category, with a small effect size (see Section 6.2.2.2). On the EBB scale, statistical significance was reached for the difference between the measure values of the five-minute and 30-second conditions, and the ten-minute and 30-second conditions with moderate effect sizes (see Section 6.2.2.1).

Regarding the interaction between proficiency and planning (see Section 6.2.2.3), the A level proficiency group (those that scored below 30 on the QPT; see Section 5.4.1) recorded a statistically significant difference between the measure values of the ten-minute and 30-second conditions, with a small to moderate effect size of .40. Statistical significance was not reached for the differences in measure values after planning at the B level of proficiency (those that scored above 30 on the QPT, see Section 5.4.1).

The analysis of the planning impact on the four tasks demonstrated that including extra planning time (five minutes and ten minutes) made more of a difference to the results of the picture-based narrative tasks than the non-picture-based description tasks. Overall, the increases in measure values after planning were largest on Task 4 and smallest on Task 2 (6.2.2.4). On Tasks 1, 3 and 4, the ten-minute planning condition resulted in the highest scores, whereas on Task 2 the highest scores were awarded under the five-minute condition.

6.3 CAF Results

This section reports the findings of the analysis of complexity, accuracy and fluency (CAF) in the main study. Firstly, the CAF measures are presented and the statistical procedures taken in the analysis are restated. Following this, descriptive statistics are presented in terms of planning time, proficiency level as assessed on the QPT, and task type. Finally, the results of the statistical analysis of the CAF results are reported to examine the interaction between planning time, proficiency and task type.

6.3.1 CAF measures

The test samples were transcribed by the researcher according to guidelines set out in Section 4.2.2.3. Following this, a second transcriber transcribed 10 test samples and coded for all CAF measures. Inter-coder reliability (TOTAL AGREEMENT/ n x 100) was 93.3 per cent. The following measures of CAF were included in the analysis (see Section 5.4.8):

Complexity

- Guiraud's Index (G.INDEX)
- Lexical sophistication assessed through the VocabProfile program (K1/K2/AWL/NONE)
- Clauses per AS-unit (C.AS)
- Idea units (IDEAS)

Accuracy

- Percentage of correctly supplied articles in obligatory contexts (ART)
- Percentage of correctly supplied prepositions in obligatory contexts (PREP)
- Percentage of correctly used tense (TENSE)
- Percentage of correctly supplied verbs in obligatory contexts (not omitted/ correct semantic usage/ including do and be and infinitive) (VERBS)
- Percentage of errors that are self corrected (SELF)
- Percentage of correctly supplied pronouns in obligatory contexts *
- Mean number of errors per AS unit (M.N.E)

Fluency

- Mean number of hesitations per AS-unit (MNH)

- Phonation Time Ratio (PTR)
- Percentage of hesitations that are filled (F.HES)
- Percentage of pauses that are filled (F.PA)
- Mean Length of Utterance (MLU)
- Total Speaking Time (TST)
- Speech Rate (SPR)

* excluded from further analysis (see Section 6.3.2)

6.3.2 Statistical approach

In order to investigate the interactions between pre-task planning time, task and proficiency in CAF results, statistical analytical procedures capable of calculating the effect of three independent variables upon the dependent variables were required. The MANOVA program was therefore selected for this purpose. A series of Shapiro-Wilk tests (p < .05) revealed that the variables were non-normally distributed. Tabachnick and Fidell (2007) explain that as long as non-normality is not due to outliers, a sample size of above n = 20 should ensure robustness for MANOVA analysis. The data was therefore searched for outliers. Inspection of boxplots and histograms revealed outliers in several variables. Outliers were located in K2, AWL, M.N.ER, M.N.H, T.S.T, SP.R, M.L.U, P.T.R, F.HES, SELF, VERBS, PRO, NONE, IDEA. The remaining variables (G.INDEX, K1, C.AS) were normally distributed or exhibited non-normal distribution due to skewness, which does not pose a problem for the MANOVA programme.

Two solutions are available for dealing with non-normal data for a parametric test, outlier removal or transformation (Tabachnick and Fidell, 2007). Transformation was attempted using the LN and Log10 command of SPSS (version 22). However, the outliers remained in the dataset. Transformation of the data did not create normality and so the outliers were removed. Outliers were removed for the following variables K2 (*n* = 3), AWL (*n* = 27), Tense (*n* = 7), Verbs (*n* = 7), Self (*n* = 14), M.N.ER (*n* = 4), M.N.H (*n* = 6), P.T.R (*n* = 8), F.H. (*n* = 1), M.L.U (*n* = 10), T.S.T (*n* = 13), SP.R (*n* = 4), IDEA (*n* = 4). Removal of outliers did not create normality in the percentage of correctly supplied pronouns. The majority of non-outlier participants used pronouns correctly; the range was from 0 to 100 per cent accuracy, however the mean value was 96.1 and was median 100. Pronouns were subsequently removed completely from the analysis.

6.3.3 Descriptive statistics

Distributions of CAF descriptive statistics are presented in Tables 64-70 (see Appendix 5). Table 64 presents the descriptive statistics by planning time. Tables 65 to 70 report the descriptive statistics by planning time on Tasks 1-4, and proficiency levels A and B.

6.3.4 Results of the statistical analysis

Wilks' Lambda is the most commonly reported MANOVA statistic in the literature and was thus used in the analysis (Huang, 2013). The MANOVA results are presented in Table 59. Statistical significance was reached for pre-task planning time

(time) (Wilks' λ= .239, p = .015) with a small to moderate effect size of .379, task (Wilks' λ= .203, p = .002) with a moderate effect size of .412 and proficiency (Wilks' λ= .229, p < .001) with a moderate effect size of .522. Significant interactions were found for time and task (Wilks' λ=. 020, p < .001) with a small to moderate effect size of .353. However, the remaining interactions (time and proficiency, task and proficiency, task and proficiency and time) did not reach statistical significance. In sum, the MANOVA results indicate that variation in planning time affected the CAF results. There was a statistically significant interaction between planning time and task but the interaction between planning and proficiency was not statistically significant. Accordingly, it can be summarised that the increases in planning time did not benefit one proficiency group more than the other in terms of CAF.

Table 59 MANOVA results: CAF

	Wilks' Lambda	F	Hypoth *df*	Error *df*	*p*	Effect size (η_p^2)
Within-participants effect						
Time	.239	1.568	60	152.991	.015*	.379
Task	.203	1.801	60	152.991	.002*	.412
Time and Task	.020	1.481	180	438.986	.001*	.353
Between-participants effect						
Proficiency	.229	2.781	40	102	.000*	.522
Time and Prof	.132	1.056	120	302.212	.351	.286
Task and Prof	.101	1.224	120	302.212	.086	.317
Time, Task and Prof	.031	1.138	200	480.783	.133	.294

* = significant at p < .05

In order to examine the impact of the planning conditions on the CAF results, further statistical tests were required. Individual analysis of the dependent variables was conducted with a one-way ANOVA with planning time as the independent variable. Additionally, a two-way ANOVA was run with time and task as the independent variables. A Bonferroni adjusted alpha level of .05/20 = .0025 was then set to determine statistical significance. The Bonferroni correction does ensure against type one error however this value is rather conservative and may produce conclusions that are susceptible to type two error. Nonetheless, the adjusted alpha level was deemed suitable to account for the number of tests that were completed (see Section 2.7.2.4). Results are presented in Table 60 and Table 61.

Table 60 One-way ANOVA results: time

Source	Measure	Sum of Squares	df	Mean square	F	Sig.
Time	SP.R	1904.756	3	634.919	2.929	.040
	G.INDEX	3.692	3	1.231	3.166	.030
	SELF	121.027	3	40.342	1.090	.359
	CAS	.273	3	.091	.891	.450
	K1	28.061	3	9.354	.282	.838
	K2	20.441	3	6.814	1.742	.166
	AWL	.654	3	.218	1.828	.150
	NONE	50.157	3	16.719	.514	.674
	ART	523.618	3	174.539	.222	.881
	PREP	923.305	3	307.768	.475	.700
	TENSE	257.693	3	85.898	1.355	.264
	VERBS	302.383	3	100.794	.448	.720
	M.N.ER	.107	3	.036	.210	.889
	P.T.R	730.649	3	243.550	3.308	.025
	M.N.H	4.122	3	1.374	3.154	.030
	F.HES	1388.362	3	462.787	1.859	.144
	F.PA	1832.860	3	610.953	1.393	.252
	T.S.T	4888.069	3	1629.356	3.045	.034
	M.L.U	7.455	3	2.485	2.681	.053
	IDEA	121.515	3	40.505	6.722	.000*

* = significant at $p = .0025$

The one-way ANOVA results are first discussed. The ANOVA results revealed a statistically significant impact of the pre-task planning variable on the overall number of ideas produced (F=6.722, $p < .001$). Even though the order of the tasks and planning conditions were counterbalanced between the test takers, it was important to establish that this result was due to variation in planning time rather than a practice effect. A one-way ANOVA was run with task order (i.e. the first task that each test taker completed, the second, third and fourth) as the independent variable and the number of idea units as the dependent variable. The results showed that task order did not have a statistically significant impact on the number of idea units (F=1.159, p = .327). Figure 17 demonstrates that the five-minute planning condition generated the most idea units (10.58) and the lowest number of idea units was produced under the one-minute condition (7.82). The number of idea units produced under the one-minute condition was considerably less than the number under the 30-second condition (9.45). The ten-minute planning condition did result in a higher number of idea units (10.14) but this value was slightly lower than the five-minute condition. A Tukey post hoc test indicated that the differences in the number of idea units produced under the one-minute and five-minute planning conditions was statistically significant (p = .028). However, the remaining differences in the number of idea units produced did not reach statistical significance. Increasing planning time from one minute to five minutes therefore increased the number of idea units produced. Based on these results, planning permitted test takers to generate more content to complete the task.

Figure 17. Idea unit values by planning conditions

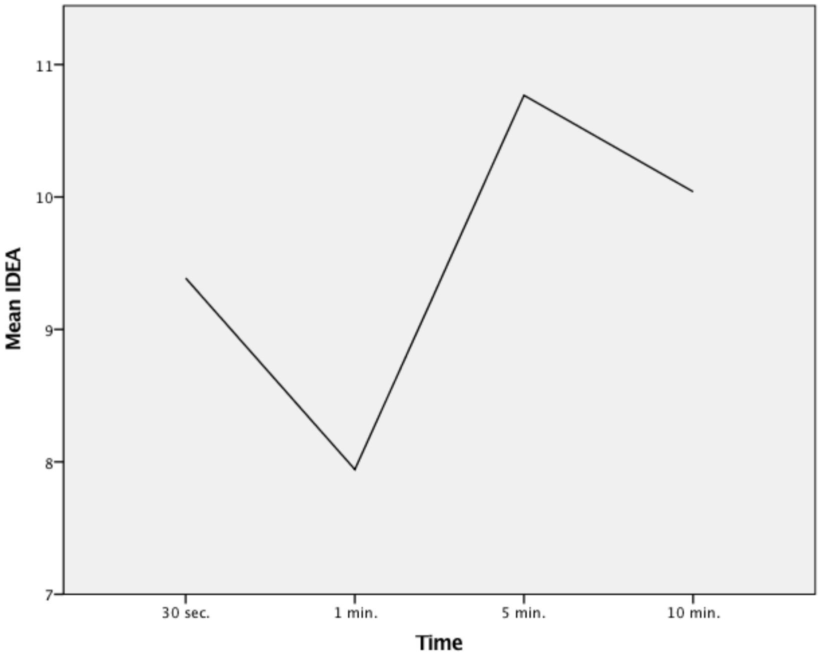


Table 60 demonstrates that the impact of planning on the remaining CAF measures did not reach statistical significance. This result is surprising given the effects recorded on equivalent measures during piloting (see Sections 4.2.6 and 4.3.5). It may be the case that the lack of impact was due to the adjusted alpha level, which is conservative at p = .0025. Speech rate, Guiraud's Index, phonation time ratio, mean number of hesitations and total speaking time did meet a statistical significance level of p = .05 that is commonly adopted in the literature (see Section 2.7.2.4). However, the adjusted alpha level was applied to ensure that the interpretation of the results was not susceptible to type one error. In sum, results of the one-way ANOVA indicated that the pre-task planning variable permitted test takers to generate extra task content but this did not impact their delivery of the content in terms of CAF.

The results of the two-way ANOVA are reported in Table 61. The two-way ANOVA was conducted to compare the effect of different amounts of planning time on CAF results on the four tasks. Results demonstrated statistically significant differences in the percentage of K2 vocabulary between the four tasks when completed with different amounts of planning time (*F*=7.267, $p < .001$). Jaeger (2007) identifies problems with the use of ANOVA to test differences between percentages (i.e. in this study the percentage of vocabulary that was K2). Specifically, with a 95% confidence interval, percentage values may ‘exceed beyond interpretable values of 0 to 100’, leading to ‘spurious results’ (2007, p. 435). To account for this potential limitation, multiple regression was also used to predict the amount of K2 vocabulary attributable to planning and task conditions. The independent variables (planning condition and task number) statistically significantly predicted K2 vocabulary use, *F*(2,182)=13.097, $p = .00$, $R^2 = .126$. Both variables added to the prediction at $p = .05$ indicating that variation in planning time and task number affected the test takers’ use of K2 vocabulary, although the impact was minor: different planning and task conditions accounted for 12.6 per cent of the total variation in K2 vocabulary use. This means that planning time contributed to variation in the amount of K2 vocabulary test takers used although the size of the contribution was minimal. The K2 results are described in detail later in this section.

In addition to the proportion of K2 vocabulary, the number of idea units produced on each task under the four planning conditions approached statistical significance (*F*=3.095, $p = .003$) but did not reach the adjusted alpha level. Nonetheless, given the level of statistical conservatism established by the Bonferroni correction, the results will be discussed at length because they may prove informative for future research.

Table 61 Two-way ANOVA results: time and task

Source	Measure	Sum of Squares	df	Mean square	F	Sig.
Time x Task	SP.R	1395.955	9	155.106	.725	.684
	G.INDEX	3.465	9	.385	.990	.456
	SELF	310.927	9	34.547	.933	.502
	CAS	.531	9	.059	.578	.811
	K1	462.007	9	51.334	1.549	.148
	K2	255.798	9	28.422	7.267	.000*
	AWL	1.724	9	.192	1.605	.131
	NONE	253.783	9	28.198	.866	.559
	ART	3845.251	9	427.250	.544	.838
	PREP	6151.953	9	683.550	1.056	.406
	TENSE	641.565	9	71.285	1.125	.357
	VERBS	2222.467	9	246.941	1.097	.377
	M.N.ER	1.522	9	.169	.998	.450
	P.T.R	754.106	9	83.790	1.138	.348
	M.N.H	3.364	9	.374	.858	.566
	F.HES	3961.928	9	440.214	1.768	.090
	F.PA	2287.508	9	254.168	.579	.810
	T.S.T	5896.563	9	655.174	1.225	.294
	M.L.U	10.160	9	1.129	1.218	.298
	IDEA	167.856	9	18.651	3.095	.003

Figure 18 presents the differences in K2 vocabulary between the planning conditions on the four tasks. The presence of K2 vocabulary in the transcript indicates that the test taker was able to use less frequent words to complete the task (see Section 2.7.2.1). Task 1 elicited the highest percentage of K2 vocabulary under the one-minute planning condition (8.18) and the lowest percentage occurred under the ten-minute planning condition (2.59). The mean difference between these values is substantial and indicates that test takers used considerably more K2 vocabulary without extra planning. In contrast, Task 2 elicited similar levels of K2 vocabulary after the one-minute (5.37) and ten-minute (5.28) conditions, whereas the lowest amount was produced under the five-minute condition (.47). The value under the five-minute planning condition was considerably lower than the values under the 30-second, one-minute and ten-minute planning conditions. The impact of planning time therefore varied markedly with regards to the lexical sophistication that was produced on the description tasks. Concerning the narrative tasks, Task 3 involved the highest amount of K2 vocabulary under the five-minute planning condition (7.66), whereas on Task 4 it was the 30-second condition (6.77). In the case of Task 4, the least amount of planning time generated the highest percentages of K2 vocabulary. Both narrative tasks featured the lowest level of K2 vocabulary under the one-minute planning condition (Task 3= 4.95, Task 4= 2.97), which supports the overall impression gleaned from the one-way ANOVA that the one-minute condition resulted in the weakest performance. Both narrative tasks recorded similar levels of K2 vocabulary after the ten-minute condition (Task 3= 6.51, Task 4= 6.38). On this evidence, the pre-task planning effect seems to be more consistent on the picture-based narrative tasks than the non-picture-based description tasks with regards to lexical sophistication.

Figure 18. Percentage of vocabulary that is K2 by time and task

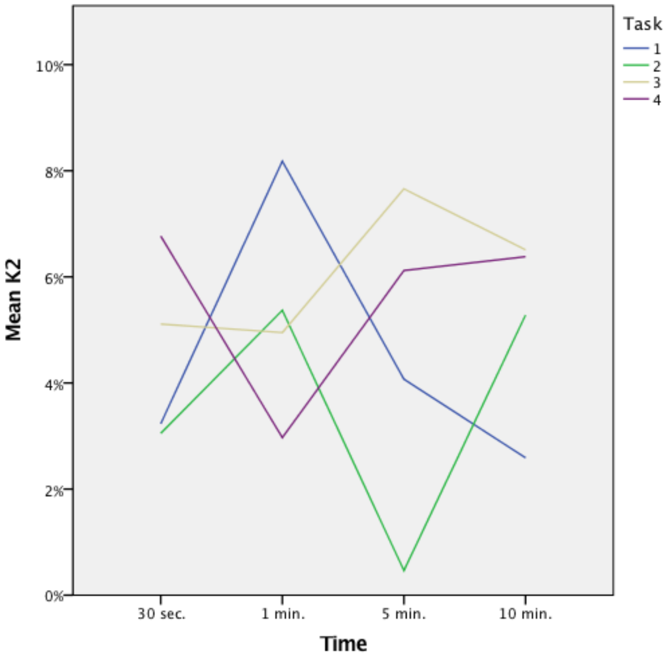
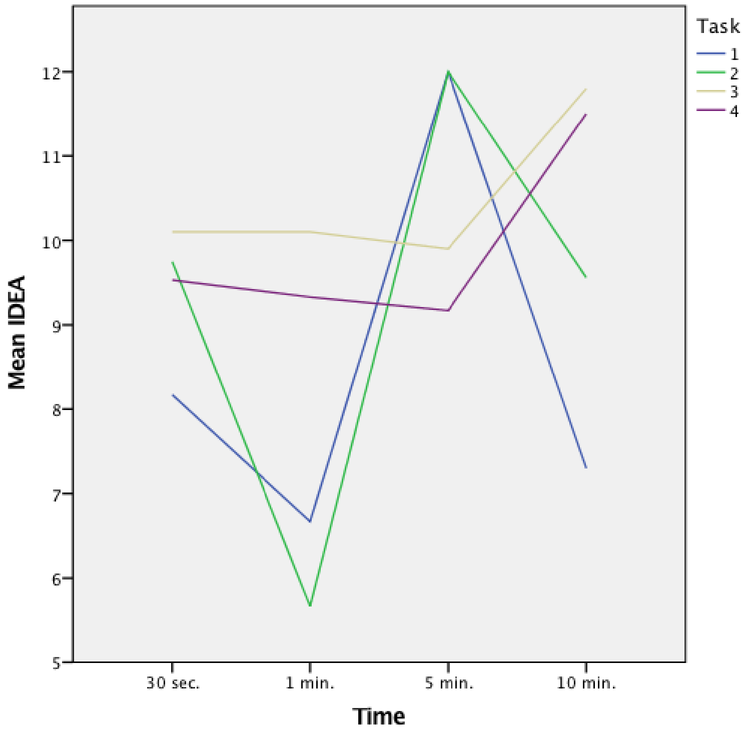


Figure 19 presents the difference in the number of idea units between planning conditions on the tasks. Beginning with the non-picture-based description tasks, Tasks 1 and 2 involved the highest number of idea units under the five-minute planning condition (1= 12, 2= 12) and the lowest under the one-minute planning condition (1= 6.67, 2= 5.67). This sense of symmetry between the two description tasks contrasts with the results of the K2 vocabulary analysis in which the planning impact varied considerably between Tasks 1 and 2. However, whereas Task 1 involved few idea units under the ten-minute planning condition (7.30), the corresponding figure on Task 2 was relatively high (9.56). Furthermore, Task 2 involved substantially more idea units under the 30-second planning condition (9.75) than Task 1 (8.17). These results indicate that when completing the description tasks, test takers tended to produce the most content after planning for five minutes.

Turning to the picture-based narrative tasks, the various planning conditions appear to have made little difference to the number of idea units produced with the exception of the ten-minute planning condition, which recorded markedly higher numbers than the other conditions (Task 3= 11.83, Task 4= 11.50). Furthermore, the five-minute planning condition recorded the lowest number of idea units on each narrative task (Task 3= 9.90, Task 4= 9.17). The test takers only seem to have been able to produce more content on the picture-based narrative tasks when ten minutes planning time was available. Increases in idea units on the picture-based narrative tasks may only be possible when test takers have a substantial amount of time to plan their speech. To compare the two task types, the pattern of distribution between the planning variables was more uniform on the picture-based narrative tasks than the non-picture-based description tasks indicating that test takers responded more consistently to differences in planning time when completing picture-based narrative tasks.

Figure 19. Idea unit values by time and task



When completing the non-picture-based description tasks, test takers produced the largest numbers of idea units under the five-minute planning condition but less K2 vocabulary. Conversely, the K2 vocabulary values peaked at one minute on the description tasks, whereas this planning condition generated the lowest number of idea units. In comparison, the picture-based narrative tasks tended to involve more idea units under the ten-minute planning condition, whereas the five-minute planning condition resulted in high percentages of K2 vocabulary on Task 3. On Task 4, the percentages of K2 vocabulary were relatively similar under the five and ten-minute planning conditions.

6.3.5 Summary of the findings

Variation in pre-task planning time made little difference to the results of the CAF measures. However, a statistically significant result was observed in the amount of idea units produced; the five-minute planning condition facilitated the generation and production of more content to complete the tasks in relation to the one-minute condition. Variation in planning time did not impact levels of grammatical accuracy, grammatical complexity, lexical variety or fluency. This result contrasts with the results that were observed during piloting, which demonstrated gains in fluency after both five minutes and ten minutes planning time. The lack of a clear impact on the fluency measures in this study may be due to the conservative significance level. Results of the one-way ANOVA demonstrate that an alpha level of $p < .05$ was reached in speech rate, phonation-time ratio, total speaking time and mean number of hesitations.

In terms of the interactions between the independent variables, interaction was observed between time and task. Extra planning time (five minutes and ten minutes) led to increases in lexical sophistication (K2) on Task 3 and to some extent on Task 4. In terms of idea units, the ten-minute condition caused the most substantial increases in the picture-based narrative tasks. On the non-picture-based description tasks, the largest increases were observed under the five-minute planning condition. However, the overall impact of variation in planning time was not consistent and extra time led to decreases in lexical sophistication and the number of idea units especially on the description tasks. For instance, on Task 1 the largest amount of K2 vocabulary was produced under the one-minute condition. The implications of these results are discussed at length in the Discussion.

The analysis did not uncover a statistically significant interaction between pre-task planning and proficiency. This result suggests that proficiency was not a factor that contributed to the impact of planning on the CAF measures. Furthermore, the tests conducted to identify an interaction between pre-task planning, task and proficiency did not reach statistical significance.

7 Discussion

7.1 Introduction

This chapter discusses the research findings reported in Chapter 6. It separates the discussion in to six sections corresponding to the research questions, which are restated at the beginning of each section. Each section begins with an overview and interpretation of the quantitative research findings and examines transcript samples that provide insights into these findings. The transcript samples were selected to be representative of task performance in the test taking population and illustrate how task completion varied when the planning condition was manipulated. Following this, the discussion interprets these findings with reference to the literature review.

Section 7.2 discusses the overall impact of variation in planning time on the test scores and the complexity, accuracy and fluency (CAF) results. Section 7.3 identifies the planning condition that had the largest impact on the test scores and CAF results. Following this, Section 7.4 compares planning effects on the EBB scale scores and the analytic scale scores. Section 7.5 discusses the impact of variation in planning time and test taker proficiency on test scores and CAF results. Section 7.6 discusses the interaction between task type and planning time and the impact of this interaction on CAF results and test scores. Section 7.7 describes the impact of variation in planning time on test taker performance on Task 4.

7.2 Does variation in planning time operationalized as 30 seconds, one minute, five minutes and ten minutes impact the results of a language test when assessed with a)

an EBB scale b) an analytic scale c) measures of complexity, accuracy, and fluency (CAF)?

To restate the results of the multi-faceted Rasch measurement (MFRM), variation in planning time did have an impact on test scores (see Section 6.2.1). However, the impact was minor. The logit scale contained five logit levels (ranging from -3 to 2) and the maximum difference between the planning conditions was .36 of a logit on the scale (see Table 40). This difference was between scores that were awarded under the five-minute planning condition and scores awarded under the one-minute, and 30-second planning conditions. In both cases the effect size was small to moderate: .38 and .40 respectively (Cohen, 1988). In sum, the overall picture that the MFRM provided was that the addition of extra planning time before the tasks marginally increased test taker scores.

The results of the CAF analysis showed that the planning time variable had a limited impact on the test takers' performance (see Section 6.3.4). The one-way ANOVA demonstrated that the only result that reached statistical significance was the total number of idea units produced (Table 60). This result showed that planning for extra time (five minutes and ten minutes) was associated with high numbers of idea units. In addition to this, the two-way ANOVA demonstrated that there was a statistically significant interaction between the amount of K2 vocabulary produced, task number and planning time (Table 61). This result is discussed in detail in Section 7.6. The increase in the number of idea units indicates that extra pre-task planning time had more of an effect on task content than the language forms used to express this content. In other words, the test takers generated more ideas after extra planning time but this made little difference to the complexity, accuracy or fluency of their speech. To exemplify this, examples of test transcripts are presented in sample TT1a (produced by test taker 1 on Task 1 '*Tell me about something interesting you have recently heard in the news*' after planning for one minute) and sample TT1b (produced by test taker 1 on Task 2 '*Tell me about an event that has changed your life*' after planning for ten minutes). AS units are separated with vertical bars, two colons represent a new clause within an AS unit, and unfilled pauses and hesitations are indicated by the number of seconds in parenthesis.

Sample TT1a, Task one, one minute:

I live in karsiyaka | er and we have a famous bazaar in karsiyaka | (1) we all use our mobile phones | this is very er useful for us | (1) and sometimes we have to charge our phone or | er we need to have wifi connection | (1) in karsiyaka bazaar we have wifi connection po er points | (1.5) (if you have) er (2) if you have to do a research for example on the internet :: you can go to this point :: and have the wifi connection | and make your research

Sample TT1b, Task two, ten minutes:

I have a problem (with my legs) about my legs | (I have a) I have a balance problem | (1) er I couldn't walk like other people's walking style | (1.5) er when I was in the primary school :: (1) my friends asked to me :: what happened :: er what's your problem :: and I felt upset | (1.5) er but I realised :: that something is different (1) about me | and I asked (1) er to my family my story | (1) and they explained me :: (1.5) this is a disease :: (1) called cerebral palsy | (1.5) er this is about brain and muscle connections | (1) when I was a baby :: I couldn't breathe (n) for a minute | (1) and er I couldn't have enough oxygen | (1.5) so (my) er the right side of my brain (1.5) had bad effects | (1) and this effects er (1.5) provide (1.5) my er (1) left leg | (1) my left leg is weak | (2) the other leg is er more powerful than (3.5) er (1) | and when I er (2) (her) learn this story :: I researched it (1) | and I realised :: that this is an important disease :: but I was very lucky (1.5) | er since I learnt d my story :: (I am) er (2) I have more self confidence :: and I'm more social

Comparing TT1a and TT1b, it is clear that the test taker produced more idea units under the ten-minute planning condition than the one-minute condition. According to the guidelines for identifying idea units set out in Luoma (2004), and

Frost, Elder and Wigglesworth (2011), sample TT1a contains four idea units, whereas sample TT1b contains nine. The test taker doubled the number of ideas she produced after taking part in extra planning, which generated a richer description of the task content.

Overall, the analysis showed that the five-minute planning condition generated the highest average number of idea units (10.58). This was followed by the ten-minute planning condition, which generated a value of 10.14. Under the 30-second planning condition the average number of idea units was 9.45, and 7.82 under the one-minute condition. Samples TT27a and TT27b provide examples of the difference between the number of idea units produced under the 30-second planning condition and the five-minute planning condition. Sample TT27a is the transcript of test taker 27 completing Task 1 under the 30-second planning condition, and sample TT27b is the transcript of test taker 27 completing Task 2 under the five-minute planning condition.

Sample TT27a, Task 1, 30 seconds:

It's not new :: but I heard explosion bomb in Ankara | I think :: it's very bad (1.5) for Turkish peoples :: because lots of people die :: and lots of family (is) are very er sad :: because they lose our (1) er religion | (1.5) and (1) our economy (Turkish) p Turkish (1) er economy is go down :: because this explosion is very bad

Sample TT27b, Task 2, five minutes:

when I was in (high) er primary school :: (1) er I started play handball | it's kind of spor :: you play your hands | (1) er and er I (1.5) er went lots of away match | (we stayed) er it's team sport you play team :: and er you stay (1) generally one week in the away | and er (1) it's very good for my life :: because er I stayed (1) er alone | I go (1) away and my parents (not) er aren't near me | (1) I think :: it's good experience for my life :: because I learn :: how can I live alone :: or what can I do :: I am in the alone

In sample TT27a, the test taker produced three idea units, whereas in sample TT27b the total number of idea units was seven. After the extra planning time, the test taker elaborated on the subject of handball (it's a sport, you play it using your hands), and how the experience of travelling to play handball with a team has benefitted his life (becoming independent from his parents, learning to live alone). In contrast, in sample TT27a the test taker stated that a bomb exploded in Ankara and it affected society and the economy negatively. This comparison reflects the common finding that more elaboration was provided on the task content after the test taker had planned the speech, which may have had some bearing on the test takers' rating scale scores. Ellis and Barkhuizen (2005, p. 154) suggested that the number of idea units is an indication of the 'propositional completeness' of a text. Relating this to the rater scores, the implication of this finding is that greater propositional completeness may have caused raters to assign modestly higher scores to the test samples. Therefore the raters' perception of second language proficiency may be linked to the test takers' ability to generate ideas. This finding is discussed in more detail in Sections 7.6, 7.7, and 7.8.

In the literature review (see Section 2.7), it was suggested that the conflicting accounts of the impact of planning in the fields of language testing and task-based language teaching (TBLT) were due to variation between the fields in the respective approaches to measurement. TBLT studies invariably assess planning effects with measures of CAF and report positive results, whereas studies with a language testing focus typically use a rating scale and the impact is less consistent. This trend indicated that increases in planning time in this study would impact the CAF measures more than the test scores on the rating scales. Contrary to expectations, the opposite proved true in this study. The increases in measures of complexity, accuracy and fluency that have been widely reported in the TBLT literature, and less frequently in the language testing literature (see Section 2.2) were not present in the results. Rather, planning had very minor impacts on the overall test scores (i.e. when all facets were included in the MFRM), which may be attributed to the rise in the number of idea units (the interaction between the rating scale and the increase in idea units is discussed in detail in Section 7.4). This was an unexpected result that indicates there were important differences between this study and those discussed in the literature review.

The clear difference between the findings of this study and the trend of results reported in the literature needs to be accounted for. Shortcomings in the analyses employed in the TBLT and language testing studies may account for much of the disparity. A conservative approach to interpret the significance of CAF results, using Bonferroni correction, was adopted for this study. This resulted in a critical alpha value of p = .0025. In contrast, many of the TBLT and language testing studies (see Section 2.2 and Section 2.7.2.4) do not account for the increased chances of committing a type one error when completing multiple statistical tests on CAF results. For instance, the reported increases in CAF results after planning in Ellis (1987), Crookes (1989) and Sasayami and Izumi (2012) do not reach statistical significance using an adjusted alpha value. In addition, using an Bonferroni adjusted alpha value to interpret the results reported in Foster and Skehan (1996), an influential study in the second language speech planning literature, the only results that reach statistical significance are the number of pauses, and the amount of task silence (see Section 2.7.2.4). In the current study, with an unadjusted alpha value of p = .05, the results of Guiraud’s Index, speech rate, phonation-time ratio, mean number of hesitations, and total speaking time would have been statistically significant. The interpretation of the findings would have been very different. Without the adjusted alpha level, the results of this study would correspond very closely to the positive accounts of planning made in the literature.

Based upon empirical research findings (e.g. Foster and Skehan, 1996, Yuan and Ellis, 2003), task-based learning researchers (Robinson, 2005, Skehan, 2009) emphasize that pre-task planning has an important influence on language learners' ability to produce L2 speech. However, the findings of this study suggest that the overall impression of second language ability that this particular test generates is not affected by variation in pre-task planning time. The quality of language production does not significantly vary according to differences in planning time (in terms of CAF) but planning permits test takers to increase the number of ideas they express during the task and this has a minor impact on test scores. In light of Fulcher's (2003, p. 64) argument that 'gross changes' need to be made to the task in order to affect test scores, these findings suggest that pre-task planning does not have an important impact on the assessment of test performance. However, this interpretation is discussed further in Section 7.6, and 7.7.

In contrast to the TBLT literature, there is broad overlap in the research findings of this study and the language testing literature where planning either has a limited impact on the test scores (e.g. Nitta and Nakatsuhara, 2014, Weir et al., 2006, Xi 2005) or does not impact the test scores at all (Elder et al., 2002, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006, Iwashita et al., 2001).

Language testing researchers have proposed that the ability to recall information generated during planning is constrained by limitations in working memory capacity. Elder and Wigglesworth (2006) argue that test takers cannot sustain the levels of complexity, accuracy and fluency that planning may foster throughout an entire test. The effect of planning diminishes as the task progresses and the test taker becomes increasingly reliant on online planning. As raters consider the entire test performance when assigning grades, the potential for planning to impact test scores is minimal. Future research will be required to confirm this because this study did not uncover an impact of pre-task planning on CAF measures. However, the way that speech planning affects the first few utterances of the task performance is discussed in detail in Section 7.7.

Elder and Wigglesworth (2006, p. 21) recognise that the planning variable in their study may have increased the ‘propositional complexity of the discourse’ but question the relevance of this to the results of their test on the grounds that it has little bearing on second language knowledge and proficiency. However, the current study demonstrates a clear relationship between planning and an increase in the total number of idea units, which in the absence of statistically significant effects of the planning variable on measures of CAF, may explain the increases in raters’ scores after planning time was increased. Therefore, propositional complexity may be an important factor in raters’ assessment of language proficiency. For a detailed discussion of this factor, see Sections 7.6, 7.7, and 7.8.

A potential cause for the relatively limited impact of planning on test scores that this study has not examined is that the participants’ cognitive orientation to the task may vary substantially between the high-stakes, language testing context and the low-stakes, pedagogical classroom context (Ellis, 2005). The high-stakes context promotes a focus on form because test takers are keen to avoid the penalties involved in making mistakes. This focus on form is absent in classroom environments where the participants are free to engage in consequence-free experimentation with language forms that are not well rehearsed. The results of the pilot studies (see Sections 4.2 and 4.3) indicated that variation in planning time would have an impact on the main study results. As a result, measures to explore Ellis' hypothesis were not taken in the main study. However, the findings reported in the main study indicate that planning did not affect speech performance a great deal and that Ellis' hypothesis may be credible. Future research that compares planned speech performance elicited in a language classroom with planned speech performance on a language test may serve to shed further light on the relationship between context and pre-task planning. Areas for future research are mapped out in the conclusion (see Section 8.5).

7.3 Which amount of planning time (30 seconds, one minute, five minutes, ten minutes) most substantially impacts test scores and CAF results?

The results of the MFRM indicate that the five-minute planning condition led to the highest scores on the logit scale with a measure value of -.20. This was followed by the ten-minute condition with a measure value of -.12. The difference between planning for 30 seconds and planning for one minute did not affect the scores on the logit scale; both planning conditions resulted in measure values of .16. The results of the Welch's t-tests showed that the differences between scores awarded under the five-minute and one-minute planning conditions, and secondly the five-minute and

30-second planning conditions were statistically significant with small to moderate effect sizes of .38 and .40 respectively (Cohen, 1988). Furthermore, the differences between scores awarded under the ten-minute and one-minute planning conditions, and also the ten-minute and 30-second planning conditions were statistically significant with small effect sizes of .29 and .30 respectively. These results indicate that variation in the amount of planning time had a small to moderate effect (Cohen, 1988) on the test scores. The largest effect size was observed between the test scores awarded under the five-minute planning condition and test scores awarded under the 30-second planning condition. In sum, the five-minute planning condition resulted in the highest test scores when all test facets were included in the MFRM (test takers, raters, tasks, proficiency groups). Even so, the difference between scores awarded under the five-minute and ten-minute planning conditions did not reach statistical significance. Based on these results, it is not possible to categorically conclude that the five-minute condition was the most advantageous planning condition.

To restate the CAF results, variation in planning time increased the number of idea units. The highest number of idea units was recorded under the five-minute planning condition (10.58), which was a marginal increase over the number of idea units produced under the ten-minute planning condition (10.14), and more substantially over the 30-second (9.45) and one-minute (7.82) planning conditions. The value of 7.82 under the one-minute planning condition was particularly low and indicated that planning for one minute in contrast to 30 seconds was detrimental to the development of ideas. Test takers produced more idea units when performing the tasks relatively spontaneously (i.e. after 30 seconds).

It is possible that during the one-minute planning condition, test takers began to generate plans for the task, which they had to abandon abruptly when the planning time ended. This interpretation is based on Skehan's (2009) model, which states that planning may serve to complexify the process of speech production by causing the test taker to attempt to a) generate more ideas and b) increase the complexity of these ideas (see Section 2.5). Test takers may have underestimated how quickly the one-minute planning condition would pass and started to develop plans that they could not complete before the task began. This means that they were suddenly reliant on online planning to complete the task. In contrast, in the knowledge that they would not have time to plan a substantial amount of information (i.e. in 30 seconds), test takers may have prepared themselves to begin speaking almost immediately by quickly scanning the images (Tasks 3 and 4) or recalling simple information with which they were very familiar (Tasks 1 and 2). Future research involving stimulated recall methodology might help to resolve this issue by comparing test taker reports under different planning conditions (Sangarun, 2005). This is discussed in detail in the areas for future research section of the conclusion (see Section 8.5).

In much of the language testing literature, planning time is restricted to one minute (Elder and Wigglesworth, 2006 Weir et al., 2006, Wigglesworth, 1997, Xi, 2005, 2010). The similarity in scores awarded under the 30-second and one-minute planning conditions in this study indicates that the amount of planning time may need to be increased beyond one minute to have an impact on the test scores. For instance, in Elder and Wigglesworth (2006) no statistically significant difference was observed between scores awarded under one-minute and two-minute pre-task planning conditions. Neither did the variation in planning time affect CAF results. The researchers used non-picture-based description tasks (similar to Tasks 1 and 2 in the current study). In light of the findings of the current study, increasing planning time to five minutes may have led to different results.

The finding that the five-minute planning condition was associated with the highest scores was unexpected. The TBLT research consistently demonstrates that a period of ten minutes pre-task planning has a positive impact on the results of CAF measures (see Section 2.2). This has created the belief among some that ten minutes may be the optimal amount of planning time. Li et al. (2014) suggest that greater increases in speech complexity may have been observed in their study if they had provided their participants with the opportunity to plan for ten minutes. The authors argue that the inclusion of a ten-minute planning condition may encourage test takers to consider complex syntactic structures in detail, building confidence and reducing the need for speech monitoring during the task. Attentional resources would thus be available for the production of complex language. However, the results of the current study contradict these predictions because increases in syntactic complexity (i.e. the mean number of clauses per AS unit) were not observed under the ten-minute planning condition. It may be the case that the addition of pre-task planning time is insufficient to increase syntactic complexity and that L2 learners need to be trained to produce subordinate clauses during the task (Mochizuki and Ortega, 2008). Such training would be inappropriate under assessment conditions but would be suitable as a classroom activity (e.g. for test preparation purposes).

Elder and Iwashita (2005) and Iwashita et al. (2001) also suggest that their results may have been different if they had included more planning time. Iwashita et al. (2001) propose that planning affects speech production most clearly when test takers are provided with long periods of time to plan (e.g. ten minutes) for tasks that are complicated and elaborate. This is because the more challenging the task, the more benefits can be derived from planning (see Section 2.5.1.4). The findings of the current study support this claim. In the current study, results of the MFRM show that test takers received the lowest overall scores on the picture-based narrative tasks (i.e. picture-based narratives were the most challenging) and also benefitted most from the ten-minute planning condition when completing these tasks (see Section 6.2.1). When complicated tasks (picture-based tasks: Task 3 and 4) and relatively simple tasks (non-picture-based tasks: Task 1 and 2) were included in the MFRM, the largest gains in scores were observed under the five-minute planning condition. This indicates that the complexity of the task that test takers completed had an important bearing on the impact of the planning variable (see Sections 7.6 and 7.7). Scores on complicated tasks were most impacted after test takers planned for ten minutes, whereas scores on the less demanding task (i.e. Task 2) were most impacted after test takers planned for five minutes.

To hypothesise about the reason for this, when test takers complete a relatively simple task (i.e. describing an event that was important in their lives: Task 2), there is little need to engage in extensive planning. The task information may be well rehearsed and the test taker may opt to describe an event that requires language with which they are relatively familiar. Including extensive periods of planning (i.e. ten minutes) for such tasks may be unnecessary and even hinder test taker performance. In the current study, test takers may have lost concentration and motivation during the ten-minute planning condition for Task 2 and this may have had a negative impact on their ability to complete the task (Field, 2011). Further research will be required to confirm this (see Section 8.5).

7.4 Does the impact of the four planning conditions on test scores vary between the analytic scale and the EBB scale?

7.4.1 EBB scale

Wigglesworth (1997) discusses the possibility that variation in planning time did not impact the results of her study due to mismatches between the rating scale content, the changes that planning instigated, and the raters' own internal criteria (see Section 2.7.3). This study attempted to resolve this issue with the use of an assessor-oriented, EBB scale (see Section 2.7.3.3). The MFRM of the EBB scale scores demonstrated that the five-minute planning condition resulted in the highest overall (i.e. accounting for the entire participant sample on all four tasks) fair average value (3.08), followed by the ten-minute (3.06), one-minute (2.92), and 30-second (2.77) planning conditions. The results of the Welch's t-tests demonstrated that the differences between the scores awarded under the five-minute planning condition and the scores awarded under the 30-second planning condition were statistically significant with a moderate effect size of .46. Furthermore, the difference between scores obtained under the ten-minute planning condition and those obtained under the 30-second planning condition were statistically significant with a moderate effect size of .43. Importantly, the differences between the scores awarded under the five-minute and ten-minute conditions did not reach statistical significance.

At first glance, these results indicate that planning did not have a substantial impact on test scores. This may be due to the binary nature of the EBB scale. When raters were required to make binary, holistic distinctions, rather than identify a level of competency in one category on the analytic scale, planning may have had little effect on the decision. For example, using the analytic scale a score of 2 on the fluency category indicates that a test taker was less fluent than one who attained a score of 3. In contrast, on the EBB scale, the distinction between a score of 2 (i.e. the fair average value under the 30-second planning condition, and the one-minute planning condition) and a score of 3 (i.e. the fair average value under the five-minute planning condition, and the ten-minute planning condition) involves a more holistic decision about whether the task had been completed successfully. Specifically, at level 3, test samples exhibit 'satisfactory task completion with minimal strain for the listener'. At level 2, task completion is unsatisfactory and causes the rater strain. This is a major difference and the impact of planning needed to be substantial to affect raters' decisions on this aspect of the EBB scale.

The results of an analysis of the individual test takers' scores showed that variation in planning time increased five of the test takers scores from level 2 to level 3 on the EBB scale (see Section 6.2.2.1, Table 47). For these test takers, planning for extra time enabled them to complete the task successfully and played a major role in their test scores. In addition, the analysis showed that overall 12 test takers were placed in to different levels on the EBB scale under different planning conditions. For examples of the differences in performance after extra planning time, samples TT29a and TT29b are the transcripts of test taker 29 under the 30-second planning condition on Task 3 (level 2 on the EBB scale) and under the five-minute planning condition on

Task 4 (level 3 on the EBB scale). Turkish words are italicized and the English translation is provided in brackets.

Sample TT29a, Task 3, 30 seconds:

one woman is washing dishes (1) er in (your) her house | (1) and after (she) m *Sey* [thing] (1) (she) er (her) she washing her clothes | and (1.5) I forgot *Sey* [thing] m this mean | er she washing her own clothes | (1.5) after er (ch childrens see) children saw balloons | and er (they) they want to buy | and they (1) buying one balloon | (1) after children (1) take clothes (for rope) on the rope | (1.5) and (2.5) and er other children er painting (1.5) (on) on the (1) maybe ball maybe balloon ha balloon | other childrens (painting face) drawing face on the balloon | and (last) last picture is :: (1) they er (1) scared for they mothers | and woman is (1) looks like scared | (2) and (1) ha cat is running last (1) this

Sample TT29b, Task 4, five minutes:

in first photo woman (1) reading a book :: and er (baby s) baby is sleeping on the (ba on the) baby's (bed s) bed | and after (1) er woman is taking a nap | (1) and baby still sleeping | and m children two children little one girl and boy (see the) see the la maybe mother | (1) er look and mother is sleeping | (and) er (and they) and then little girl is taking baby on the baby's bed | (1.5) and after er boy is (1) put maybe ball maybe balloon (1) put the baby's bed | and (1.5) er little girl (1) er take the baby | and after they er little girl boy and baby hiding (1) er behind the mother | and mother is scared :: because of er mother she look baby's bed :: and baby is not here :: (2) (there is a b) there is a ball | and (she scared) she was scared (2) and finished

In sample TT29a, the test taker struggled to generate lexis to describe the scene and openly told the examiner that she had forgotten a word (*after (she) m sey [thing] (1) (she) er (her) she washing her clothes | and (1.5) I forgot sey [thing] m this mean*). However in sample TT29b, the same test taker did not experience this kind of problem suggesting that she may have successfully identified the vocabulary she wanted to use during the planning stage. In addition in sample TT29a, the test taker omitted a crucial element of the narrative (the children showing the mother the figure they have made), but rather explained that the mother was scared. This may have been an element of the narrative that the raters considered obligatory for successful task completion. In contrast, in sample TT29b, the test taker offered the following explanation for the mother's surprise: *because of er mother she look baby's bed :: and baby is not here :: (2) (there is a b) there is a ball*. The comparison indicates that after extra planning time, the test taker was better able to generate vocabulary for the task and provide explanations for the events of the narrative. This may have impacted the raters' decisions when assessing overall task success.

The EBB scale required raters to consider the extent to which the task completion was satisfactory. Section 7.3 explained that extra planning time helped test takers produce more idea units (a measure of propositional complexity and completeness), which may have influenced the raters' impression of task completion using the EBB scale. This may explain why some of the test takers were placed in to level 3 after completing extra planning. In contrast, the analytic scale describes language complexity, accuracy and fluency without referring to task completion. As a result, linking the increases in idea units with the increases in scores on the analytic scale is difficult. As discussed in Section 2.7.3.2, according to Brown's (2006) research, raters are often influenced by elements of the task performance that do not feature in the rating scale. The analytic raters in this study may have been influenced by the increase in idea units despite the absence of explicit reference to 'propositional completeness' (Ellis and Barkhuizen, 2005, p. 154) in the scale contents. The absence of any reference to task success or completion in the analytic scale indicates that the EBB scale is better able than the analytic to account for the differences in task performance after variation in planning time (see Section 7.4.2).

7.4.2 Analytic scale

To restate the MFRM of the analytic scale scores, the overall fair average values were highest under the five-minute planning condition (8.83), followed by the ten-minute (8.65), 30-second (8.31), and one-minute (8.11) planning conditions. A series of Welch's t-tests showed that the following differences between scores were statistically significant: the five-minute and one-minute planning conditions with a moderate effect size of .53, the five-minute and 30-second planning conditions with a small to moderate effect size of .40, and the ten-minute and one-minute planning conditions with a small to moderate effect size of .39 (Cohen, 1988). Once again the differences between scores awarded under the ten-minute and five-minute conditions were not statistically significant.

The scores awarded under the four planning conditions were situated in the same band of the analytic scale and the effect size of variation in the amount of planning time was generally small to moderate. However, when the individual test takers' measure values in the MFRM were calculated for each planning condition (see Section 6.2.2.1, Table 45), it was clear that planning played an important role in the test scores. In total, 26 of the test takers received different scores on the analytic scale as a result of increases in planning time and in three cases the test takers were placed into three different levels (i.e. levels 9, 8, and 7 for test takers 22 and 44, 47). To provide examples of this, samples TT22a and TT22b are the lowest and highest scoring transcripts of test taker 22. Sample TT22a was completed under the one-minute planning condition and was placed in to level 7 on the analytic scale. Sample

TT22b was completed under the five-minute planning condition and was placed in to level 9.

Sample TT22a, Task 2, one minute:

experience is the most important thing in er my life | er and er for example er if I stay er dormitory :: er I can win er a lot of experiences | mm (1) but er (I live with my family) if I live with my family :: er I er don't gain er experience | mm and er thus er (1) (I) er we can er more er confidence :: I think | er (1) (in the future) er (is in the future) er for in the future is more better er for m me

Sample TT22b, Task 4, five minutes:

the woman is er (sitting) m (on the) er (1) sitting on the er (1) and in front of the baby | er and then er the woman er :: I think er children's er (mother) er mother er fall er asleep | er and two children er come to er living room | m suddenly er two children er take er to baby m baby's basket | er and then er (1.5) two children m (1) they er ha put the baby m (1) instead of m put the toys er baby | er and they run away | er (the) er (mother) (is) the mother awakes :: er and the woman is frightened

As the analytic scale contains descriptors of complexity, accuracy and fluency, a comparison of CAF results between samples TT22a and TT22b may provide a basis for interpreting the difference between the test scores. The CAF results demonstrate that sample TT22b contained a higher percentage of K2 vocabulary, involved more accurate use of verbs and prepositions, recorded higher values in total speaking time and speech rate, and contained more idea units. However, sample TT22a involved slightly more clauses per AS unit. This may be due to the test takers' use of the conditional clause structure in sample TT22a, which increased the average number of clauses per AS unit. Furthermore, sample TT22a recorded a higher Guiraud's index value. Sample TT22b recorded lower values in accuracy of articles and contained a higher value in the mean number of errors per AS unit. The fluency results demonstrate that the phonation time ratio and mean length of utterance results were also slightly higher in sample TT22a. The inconsistency between CAF results and test scores indicates that raters may have placed greater value on certain features of the speech. For instance, the raters may have considered accuracy of the verb phrase to be more important than accuracy of the article system. This is certainly a possibility. An error within a verb phrase has more potential to impede communication than an error with an article (Foster and Wigglesworth, 2016). However, this interpretation requires further investigation. Areas for future research are discussed in the conclusion (see Section 8.5).

Using the same analytic scale, Iwashita et al. (2001), Elder et al. (2002), and Elder and Iwashita (2005) found no difference in test scores after increases in pre-task planning time. In contrast, the findings of the current study indicate that planning did impact test scores, but that the overall impact was relatively minor. The fair average values fell in to the same band regardless of variation in planning time. Interpreting this difference between the studies, it is important to bear in mind that different amounts of planning time were provided in Iwashita et al. (2001), Elder et al. (2002), and Elder and Iwashita (2005) and the current study. Elder and Iwashita (2005) discuss the possibility that their unplanned condition (75 seconds to read the test rubric) may have been sufficient time for test takers to prepare a language plan and that increasing planning time to three minutes therefore made little difference to test scores. While this is certainly a possibility, the current study uncovered a difference in scores of .11 logits (.20 fair average on the analytic scale) on the same scale awarded after minimally different lengths of planning time (30 seconds and one minute), indicating that even slight changes to pre-task planning can impact test scores.

In addition to the overall analysis of the analytic scale results, three MFRM analyses were completed on each category of the scale (complexity, accuracy, and fluency; see Section 6.2.2.2). To summarise the results of the complexity category, variation in planning time caused minor increases in the fair average values. The results demonstrate that the five-minute planning condition resulted in an overall fair average value of 2.80, the ten-minute condition was 2.77, the 30-second condition was 2.65, and the one-minute condition was 2.56. The only result that reached statistical significance was between the scores awarded under the five-minute planning condition and those awarded under the one-minute planning condition with a moderate effect size of .44 (Cohen, 1988). The variation in planning time made little difference to the overall fair average values. At level 2 on the complexity scale, the test taker 'produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty' (See Appendix 1). The test takers were unable to produce substantially more complex language after planning.

This result is not surprising. Complexity scores were generally low regardless of the amount of planning time. This may be a product of the test takers' limited proficiency. As Kawauchi (2005), and Mochizuki and Ortega (2008) report, when L2 resources are limited, planning does not have a large impact on speech complexity. This also seems to be the case in the current study. The test takers may not have had the linguistic means to use complex structures and planning for increased lengths of time did not affect the complexity of language use. According to the scale, at level 2 test takers typically attempt to use coordination and subordination but do so with hesitancy and difficulty (see Appendix 1). The results of this analysis suggest that planning has little bearing on this aspect of the test takers' speech.

The results of the MFRM of the accuracy category data indicate that planning did not have a statistically significant impact on scores. The fair average values show that test takers were generally placed in to level 2 on the scale (see Section 6.2.2.2). At this level, test takers demonstrate limited linguistic control and major errors are present in their speech (see Appendix 1). Variation in pre-task planning time had no impact on this aspect of their speech. This finding corresponds to many results reported in the literature where pre-task planning was not shown to impact speech accuracy (Crookes, 1989, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006, Nielson, 2013, Wigglesworth, 1997, Yuan and Ellis, 2003).

The finding that planning had no impact on accuracy contradicts predictions made by Ellis (2005, 2009). Ellis suggested that the extra cognitive resources made available through planning help test takers improve their levels of accuracy by permitting a focus on form. However, the results of the current study indicate that if planning does promote a focus on form, this does not lead to any discernible differences in the levels of test takers' accuracy (measured through CAF and rater assessment). Rather, at least at this level of language ability, levels of accuracy are stable and do not vary in line with increases in planning time. To exemplify this, Table 62 presents transcript extracts from two test takers. The table provides the focus of the grammatical analysis, the task number and planning time, a sample of the transcript, the individual accuracy scores awarded by the raters, and the fair average accuracy grade generated through *Facets*.

Table 62 Examples of speech inaccuracies and rater scores

Test taker and focus	Task and Planning	Extract	Accuracy grades	Fair average
4: subject verb agreement	3 (30 sec.)	*the man (1) er sell balloons*	3,3,2	2.64
	4 (five min.)	*the girl er hold up a baby*	3,3,2,3	
27: articles	3 (one min.)	*they go and buy balloon*	3	2.55
	4 (ten min.)	*and boy put ball (1.5) (in) into*	3,3	

Test taker 4 completed Task 3 (Balloon Task; see Appendix 3) and Task 4 (Baby Task; see Appendix 4) using the present tense. The extracts show that the test taker made similar subject verb agreement errors under both the 30-second planning condition and the five-minute planning condition. Planning did not make any difference to this test taker's levels of accuracy in subject verb agreement. This was reflected in the individual grades assigned to this test taker, which were very similar between the planning conditions. Test taker 27 completed Task 3 after planning for one minute, and Task 4 after planning for ten minutes. However, as the extracts demonstrate, the lack of accuracy in the test taker's use of articles was constant regardless of the variation in planning time. In summation, the test takers' levels of speech accuracy and the raters' assessment of speech accuracy was not affected by the variation in planning time. This finding contradicts claims made by Ellis (2005, 2009) that the extra cognitive resources that planning makes available may be spent focussing on such forms, with the result that they are produced with better accuracy.

The analysis of the scores on the fluency category indicated that planning had the largest impact on this element of the scale. The five-minute planning condition generated a fair average value of 3.15, under the ten-minute condition the fair average was 3.05, and under the 30-second and the one-minute conditions it was 2.87. The results of the Welch's t-tests showed that the differences between scores awarded under the following planning conditions were statistically significant: the five-minute and one-minute conditions with a moderate effect size of .48, the five-minute and 30-second conditions with a moderate effect size of .51, and the ten-minute and 30-second conditions with a small to moderate effect size of .39. To summarise these findings, the five-minute planning condition resulted in the highest fluency scores and the largest difference was between the scores awarded under the 30-second and five-minute conditions.

Although the impact of planning on fluency scores was generally limited, increasing planning time from 30 seconds or one minute to five minutes had the effect of causing fair average values to move up a band from level 2 to level 3 (although the difference in fair average values was marginal at .28). To contextualise this increase, at level 2 test samples contain 'a marked degree of hesitation due to word finding delays or inability to phrase utterances easily' (see Appendix 1). A score of 2 indicates that the test taker is disfluent and struggles to produce speech in the second language, a red flag for an English-medium university. At level 3, the test taker 'speaks more slowly than a native speaker due to hesitations and word finding delays' (see Appendix 1). Therefore, the increase in planning made hesitation less marked, although speech speed was slower than that of a native speaker. That the test takers spoke more slowly than native speakers is not controversial. The test takers generally do not have the opportunity to use the L2 in their daily lives and have not acquired the skills that facilitate native like speech production (see Section 2.6).

The finding that planning improved speech fluency supports conclusions reported in the literature. Nitta and Nakatsuhara (2014) indicate that on their amended version of the Iwashita et al. (2001) scale (i.e. with extra performance levels added between the original bands), scores on the fluency category was impacted most by the introduction of extra planning time. Their results also showed that complexity scores were minimally improved and accuracy scores remained the same, much like the results of the current study.

Although the increases in fluency scores are consistent with the literature, these results are not supported by CAF results, which did not show any statistically significant difference between the four planning conditions. This finding is unusual given the match between the fluency measures, specifically the mean number of hesitations and speech rate, and the band level descriptors at levels 2 and 3, which describe hesitation and speech speed. However, as discussed in Section 7.4.1, performance features that did not appear on the scale (i.e. the number of idea units) may have influenced the analytic raters when assigning scores. In short, the fluency scores may reflect more than hesitation and speech speed. Deviation from the rating scale is indicative of construct underrepresentation, which is a threat to the validity of decisions based on the test scores (see Section 2.7.3.1). Future research in which raters explain how they scored speech samples (e.g. involving stimulated recall methodology) may help to resolve this issue. Lumley (2005) casts doubt on the possibility of identifying common rater behaviour through stimulated recall. In his research, raters reported that they had applied the same scale in different ways when grading a series of L2 writing tasks. However, to identify features of task performance that do not appear in the scale but affect scores nonetheless, stimulated recall would be an appropriate method to inform scale development. This is discussed in detail in the conclusion (see Section 8.5).

7.4.3 Comparison of the analytic and EBB results

Comparing the results of the analytic scale with the results of the EBB scale, the five-minute planning condition consistently resulted in the highest fair average values. However, it is important to note that on both scales, the differences between scores awarded under the five-minute and ten-minute planning conditions did not reach statistical significance and it is not possible to categorically conclude that five minutes was the most advantageous planning condition.

Overall, the effect sizes of the difference between the five-minute planning condition scores and the lowest scores on both scales are similar. The largest effect size of .46 on the EBB scale is between the scores awarded under the five-minute and 30-second planning conditions, whereas the largest effect size on the analytic scale is between the scores awarded under the five-minute and one-minute conditions at .53.

Using the fair average values, the average test score on the EBB scale after the one-minute, and 30-second planning conditions was level 2. At this level, task completion is unsatisfactory and strains the rater. However, extra planning time (five minutes and ten minutes) increased the fair average value to level 3. At this level the task is completed satisfactorily and causes minimal strain to the rater. Extra planning therefore had an important impact on fair average values on the EBB scale. In contrast, the fair average values on the analytic scale fell within the same band level under each planning condition. This indicates that variation in planning time had less of an overall impact when raters used the analytic scale. This may be explained by the absence of reference to task content, propositional completeness, and propositional complexity on the analytic scale (see Section 7.4.1).

7.5 Does the impact of the four planning conditions on test scores and CAF results vary between groups of test takers who have different levels of language proficiency?

The test taking population that this study samples do not have vast experience of using the L2 to communicate and generally struggle to produce spoken English, especially in high stakes assessment contexts (O'Sullivan and Green, 2011; see Section 2.4.1). Evidence of this was provided in the scores on the rating scales, which were particularly low and indicated generally limited levels of proficiency in the sample. The Oxford quick placement test (QPT; UCLES, 2001) was used to obtain an independent measure of the test takers' English language proficiency and test takers were separated in to A level (CEFR levels A1 and A2, Council of Europe, 2001) and B level (CEFR B1) proficiency groups based upon the results. Comparisons between each groups' CAF results and test scores were then made.

To review the CAF findings, there was no interaction between proficiency and pre-task planning in the results of the analysis. Regardless of variation in language proficiency, extra planning time helped the test takers to generate content, which increased the number of idea units they produced (see Section 6.3.4). In terms of test scores (Section 6.2.2.3), the MFRM demonstrated that A level participants' scores increased when planning time was increased from 30 seconds to ten minutes. The difference between the ten-minute and 30-second planning conditions was .39 on the logit scale. The Welch's t-test demonstrated that this difference was statistically significant with a small to moderate effect size of .40 (Cohen, 1988). The B level participants' scores increased with extra planning time and the five-minute condition resulted in the highest scores. However these results did not reach statistical significance.

The differences between the A level and B level results may further account for the finding that the five-minute condition resulted in the highest scores when all test takers were included in the MFRM (see Section 7.3). The results indicate that the B level test taker scores were highest under the five-minute condition (although statistical significance was not reached). When both groups were combined in the MFRM, the five-minute planning condition resulted in the highest scores but the difference between the scores awarded under the five-minute and ten-minute conditions did not reach statistical. Excluding the B level data from the MFRM to examine the impact of planning on the A level scores appears to have decreased the number of high scores awarded under the five-minute condition. Based on this interpretation, test taker proficiency was a key variable in the outcome of the study.

Presumably, completing the language tasks after planning for 30 seconds proved especially challenging for the low-level (A level) group and the opportunity to plan for ten minutes mitigated this challenge. However, variation in planning time did not affect the scores of the more advanced-level (B level) participants who may have been better equipped to deal with the challenge of producing relatively spontaneous speech (i.e. after 30 seconds planning).

In the literature review (see Section 2.6.3), the suggestion was made that in order for test takers to benefit from the opportunity to plan their speech at length, they should have acquired sufficient knowledge of the L2 to access and generate task relevant language forms. This implies that advanced-level test takers benefit most from the opportunity to plan because they have access to a wider range of language than low-level test takers. It is surprising then that it was the low-level group who benefitted from the extra planning time rather than the more advanced group.

To hypothesise about the cause of this result, narrating a series of pictures and describing personal experiences in the L2 may pose more of a challenge for low ability test takers than more advanced test takers because at these lower levels, test takers may not have access to relevant language to discuss task content. The literature review indicates that the more challenging test takers find a language task, the more pre-task planning is used to reduce the challenge (see Section 2.5.1.4). Task challenge is related to language proficiency because novices typically struggle to complete a task that experts complete with little effort (see Section 2.5.1.3). Therefore, pre-task planning affected the test takers with very limited second language proficiency most clearly because they found the tasks particularly challenging.

Genc (2012) conducted her study in a comparable context with Turkish learners of L2 English and found that planning did not affect participants' speech accuracy but did increase the amount of speech her participants produced. Genc argues that when language learners have limited second language ability, they are preoccupied with the generation of ideas and do not attend to form during planning. In light of the current findings, the indication is that pre-task planning does not affect low-level language learners in the same way that has been reported in the literature (e.g. Foster and Skehan, 1996, Skehan and Foster, 1997).

In an English as a foreign language context such as Turkey, opportunities to develop speaking skills are limited. In contrast, in English as a second language contexts, frequent communication in the L2 helps language learners develop L2 knowledge that can be accessed during planning and may lead to noticeable differences in complexity, accuracy and fluency. Without such experience, accessible resources are limited and the impact of pre-task planning on task performance may be constrained to the generation of ideas.

7.6 Does the impact of the four planning conditions on test scores and CAF results vary between picture-based narrative tasks and non-picture-based description tasks?

7.6.1 CAF

Four tasks were used in the study. These included two non-picture-based description tasks (i.e. Task 1: '*Tell me about something interesting you have recently heard in the news*' and Task 2: '*Tell me about an event that has changed your life*') and two picture-based narrative tasks (i.e. Task 3: '*Balloon task*' and Task 4: '*Baby Task*'; see Appendix 3 and 4).

To review the impact of planning on the tasks as measured by CAF, the two-way ANOVA (see Table 61) revealed statistically significant differences between the four tasks under different planning conditions in the use of K2 lexis. K2 lexis refers to vocabulary that occurs in the second most frequent thousand words in the British National Corpus and may thus be considered more advanced than K1 lexis (Laufer and Nation, 1995). The two-way ANOVA result was confirmed by the results of the multiple regression, which indicated that 12.6% of K2 vocabulary use was attributable to variation in planning time and task type. Thus, although there was a statistically significant impact of planning on K2 lexis, the impact was relatively minor. In recent research, K1 and K2 vocabulary is combined in statistical analyses to form the most common 2,000 words, indicating that researchers regard the distinction between the two levels as minimal (Laufer, Elder, Hill and Congdon, 2004). However, evaluating the examples of K1 and K2 lexis in the transcripts, it is clear that K2 lexis permits the speakers to express their ideas with greater precision. For instance, whereas K1 lexis in the transcripts included words such as *bad* and *good*, the K2 equivalents were *harmful* and *honest*. Further examples of K2 lexis in the transcripts included *self-confidence*, *disease*, *imaginary*, *shocked*, *damage*, *government*, *jewellery*, and *perfect*. The presence of such vocabulary in the samples may have had a bearing on test scores, although in the absence of data about the raters' scoring processes it is not possible to confirm this hypothesis (see Section 8.5.2). Given the evidence that planning had a minor impact on this aspect of the test performance, the K2 vocabulary results require further attention.

On Task 1 the highest percentage of K2 lexis was associated with the one-minute planning condition (8.18), and the lowest with the ten-minute planning condition (2.59). On Task 2, the highest percentage was observed under the one-minute planning condition (5.37) and the lowest under the five-minute planning condition (.47). On the two description tasks, the one-minute planning condition generated the highest use of K2 lexis. In contrast, the one-minute planning condition was associated with the lowest levels of K2 lexis on the narrative tasks (on Task 3 the K2 value is 4.95, and Task 4 it is 2.97). The highest levels of K2 lexis were observed under the five-minute planning condition on Task 3 (7.66) and under the 30-second planning condition on Task 4 (6.77). These results indicate that there is an interaction between planning time, task number and lexical sophistication in the results.

Speaking relatively spontaneously on the description tasks (i.e. after planning for one minute) increased the test takers' use of K2 vocabulary (see Figure 18). This was an unexpected result. As discussed in the literature review (see Section 2.7.2.1), no research has been conducted that seeks to establish the impact of planning on lexical sophistication. There is no opportunity to compare these results with the literature. However, the increase in the number of idea units indicates that there may be some interaction between the use of sophisticated lexis and the amount of speech the test takers produced.

The lexical sophistication measure is reported in terms of the percentage of lexis that was the first 1000 words, the second 1000 words and the academic word list. As these results are reported as a percentage, the less speech the test taker produced the more chance there was for examples of K2 lexis to impact these values. For example, if a test taker produced 50 words and five of these were K2, then the overall percentage of vocabulary that was K2 would be 10 per cent. In comparison, if a second test taker produced 200 words, 20 of these words would need to be K2 in order to record a value of 10 per cent. The results of the idea unit analysis demonstrate that test takers generally produced more ideas after planning for longer periods (although this result only approached statistical significance in the two-way ANOVA: see Table 61). This increase in idea units may have impacted on the percentages of K2 lexis.

To provide an example of the interaction between K2 lexis and idea units, in sample TT27a (see Section 7.2) the total number of words was 53 and 9.43 per cent were K2 (*explosion, explosion, lots, lots, sad*). In sample TT27b, the total number of words was 102 and 3.92 per cent were K2 (*lots, match, parents, sport*). In sample TT27a, three of the K2 words were repeated, whereas in sample TT27b the four K2 words were used once. This indicates that the amount of K2 lexis produced was relatively stable but when the test taker produced more speech this reduced the percentage of the text that was made up of K2 lexis.

Inoue (2013) discusses a similar impact in her study in which increased speech content also caused increases in the number of inaccuracies test takers produced. This finding suggests that variation in text length and the number of lexical repetitions affected the results of the VocabProfile program (Cobb, n.d) [accessed 1 October 2015 from http://www.lextutor.ca/vp/]. A possible solution for this problem may be to use an alternative approach to the measurement of lexical sophistication such as dividing the transcript into a series of segments (e.g. 40 words; Yuan and Ellis, 2003) and calculating an average value.

The impact of variation in planning time on K2 lexis on the narrative tasks followed a slightly different pattern (see Figure 18). In both narrative tasks, the ten-minute planning condition led to the highest number of idea units and also high levels of K2 lexis. This suggests that for the narrative tasks, planning for ten minutes increased both the number of idea units test takers produced and lexical sophistication. This finding is similar to the conclusions that have been reached in the TBLT literature (e.g. Crookes, 1989, Foster and Skehan, 1996, Skehan and Foster, 1997). Namely, that a period of ten minutes pre-task planning for a picture-based narrative task benefits the language learners' speech complexity specifically. It further indicates that test takers were more responsive to pre-task planning on the picture-based narrative tasks because extra planning facilitated both the generation of speech content and the production of less frequent lexis. This is discussed in detail in Section 7.7.

7.6.2 Test scores

Turning to the test scores, on Tasks 1, 3, and 4 the highest scores were observed under the ten-minute planning condition (see Section 6.2.2.4). On Task 2, the five-minute planning condition generated the highest scores. However the differences between the five-minute, 30-second, and ten-minute condition scores on Task 2 were marginal on the logit scale (at -.34, -.32, and -.25 respectively) and did not reach statistical significance (see Table 54). On all four tasks, the lowest scores were associated with the one-minute planning condition. Comparing the results between the two task types, scores on the picture-based narrative tasks were more likely to be affected by variation in planning time than the scores on the non-picture-based description tasks. This can be observed in the effect sizes of the planning conditions for each task (see Section 6.2.2.4).

To begin with the non-picture-based description tasks, Task 2 (*Describe an event that has changed your life*) was least impacted by variation in planning time, although the effect size between the five-minute and one-minute planning conditions was large (1.16). When all facets were included in the analysis, test takers received the highest grades on Task 2, indicating that this was the simplest of the four tasks (see Section 6.2.1). To clarify this, speaking about a personal topic that is familiar was less challenging than speaking about current affairs or narrating a series of images. Test takers may thus have been less dependent upon the planning time when completing Task 2. Foster and Skehan (1996) report similar findings in their comparison of the planning impact on three tasks (a personal information task, a picture-based narrative, and a decision making task). In their research, the task that involved personal information recorded the highest CAF results, and was least impacted by the addition of extra planning time.

Task 1 (*Describe something interesting that you have heard in the news*) proved to be more difficult than Task 2 and test takers benefitted more from the opportunity to take part in extra planning. This can be observed in the larger effect size of the planning variable on Task 1 (1.28), which was higher than Task 2 (1.16). Describing current affairs required references to events and people that were not part of the test takers' daily experience, and the language to express these ideas may not have been as available as the language required to describe a personal experience. This is exemplified in samples TT7a (test taker 7, Task 1, the five-minute planning condition) and TT7b (test taker 7, Task 2, the 30-second planning condition).

Sample TT7a, Task 1, five minutes:

in recently er I heard that :: er in Turkey some er politicians are in prison because of the government | er government erm doesn't want to have independence opinions | er it is so bad for Turkish people | er we haven't er real republic and | m we cant do :: whatever we want mm | we want to have a laic system in Turkey | and our politicians are and important poems are in prisons | it is very bad for us | er we don't want to continue er with these er news (1) | er we want to live er more (independence) er in independence and more free

Sample TT7b, Task 2, 30 seconds:

last year er in these er times we learned :: er my mother was ill | she went to er doctor :: and her illness was cancer | er and I was shocked :: w when I heard this | and after she had a big treatment | er she took chemotherapy :: and after radiation therapies | er and this summer everything finished | er and now she is healthy | but I learn er er a lot of things er | I changed my lifestyle er :: because of for example er I don't think a lot (of) er about everything | I think er only er enough for me er | and I want to be happier in my life :: because happiness er is the key of everything | er it was very er important experiences for me er | I want to continue this way

In sample TT7a, the test taker discussed the imprisonment of politicians and poets (referred to as 'poems' in the transcript) in Turkey. The test taker planned for five minutes and produced a total of ten idea units. In contrast, in sample TT7b, the test taker produced 13 idea units about an illness in her family after planning for 30 seconds. Despite the difference in planning time, the test taker produced more idea units in sample TT7b. This may be because the test taker had more to say about Task 2 and did not require the extra planning time to generate content for her speech.

Turning to the picture-based narrative tasks, the results showed that Tasks 3 and 4 were more substantially impacted by variation in planning time than Tasks 1 and 2. For example, on Task 3, the effect size of the difference between scores on the logit scale under the ten-minute and one-minute planning condition was large at 1.55. In addition, the difference between the five-minute and one-minute planning condition scores also recorded a large effect size of 1.45. Both values exceed those observed in the description tasks. On Task 4 the effect size of the difference in scores between the ten-minute and one-minute planning conditions was 2.49. This value exceeds the effect sizes of the planning variable on Tasks 1, 2 and 3 considerably. Comparing the logit measure values between the tasks, it is clear that completing Task 4 under the one-minute planning condition was the most difficult task condition for test takers.

Overall, these results indicate that planning has more potential to increase test scores when tasks are challenging (i.e. tasks that record low average scores are regarded as challenging). This finding broadly corresponds to claims made by Skehan (2009). Narrating a series of images that contain obligatory content was more difficult than discussing personal information (i.e. Task 2). Skehan (2009, p. 524) refers to this obligatory content as 'the non-negotiability of the task'. He states 'a narrative is necessarily input-driven, and unforgiving in what needs to be covered' (2009, p. 517). This has implications for the test taker who must access suitable lexis to meet the task demands: picture-based 'narratives seem to push second language speakers… into using less frequent lexis' (2009, p. 517). The results of this study indicate that pre-task planning may have facilitated the generation of language to describe obligatory content to the extent that raters provided higher scores to extensively planned samples (i.e. the ten-minute planning condition). This indication is discussed in detail in Section 7.7.

7.7 Which task type and planning condition has the largest impact on test scores and CAF results?

As discussed in Section 7.6, the impact of the pre-task planning variable was strongest on the picture-based narrative tasks. It was suggested that the large planning effect on these tasks was associated with the requirement to narrate a series of images involving obligatory content. In contrast, the non-picture-based tasks did not involve obligatory content and test takers were free to decide what to communicate and what to avoid during the task. Planning had less of an impact on these tasks.

Comparing the effect sizes of the pre-task planning variable on the picture-based narrative tasks, the variation in planning time had the largest impact on the test scores of Task 4. The largest effect size of 2.49 was between the scores awarded under the ten-minute planning condition and the scores awarded under the one-minute planning condition. In comparison, the difference between scores awarded under the ten-minute and one-minute planning conditions on Task 3 recorded an effect size of 1.55. This means that test takers were better able to compensate for the level of difficulty associated with certain task characteristics in Task 4 by taking part in extra planning. This section identifies these task characteristics by analysing four samples that were produced under the one-minute and ten-minute planning conditions.

Task 4 (*Baby Task*, see Appendix 4) depicts a scene in which two children replace a sleeping baby with a ball to play a joke on their mother who is also asleep. The mother wakes up to find that the baby is missing and is shocked. In the following transcript, sample TT5a, test taker 5 described these events under the ten-minute planning condition.

Sample TT5a, Task 4, ten minutes:

er (in) in this picture I see one mirror and | er I think :: there is a war :: because I see fire in the room and | (1) one woman er reading a book | one boy one little boy one baby is sleeping | (1) then she slept | two guy comes :: and er (put baby) (1) er bring baby out | (1) then they put a scary er face balloon :: where was the baby sleeping | (1.5) er (then woman) (2) then er they (they) waited er :: for the woman wake up | (1) woman wake up and :: er see a er scary face balloon | she frightened | er (two guy) two guys er (1) er smile er behind her

The test taker began by providing details about the scene. This section is relatively free of hesitations (there is one hesitation in the first four AS units) suggesting that the test taker had to some extent prepared the content of the first four AS units during the planning time. She commented on the decoration in the room and reached the conclusion that the scene may have taken place during a war (presumably in the past owing to the datedness of the furniture). After setting the scene, the test taker ordered the events in a sequence, by first describing the woman who was reading, and that the baby was sleeping. Secondly she described the woman falling asleep and the entrance of the two children. The test taker appeared to struggle to formulate language for the children's removal of the baby from the crib. She first suggested that the children 'put (the) baby' but was not satisfied with this and settled for 'bring (the) baby out'. This may be interpreted as evidence of monitoring and indicates that the test taker was concerned with the accuracy of her speech. After describing the replacement of the baby with the ball, the test taker required some time to conceptualise and formulate language to describe the next scene. This is evidenced by the increase in pausing at this point and the presence of the false start (*(1.5) er (then woman) (2) then er they*). As the test taker progressed, the number of hesitations increased until in the final AS unit, she hesitated four times in order to produce a five word sequence. This suggests that the test taker had become reliant on online planning (Yuan and Ellis, 2003) at this stage and either did not plan for this stage of the narrative or experienced difficulty in recalling the plan. In short, the effect of the increased planning time had diminished as predicted by Elder and Wigglesworth (2006) (see Section 7.2).

Sample TT16a, Task 4, one minute:

(I am) I am seeing a baby :: and she is sleeping | er her mother er reading a book | er when she is reading a book :: s she is having a nap | and other children are seeing them | er they are sleeping | (1) and they think :: that (2) er (2.5) they want to :: I think they want to play games and | er they are catching the baby (1) er (2) from the her pocket | and boy (put) (a) (the) er (3) move the *Sey* [thing] (2.5) ball in the pocket | and girl catch the baby | (1) when her mother waking up :: she was frightened | and she think that :: where is my baby (2) that's all

Sample TT16a was completed by test taker 16 under the one-minute planning condition. In this sample, the test taker directly began to relate the events of the narrative without offering any interpretation of the scene. Hesitations began relatively early in the transcript (i.e. in the second AS unit), which is indicative of online planning (Yuan and Ellis, 2003). Online planning is a feature of careful speech production that facilitates retrieval and encoding of linguistic forms, i.e. Levelt's (1989) formulation stage of speech production, but is detrimental to speech fluency. That the test taker showed evidence of online planning during the first few utterances of the task suggests that the one-minute planning condition did not facilitate the generation of language to complete the task. As observed in sample TT5a, the image depicting the removal of the baby and replacement with a ball caused the test taker some difficulty as evidenced by the increase in hesitations, false starts and use of the L1 (*and boy (put) (a) (the) er (3) move the* Sey *[thing] (2.5) ball in the pocket*). Like sample TT5a, in sample TT16a the test taker struggled to generate a suitable verb to describe the children's removal of the baby and settles on catch (*they are catching the baby*). At this stage of the narrative, the test taker became disfluent and the accuracy of the grammar began to decrease: the omission of the auxiliary verb in the continuous aspect (*when her mother waking up*) and mistakes in subject verb agreement (*and girl catch the baby*). Describing the order of the events also posed problems for the test taker who misused the continuous aspect to describe the cause and effect relationship between the mother waking up and becoming frightened (*when her mother waking up :: she was frightened*). This was an obligatory event that needed to be communicated in order to successfully complete the task. The test taker may not have had the relevant linguistic knowledge to describe this scene and was forced into attempting a structure that she was unable to produce accurately. In sample TT5a, the test taker avoided this problematic structure by opting to relate this sequence in the past tense. However, it is not immediately clear whether this is because the test taker anticipated encoding problems with this image during the planning time. Future research involving stimulated recall may help to resolve this issue (see Section 8.5.1).

Sample TT17a, Task 4, ten minutes:

lucy is a mother :: and er (her baby) er her babys name is susan | er lucy is very tired :: because her baby wasnt er sleep last night | and er finally her baby is sleeping :: and er she is reading a book | and then er suddenly she is sleeping | (1) after er her other children is (coming to) er coming room | (1) and er the boy says :: er I want to make a joke for my mum | and er you should (wake) wake up er m my sister :: and put up (he) er her | and then er (she wa) she is waking up er her sister | and er the boy s says er sh :: shut up :: my mother is er sleeping er | and he put er her sisters bed the ball | and then er their mother is wake up :: and she er cant see her baby in bed :: and er she is crying :: er where is my baby | and er the er girl (1) not little one er big one girl :: is (stop my) stop my mum :: just a joke | er and er mum is very angry :: and she saying :: er (you sh) you must go your er room :: and im not allowed to anything | er you are not watching t tv and nothing this (laugh)

In sample TT17a, test taker 17 completed the task under the ten-minute planning condition. This sample is unique in terms of the levels of personal interpretation the test taker opted to include. The test taker assigned names to two characters, provided explanations for the state of affairs (*Lucy is very tired :: because her baby wasn't er sleep last night*) and also provided direct quotations (*shut up :: my mother is sleeping*). The sample contains a relatively high amount of content (14 idea units), which may have been facilitated by the pre-task planning: the ten-minute planning condition on the narrative tasks resulted in considerably more idea units than the other three planning conditions (see Figure 19). The test taker did not seem to experience any difficulty when formulating the problematic image involving the replacement of the baby with the ball, although she uses the vague term '*put*'. However, the test taker experienced difficulty in formulating the correct language to distinguish between the characters in the text, which forced her to explain the use of pronouns ('*not little one er big one girl*'). This contradicts claims made in Robinson (2005) that increasing the number of characters in a language task will increase the test takers' language complexity by forcing test takers into using relative clauses (provided that the test taker has relative clauses in her repertoire). Instead, the requirement to distinguish between characters caused the test taker to use inaccurate language, and hesitate under pressure to explain the pronoun.

Sample TT18a, Task 4, one minute:

maybe first picture er the m (the) woman is er grandmother or mother | mm the baby is sleeping | mm the second picture erm a boy and a girl child er come to the room | and er s er woman is sleeping | (1) er third picture er children er ba ba the baby woke up | er and er (1.5) er four picture er children er put the ball (1) er b on bed | er (4) er six picture er mother is wake up | and a m daughter see the baby | (she) (1) (of) she wondered :: (the) where the baby (1) that's all

In sample TT18a, test taker 18 completed the task under the one-minute planning condition. The test taker hesitated in the first AS unit and filled hesitations frequently occurred throughout the sample. However, unfilled hesitations and pauses (i.e. in excess of one second) did not occur in the transcript until the fifth AS unit. From this stage onward, the number of filled and unfilled hesitations and pauses increased and the test taker became disfluent. The increase in the frequency and length of pauses and hesitations observed after the fifth AS unit indicates that the test taker required online planning as the task progressed to generate task content and relevant language. In the same way that sample TT16a was completed, the test taker described the images without offering any interpretation of the events. The description was minimal and there was little attempt to sequence the events in order (e.g. with adverbial phrases such as then and after). Rather, the test taker stated the number of the picture that was described. The test taker did not provide a description of the scene in which the children take the baby from the crib (*children er ba ba the baby woke up*). Perhaps realising that the relevant language was unavailable, and under pressure to complete the task, the test taker opted to narrate an event that was not depicted in the images but that could be expressed using his language knowledge. This means the test taker may have been forced into improvising content because the events depicted in the image required language that he was unable to generate.

7.7.1 Summarising Task 4 characteristics and the impact of planning

According to the results of the MFRM, when test takers found a task challenging the potential for pre-task planning to impact the test scores was strongest. For this reason, this section has discussed the characteristics of Task 4 that appear to have posed a challenge to the test takers (based on an analysis of transcripts, see Section 8.5.1). The task characteristics that have been identified include:

- Limitations in the lexicon to describe obligatory content in the narrative sequence. Specifically, the verb *remove* or *replace* was most suitable in the context but test takers generally used vague terminology such as *catch*, *put*, and *bring*. Extra planning time may be used to consider lexis to describe such content.
- Limitations in grammatical ability to sequence obligatory events, such as a cause and effect relationship. Obligatory content caused test takers to attempt language that was beyond their current level of ability, which may have lead to mistakes and errors. It may also have caused test takers to deviate from the events depicted in the images and discuss events that did not occur in the narrative. This problem might have been avoided when the test takers were able to generate their own content (i.e. Tasks 1 and 2). Extra planning time may enable test takers to avoid problematic structures and consider alternatives.
- Having to distinguish between two similar characters led to ambiguous pronoun usage, which may require explanation, lead to increases in the number of pauses and hesitations, and derail the flow of the narration.
- Having to process and describe the images simultaneously caused the test taker to hesitate during the early stages of the narrative. This appears to have been mitigated by the opportunity to plan. Although this interpretation was not supported by the results of the statistical analysis of the CAF measures,

the analysis of the samples indicates that after extra planning time, test takers hesitated less during the early stages of the narrative.

- Processing and describing the images simultaneously may have prevented the test taker from interpreting the images in terms of the setting and the characters' motivations. Without extra planning time, the test takers provided minimal descriptions of the images with little connection between the events. Under the ten-minute planning condition, the test takers produced more ideas about the task content, which resulted in a detailed description.

7.8 Summary

This chapter has discussed the results of the main study. Overall, the findings partially support the claims made by Elder et al. (2002), Elder and Iwashita (2005), Iwashita et al. (2001), and Wigglesworth (1997) that planning has little meaningful impact on the results of task-based language assessment. Test takers did not benefit extensively from increases in planning time in terms of complexity, accuracy and fluency. However, the findings indicate that extra planning time did permit test takers to elaborate on the task content, which increased test scores. In addition, contrary to the findings of much of the TBLT literature (Foster and Skehan, 1996, Skehan and Foster, 1997), the five-minute planning condition had the largest impact on test scores when all test facets were included in the analysis, although statistical significance was not reached for the differences between scores awarded under the five-minute and ten-minute planning conditions.

In addition to the overall MFRM analysis (i.e. accounting for all test facets), detailed analysis of the results of the independent variables revealed interesting effects of pre-task planning. Firstly, the effect of the pre-task planning variable on both rating scales was discussed. The findings indicate that the extra planning time may have caused increases in scores at an important stage of the EBB scale, namely the point where raters consider whether the task was completed successfully. The results suggest that the extra planning time (the five-minute and ten-minute planning conditions) caused a change in test taker performance whereby they advanced from unsuccessful task completion to successful task completion. This may have been a product of the increase in propositional completeness, as indicated in the increase in idea units, that planning facilitated. The binary, holistic approach to rating necessitated by the EBB scale therefore reveals that planning played an important role in the test scores. In contrast, the results of the analytic scale indicated minor increases in complexity and fluency scores and offered little information about how successful the task completion was.

The finding that the lower-level participants benefitted most from the extra planning time was also discussed. This was an unexpected result that suggested low-level test takers might have required extra planning time to successfully complete the tasks. The literature review indicated that advanced test takers have more knowledge of the target language to access during planning and should benefit most from the opportunity to take part in extra planning. However, this interpretation is not supported by the results of this study. Rather, low-ability test takers required support to complete the tasks and when support was provided in the form of pre-task planning they achieved marginally higher scores.

The chapter discussed the finding that the planning variable had the largest impact on the picture-based narrative tasks. According to the literature review (see Section 2.5.1.2), describing a series of images causes problems when test takers do not possess adequate linguistic knowledge. This effect was partially mitigated in Tasks 1 and 2, as learners were free to avoid any potentially complicated language structures. However, extra planning time still impacted on the test scores on these tasks positively.

Pre-task planning had more of an impact on the picture-based narrative tasks because it permitted test takers to process and plan for obligatory language structures that would otherwise cause problems during spontaneous speech production. The largest effect sizes between scores on the picture-based narrative tasks were between those awarded under the ten-minute planning condition and those awarded after the one-minute condition. To clarify this, test takers benefitted from the largest amount of planning time when completing the picture-based narrative tasks. Of the two narrative tasks, the task that was most impacted by the planning variable was Task 4. A list of task characteristics that contributed to the task challenge was identified and the potentially mitigating effect of planning the speech was discussed.

Overall, the findings of this study indicate that pre-task planning had more of an impact on the quality and quantity of the task content than complexity, accuracy and fluency. In short, planning increased the ‘propositional completeness’ (Ellis and Barkhuizen, 2005, p. 154) of the task content. This is a feature of the test taker’s task completion that cannot be investigated with measures of complexity, accuracy and fluency, or with a rating scale based on these constructs. This flaw in the CAF method has been discussed in the literature (Pallotti, 2009, Fulcher, 2015). However, regardless of the absence of relevant descriptors in the analytic rating scale (Iwashita et al., 2001), the degree of propositional completeness in the test samples may have played a role in the raters' assessments. Because the EBB scale contained criteria relating to the degree of task success, the findings of the EBB score analysis provided more interpretable information concerning the impact of pre-task planning time on propositional completeness and hence test scores.

Kuiken and Vedder (2014) have emphasised the importance of task content to raters' perceptions of L2 ability in their research about L2 writing assessment. Their research findings suggested that raters placed more importance on 'communicative adequacy (content, use of arguments, rhetorical organization, style and general comprehensibility) than to linguistic complexity' (2014, p. 341). This led the researchers to develop a new rating scale for writing assessment that accounted for what they refer to as functional adequacy, involving task features such as content, task requirements, comprehensibility and coherence and cohesion (Kuiken and Vedder, 2017).

In research on second language speaking assessment, Sato (2012, p. 235) found that 'content elaboration/development' made a major contribution to the raters' assessment of 'overall communicative effectiveness'. Sato (2012, p. 237) concludes that 'the quality of ideas that test-takers attempt to convey should be treated as a criterion in oral assessments' and that 'narrowly restricting our focus to linguistic features may lead to erroneous inferences about L2 learners' ability to communicate effectively'. The results of this study indicate that task content did play a role in the raters' assessment despite the fact that it was not part of the scale criteria (i.e. the analytic scale). In addition, it was the content related aspects of the speech (i.e. idea units) that were impacted by the addition of extra planning time to the language task. Therefore, efforts to develop rating scale criteria that describe the propositional aspect of spoken proficiency would not only reveal more about the impact of pre-task planning, but also increase the validity of second language speech assessment in general.

8 Conclusions

8.1 Introduction

This study sought to determine the impact of variation in planning time in an admission test for an English-medium university in Turkey (see Section 1.1). Planning before a second language task is widely regarded as advantageous for the speech production process (Ellis, 2005, 2009, Robinson, 2005, Skehan, 2016). However, in language testing contexts the evidence for the benefits of pre-task planning is limited (Wigglesworth, 1997, Elder and Iwashita, 2005, Elder and Wigglesworth, 2006, Nitta and Nakatsuhara, 2014).

There were two motivating factors for this investigation (see Section 1.1). The first was to establish whether the addition of pre-task planning in the speaking tasks of the university admission test would 'bias for best' (Swain, 1985, p. 42). In second language assessment, it is important to elicit the best possible performance from test takers to ensure that they have a fair chance of passing the test.

The second motivating factor concerned test validity, specifically Weir's (2005) and O'Sullivan's (2016) context and cognitive elements of validity. The target language domain of undergraduate English-medium instruction involves various situations in which students are required to plan their speech in detail (Wigglesworth and Elder, 2010). Therefore, a test that is assumed to assess prospective students' ability to function in this domain must replicate the conditions as much as possible by including tasks that feature planning before speech production.

The research questions that this study aimed to answer were:

1. Does variation in planning time operationalized as 30 seconds, one minute, five minutes and ten minutes impact the results of a language test when assessed with

   a) an EBB scale

   b) an analytic scale

   c) measures of complexity, accuracy, and fluency (CAF)?

If the answer to research question 1 is affirmative,

1.1 Which amount of planning time (30 seconds, one minute, five minutes, ten minutes) most substantially impacts test scores and CAF results?

1.2 Does the impact of the four planning conditions on test scores vary between the analytic scale and the EBB scale?

1.3 Does the impact of the four planning conditions on test scores and CAF results vary between groups of test takers who have different levels of language proficiency?

2. Does the impact of the four planning conditions on test scores and CAF results vary between picture-based narrative tasks and non-picture-based description tasks?

If the answer to research question 2 is affirmative,

2.1 Which task type and planning condition has the largest impact on test scores and CAF results?

To answer these questions the study adopted a quantitative methodology in which speech samples were collected from test takers under different pre-task planning conditions and assessed by trained raters, and with measures of complexity, accuracy and fluency (CAF). The literature review identified four variables that might impact the outcome of this investigation.

The first variable was the task type. The tasks under investigation were divided into two categories, which were picture-based narrative tasks and non-picture-based description tasks. The participants completed two picture-based narrative tasks comprised of six images (see Appendix 3 and 4) and two non-picture-based description tasks: '*Tell me about something interesting you have recently heard in the news*' and '*Tell me about an event that has changed your life*'.

The second variable was pre-task planning time. Based on a review of the literature (see Section 2.5.2), four planning conditions were investigated in this study: a ten-minute planning condition, a five-minute planning condition, a one-minute planning condition, and a 30-second planning condition.

The third variable was the levels of second language proficiency in the test taking population. The test takers were categorised into two groups: a low level group (A1 and A2 levels on the CEFR) and a higher performing group (B1 level on the CEFR) according to their results on the Oxford quick placement test (UCLES, 2001; see Section 5.4.1).

The final variable was the approach to measurement, which was categorised into CAF measures and rater scores. The rater scores were subcategorised according to the rating scale that was used: an intuitively derived analytic scale comprising descriptions of complexity, accuracy and fluency at five levels of ability and an empirically based, rater oriented EBB scale (see Section 2.7).

The results of the analyses are reviewed in Section 8.2. Following this, Section 8.3 discusses the implications and contributions this study has made to the language testing literature (Section 8.3.1), the task-based language teaching (TBLT) literature (Section 8.3.2) and to our understanding of language learners/test takers (Section 8.3.3). The limitations of the research are discussed in Section 8.4 and areas for future research are set out in Section 8.5. Section 8.6 offers final comments on the study.

8.2 Review and interpretation of the results

Table 63 summarises the results of the main study. The table presents the research questions and provides a short synopsis of the results in the following column. Following the presentation of the table, this section looks at each research question in turn and suggests the conclusions that can be drawn from the study.

Table 63 Summary of results (continued on page 286)

Research Question	Result
Does variation in planning time operationalized as 30 seconds, one minute, five minutes and ten minutes impact the results of a language test when assessed with *a) an EBB scale* *b) an analytic scale* *c) measures of complexity, accuracy, and fluency?*	Increases in scores were observed on both scales and in the mean number of idea units produced after variation in planning time.
Which amount of planning time (30 seconds, one minute, five minutes, ten minutes) most substantially impacts test scores and CAF results?	The difference between scores awarded after the 30-second/one-minute and five-minute conditions was .36 on the logit scale. The largest number of idea units was produced under the five-minute condition (10.58), an increase of approximately 3 idea units over the one-minute condition (7.82).
Does the impact of the four planning conditions on test scores vary between the analytic scale and the EBB scale?	Scores on the analytic scale increased from 8.11 under the one-minute condition to 8.83 under the five-minute condition. Scores on the EBB scale increased from 2.77 under the 30-second condition to 3.08 under the five-minute condition.
Does the impact of the four planning conditions on test scores and CAF results vary between groups of test takers who have different levels of language proficiency?	A-level (a score below 30 on the QPT; see Section 5.4.1) test taker scores increased by .39 on the logit scale when planning time was increased from 30 seconds to ten minutes. B-level (a score of 30 and above on the QPT) increases were not statistically significant.
Does the impact of the four planning conditions on test scores and CAF results vary between picture-based narrative tasks and non-picture-based description tasks?	The impact of variation in planning was larger on the picture-based tasks. The largest number of idea units on the non-picture-based tasks was produced under the five-minute planning condition. The largest

	number of idea units on the picture-based tasks was produced under the ten-minute condition.
Which task type and planning condition has the largest impact on test scores and CAF results?	The largest effect size was observed on scores awarded on Task 4 (Baby task).

Continuation of table 63

Does variation in planning time operationalized as 30 seconds, one minute, five minutes and ten minutes impact the results of a language test when assessed with

a) an EBB scale

b) an analytic scale

c) measures of complexity, accuracy, and fluency?

The results of the multi-faceted Rasch measurement (MFRM) demonstrated that variation in planning time did have an impact on the test scores on both scales but the increases in measures of complexity, accuracy and fluency (CAF) that are consistently reported in the TBLT literature were not reproduced (see Section 2.2). Applying a Bonferroni correction to the interpretation of results, variation in planning time did not have a statistically significant impact on the test takers' language use in terms of CAF (see Section 6.3). However, the idea units analysis showed that increases in the amount of planning time (from 30 seconds and one minute to five minutes and ten minutes) did increase the quantity of speech that the test takers produced. This may have had a bearing upon the test scores, which increased when extra planning time featured as part of the task (see Section 6.2).

The absence of a statistically significant impact of the planning variable on CAF results is possibly a product of limitations in the participants' second language proficiency caused by lack of sufficient exposure to the language (see Section 2.4.1). In the literature, pre-task planning commonly increases the complexity, accuracy and fluency of speech produced by language learners that are resident in an English speaking country (e.g. Foster and Skehan, 1996) or have extensive experience of studying in an English-medium educational environment (e.g. Bui and Huang, 2016).

However, the current study investigated the effect of planning on the speech production of language learners that have little opportunity to develop competence in the second language. If appropriate language forms are not available for access during the planning stage, pre-task planning is unlikely to impact CAF (Kawauchi, 2005). The results showed that rather than using more complex, accurate or fluent language, the test takers were able to generate more content during planning and achieved significantly higher test scores (although the effect was marginal). In short, extra planning time before a test task did not affect the quality of the test takers' second language speech but seemed to help them to generate more ideas.

Which amount of planning time (30 seconds, one minute, five minutes, ten minutes) most substantially impacts test scores and CAF results?

In contrast to the general trends reported in the literature, the analysis indicated that when all test facets were included in the MFRM, the five-minute planning condition caused the largest increase in test taker scores (see Section 6.2.1). This was an unexpected result. In the TBLT literature, a period of ten minutes planning time has most consistently been shown to have a large impact on measures of CAF (see Section 2.2). It was therefore anticipated that the ten-minute planning condition would result in the largest increases in test scores. Similar predictions that ten minutes would have resulted in larger impacts on test taker language use have been made elsewhere in the literature where the anticipated effects of the planning variable were not observed (Elder and Iwashita, 2005, Iwashita et al., 2001, Li et al., 2014). In the current study, the ten-minute planning condition increased scores over the one-minute and 30-second planning conditions, however when all test facets were included in the MFRM, the five-minute condition resulted in the highest scores.

To hypothesise about the reason for this, when all test facets were included in the MFRM the difference between scores awarded under the five-minute and ten-minute conditions was marginal and did not reach statistical significance (see Section 6.2.1). Separate MFRM analyses of the test facets (task number and proficiency group) revealed that the ten-minute planning condition had resulted in the highest scores in some cases. For instance, whereas the B level test takers recorded the highest scores after the five-minute planning condition (the results did not reach statistical significance), the A level group recorded statistically significant increases in scores under the ten-minute planning condition in relation to the 30-second planning condition (see Section 6.2.2.3). In addition, on three of the tasks, the ten-minute planning condition resulted in the highest scores with large effect sizes (see Section 6.2.2.4). That these results became clear from MFRM analyses involving less diverse data (i.e. when the analysis involved one task or when test takers with similar levels of proficiency were assessed) indicates that test taker proficiency and task type are key variables in the outcome of pre-task planning for a speaking test.

Does the impact of the four planning conditions on test scores vary between the analytic scale and the EBB scale?

Comparisons between the analytic scale and EBB scale results indicated that the five-minute planning condition resulted in the highest fair average values on both scales (although the difference between scores awarded under the ten-minute and five-minute conditions did not reach statistical significance; see Section 6.2.2.1). On the EBB scale, the five-minute planning condition increased fair average values from level 2 (the fair average value under the 30-second and one-minute planning conditions) to level 3, indicating that planning had an important impact on raters' decisions regarding task completion (see Section 7.4.1). In contrast, the fair average values under each planning condition fell within the same level (level 9) on the analytic scale (see Section 7.4.2). However, the planning variable made important differences to individual test taker scores on the analytic scale and in some cases test takers were placed into three different levels according to which planning condition was involved (see Section 6.2.2.1).

In light of the CAF findings, there are two possible explanations for the increases in test scores after extra planning time a) increases in test scores may be associated with the increases in idea units b) test taker language improved in ways that were not captured by the CAF measures. Of the two possibilities, the former seems the most plausible. The CAF measures correspond very closely to the contents of the analytic scale, which describes complexity, accuracy and fluency at five levels of ability. It would be natural to assume that increases in test scores on the analytic scale would be reflected in at least some of the CAF measures. In contrast, the EBB scale refers to task success and this may be linked to the number of idea units the test taker produces (i.e. in order to successfully complete the task a minimum number of idea units may be required). A similar link between idea units and the analytic scale criteria cannot be made because the analytic scale only contains descriptors of complexity, accuracy and fluency, indicating that the analytic raters may have been influenced by features of the task performance that were not described in the scale.

This interpretation suggests the analytic scores reflect something that is irrelevant to the scale content (i.e. the number of idea units).

Does the impact of the four planning conditions on test scores and CAF results vary between groups of test takers who have different levels of language proficiency?

Test taker proficiency was shown to constitute an important variable in the study (see Section 6.2.2.3). The A level (CEFR levels A1 and A2) test takers recorded scores that were .39 higher on the logit scale under the ten-minute planning condition than under the 30-second planning condition. In contrast, increases in scores awarded to the B level (CEFR level B1) test takers after variation in planning time did not reach statistical significance.

My interpretation of this result was that A level test takers needed extra support to complete the tasks successfully. This support seems to have been provided by the opportunity to plan for ten minutes. However, the benefits of pre-task planning only reached statistical significance when test scores were compared between the minimal amount of planning time (30 seconds) and the maximum amount of planning time (ten minutes). This finding indicates that differences in the amount of planning time needed to be very large to affect the scores of the A level group. More generally, low ability language learners may struggle to successfully complete the task types investigated in this study after short amounts of planning time (e.g. 30 seconds). Providing a period of ten minutes to plan before such tasks may be a way to bias for best, but would also increase the time it takes to complete the test and stretch university resources. Clearly, further research in to the interaction between planning time and proficiency is required.

Does the impact of the four planning conditions on test scores and CAF results vary between picture-based narrative tasks and non-picture-based description tasks?

An important finding was that the largest gains in test scores after pre-task planning were recorded when test takers completed the picture-based narrative tasks (see Section 6.2.2.4). Scores on the non-picture-based description tasks also increased after variation in planning time. However, the effect size of the planning variable was not as large as observed on the picture-based tasks.

The differences in the results for the two task types may be because picture-based narratives, unlike the more open-ended non-picture-based description tasks involve obligatory content that poses specific difficulties for test takers who may not have sufficient language to describe the contents of the images (Skehan, 2009).

Based upon the analysis of the transcript samples (see Section 7.7), pre-task planning appears to affect performance on picture-based tasks in two ways. Firstly, the ten-minute and five-minute planning conditions may allow test takers to anticipate problems and limitations in their language knowledge that will affect their performance on these tasks. The planning time may thus be used to generate solutions to these problems in the form of 'achievement strategies' (Fulcher, 2003, p. 32). Secondly, increasing the amount of planning time provides the test taker with the opportunity to generate content about the tasks such as character motivation and descriptions of the surroundings. Ultimately, this elaboration of content seems to impact the raters' assessment of the test takers' language ability regardless of the rating scale criteria (see Section 7.4).

Which task type and planning condition has the largest impact on test scores and CAF results?

Performance on Task 4 (*Baby Task*; see Appendix 4) was impacted most substantially by the ten-minute planning condition (see Section 6.2.2.4). This task posed specific difficulties to the test takers, which were to some extent mitigated by the opportunity to plan (see Section 7.6 and 7.7). The source of the difficulty was related to key task features that the test takers might not have had the means to describe. These features included the need to generate language to describe the removal of a baby from a crib and the replacement of the baby with a ball, the cause and effect relationship between the mother noticing the baby was not in the crib and becoming frightened, the need to distinguish between characters that shared the same gender, sequencing of events with adverbial phrases, and providing motivations for the characters' actions. Analysis of transcripts shows that increased planning time may have helped test takers generate language to describe these features (see Section 7.7).

8.3 Implications and contributions

This section describes the implications and contributions this study has made to the field of language testing (Section 8.3.1), TBLT (8.3.2), and to our understanding of language learners and test takers (8.3.3). It emphasises the importance of the research for the various stakeholders including practitioners interested in the teaching and testing of second language speaking ability both in the Turkish educational context and more globally, and researchers with broader theoretical concerns about language learning and language testing.

8.3.1 Language Testing

The contributions this study has made to the field of language testing can be summarised as follows:

- Planning enhances validity
- Planning facilitates a bias for best
- Planning impacts on test scores most clearly on picture-based tasks
- EBB scales represent a viable alternative to traditional rating scales for collecting reliable information about test taker proficiency

Providing evidence of the context and cognitive elements of validity is an important component of establishing the overall validity of decisions based on test scores (Weir, 2005). When the purpose of a test is to determine a candidates' ability to study in an English-medium undergraduate environment, speech planning is a key behaviour in the target domain and should be included as part of the test (i.e. the test should demonstrate the context element of validity). This is important, to the extent possible under test conditions, to elicit from the test takers the same cognitive processes as are employed in the target language domain (i.e. the test should demonstrate the cognitive element of validity). The findings of this study show that test takers who struggle to complete language test tasks under minimal planning conditions (e.g. 30 seconds or one minute) may succeed when extra planning time (e.g. five minutes or ten minutes) is provided. Including extra planning time before a language task is therefore a way to elicit the best possible performance from test takers (Swain, 1985). To appropriately establish what prospective students can do in their undergraduate programs, and to make valid decisions relating to university admission, language tests must assess planned speech. Based upon the findings of the current study, test takers produce their best performance after planning for ten minutes for picture-based tasks and after planning for five minutes for non-picture-based tasks that are based on familiar information.

The association between the increases in idea units and gains in scores indicates that raters regard the number of ideas test takers produce during the test as a key element of the test construct. Therefore, the rating scale should reflect this aspect of test performance by including speech quantity in the descriptors (rating scales are discussed at length later in this section). In addition, test takers should be encouraged to expand upon their responses during the speaking section of the test by providing richer description of the task content in the form of details and examples. An anticipated effect of this is that a) scores would become more representative of what test takers are capable of in L2 speech and b) students are encouraged to develop their speaking skills to the extent that they are able to express themselves in the L2 and hence derive more benefits from undergraduate study.

An important finding with implications for the development of language test tasks was that scores on the picture-based narrative tasks were most impacted by the planning variable. Generalising from this result, it may be the case that picture-based language tasks require additional processing that can be completed during planning. This applies to the picture-based narratives used in this study but may also be relevant for graph description tasks (Xi, 2010) and map description tasks (Crookes, 1989). Such tasks typically involve obligatory content, which the test taker may not have the linguistic means to describe but nevertheless must be communicated to successfully complete the task (Skehan, 2009). Planning time may be used to generate achievement strategies such as circumlocution or approximation to compensate for the absence of relevant language forms in the test taker's repertoire to discuss such obligatory content. Strategic competence is a key aspect of language proficiency (Fulcher, 2003) and planning time is likely to improve test takers' ability to apply this competence during the completion of picture-based language tasks. In short, where picture-based tasks are used to gather information about language ability, test scores are likely to be higher if the test involves a period of pre-task planning.

This study found that planning for five minutes before a non-picture-based language task consistently increased test scores. Reservations about the potential impact on test practicality of including large amounts of planning have prevented researchers from investigated periods of time in excess of ten minutes (Li et al., 2014). Furthermore, empirical evidence shows that planning for periods of one minute, two minutes, and three minutes does not affect test performance (Wigglesworth, 1997, Iwashita et al., 2001, Elder et al., 2002, Elder and Wigglesworth, 2006). However, the findings of this study indicate that five minutes may be sufficient time to 'bias for best' (Swain, 1985, p. 42) when the test involves a non-picture-based speaking task.

This finding has implications beyond the Turkish higher educational context. Language tests that feature non-picture-based tasks as part of the assessment may include a five-minute pre-task planning condition to elicit the best performance from test takers. However, an important caveat is that language proficiency is a key variable and the evidence suggests that low-level test takers may require ten minutes to benefit from planning time (see Section 7.5). Test developers should account for the potential interaction between test taker proficiency, task type and pre-task planning time by trialling different amounts of planning time on different tasks with representative members of the test taking population before including planning in their language tests. Further research will be required to establish appropriate lengths of planning time for different educational contexts.

An important finding was that the empirically derived EBB scale better reflected the variation in test performance after pre-task planning than the general-purpose, intuitively derived analytic scale. Because of the nature of the scale, which did not include a category for task completion, variation between test scores on the analytic scale could not be linked to increases in the number of idea units after extra pre-task planning time. An implication of this finding is that raters seem to have been influenced by features of the task performance that were not referred to in the analytic scale contents (see Section 7.4). Adherence to the rating scale is a critical component of the scoring element of validity (Weir, 2005). This study therefore underscores the importance for rating scales to reflect rater criteria. By applying Turner and Upshur's (1996) EBB method, this study demonstrates the potential for assessor-oriented, empirically derived rating scales to improve the scoring element of validity in second language speaking tests. Based on this finding, it is recommended that the EBB method be used in the university admissions test. Beyond this, the EBB method may be considered in similar foreign language contexts to collect reliable information about test takers' L2 ability for purposes of language assessment (Ducasse, 2009).

8.3.2 TBLT

The contributions this study has made to the field of language testing can be summarised as follows:

- Conservative statistical analytical procedures demonstrate that the impact of pre-task planning on complexity, accuracy and fluency is minimal.
- Participant proficiency in the L2 is a key variable in planning effects.
- The absence of relevant language to discuss obligatory task content may determine the extent to which language learners find an L2 task challenging.

In the field of TBLT, evidence for the positive impact of planning on second language speech production (measured in terms of complexity, accuracy and fluency) is pervasive. In the face of such abundance of empirical research, it is generally taken for granted that planning improves task performance in both TBLT and language testing: 'if we add it, performance improves; remove it or reduce it, and performance worsens' (O'Sullivan, 2012, p. 235). However, the application of statistical procedures to reduce the chances of committing a type one error shows that the impact of planning on measures of CAF in this educational context is minimal (see Section 6.3). Adopting a conservative approach to the interpretation of statistical significance for multiple statistical tests, this study demonstrates that pre-task planning does not affect task performance to the extent that is reported in the literature.

An important characteristic of the educational context in which this study is situated is the absence of opportunities to develop spoken proficiency in L2 English (see Section 2.4.1). The test takers in this study lacked ability as speakers of English and pre-task planning did not affect the complexity, accuracy or fluency of their speech. In this regard, the findings of this study build upon research into the interaction between language proficiency and planning time (Kawauchi, 2005, Mochizuki and Ortega, 2008). Overall, this research suggests that in order for planning to have an optimal effect on language performance, test takers must have access to and make effective use of sources of L2 knowledge during the planning stage. With low-level test takers, such knowledge has not been acquired and support may be required in order for them to make the best use of the planning time. In TBLT contexts, this may come in the form of focussed planning of target structures that would facilitate successful completion of the task (Sangarun, 2005).

Another important contribution of this research is to the concept of task challenge. Research into the interaction between task challenge and pre-task planning indicates that the more challenging a task, the more benefits can be derived from planning (see Section 2.5.1.4). However, the source of task challenge is notoriously elusive and is likely to vary between language learners (see Sections 2.5.1.2 and 2.5.1.3). In this study, the picture-based narrative tasks, which effectively obliged test takers to refer to certain objects and situations, were more challenging than tasks that required the generation of task content by test takers. This was probably because the test takers may not have had recourse to appropriate language structures to communicate the obligatory information in the picture tasks. For this reason, test takers made better use of the opportunity to plan when carrying out the picture-based tasks than when carrying out the non-picture-based tasks. In sum, an important factor that contributes to task difficulty is the absence of appropriate language in the test takers' repertoire to describe obligatory task content. Planning mitigates this difficulty and increases test taker scores on picture-based tasks.

8.3.3 Language learners/test takers

This study has generated important findings about the interaction between second language proficiency and pre-task planning. The findings suggest that low-level test takers in this context do not register statistically significant gains in language use (CAF) after extra pre-task planning time because they may have not acquired the appropriate structures. This finding contradicts many of the accounts of the beneficial impact of planning made in both the language testing literature (Li et al., 2014, Tavakoli and Skehan, 2005) and the TBLT literature (Bui and Huang, 2016, Foster and Skehan, 1996, 1999, Skehan and Foster, 1997, 2005, Yuan and Ellis, 2003). Low ability test takers require support to successfully complete the kind of tasks that were investigated in this study. Planning partially provides this support by permitting the test taker extra time to consider task content and to generate strategies to avoid potentially awkward processes of linguistic encoding. However, this does not have a statistically significant bearing on the complexity, accuracy or fluency of their language use. Contemporary accounts of the benefits of pre-task planning on task performance (Ellis, 2005, 2009, Robinson, 2005, Skehan, 2016) should therefore be revised to incorporate limitations in second language proficiency.

8.4 Limitations

Although this research presents clear conclusions regarding the impact of pre-task planning on test performance, a number of limitations need to be acknowledged. The first limitation is the sample size. The study examined speech samples produced by 47 language learners (the total number of samples was 188). These learners were divided in to two proficiency levels based on the results of the QPT (UCLES, 2001): an A level group (comprising both A1 and A2 levels on the CEFR) and a B level group (B1 on the CEFR). However the A level (*n*=28) group substantially outnumbered the B level group (*n*=12) and seven of the test takers did not arrive in time to sit the proficiency test (see Section 5.4.1). Establishing a balance between the test taker profiles by increasing the number of B level test takers would have made for a stronger comparison between the groups and hence more reliable results concerning the interaction between pre-task planning and proficiency.

Collecting further data from very proficient (C level) learners of English and comparing the results of the planning variable between these groups might have revealed more about the potential that planning has to impact task performance. On the other hand, for clear conclusions to be drawn about the impact of planning on the results of the university admissions test it was important for the research to be relevant to the specific educational context. As the test taking population in this context does not generally contain test taker profiles that have C level proficiency, examining the impact of planning across such a wide range of test taker profiles was not appropriate. Nevertheless, the literature review suggests that for planning to have a meaningful impact on the test scores, the participants must have attained a level of language ability where they have access to extensive sources of L2 knowledge (see Section 2.6.2). If this is the case, the impact of planning is likely to vary substantially between test takers.

The research provides empirical evidence of the impact of pre-task planning on the results of a university admission test for English-medium education in Turkey. The study controls for potentially important variables in the assessment process such as L1 and age, and provides important findings for the research context. While this level of specificity can be regarded as an advantage of the study, an effect is that the findings are restricted to a specific educational environment. All participants came from similar educational backgrounds and shared similar experiences of learning the second language. In addition, the participants were volunteers and are thus not a random sample of the test taking population. Ultimately, this aspect of the study limits the generalizability of the results to other educational contexts where English-medium education is also common. The literature review indicates that variation in the experiences and abilities that the test taking population brings to the test may play an important role in the results of the planning variable (see Section 2.4.1). When the opportunities are limited for test takers to develop spoken language proficiency, the planning variable makes relatively little difference to CAF and test scores. Establishing a threshold level of proficiency above which the benefits of planning become available is difficult given the similar levels of ability between test takers in this study and the lack of a systematic approach to the measurement of participant ability according to common criteria in the literature (see Section 2.6.3). Further research is required to closely examine the interaction between planning and proficiency.

8.5 Further research

8.5.1 Test takers

The findings of this study point to various areas with potential for future research. The first area relates to test takers' thought processes during the planning stage. Planning made a statistically significant increase to the number of idea units produced. This finding suggests that the test takers used the planning time to generate task content. Furthermore, examination of the transcripts suggested that test takers may have used the planning stage to anticipate problems they might encounter when describing the picture-based narratives and generated solutions. However, these interpretations rely upon indirect evidence from analysis of the transcripts and CAF results. Various questions remain unanswered. For instance, the findings of this study indicate that planning for picture-based narratives has more of an impact on test scores than planning for non-picture-based tasks. Future research may explore whether the planning processes are different when the task is picture-based and involves obligatory content in relation to tasks that require the test taker to generate their own content. During the planning stage for a picture-based narrative, does the test taker anticipate problems with linguistic encoding of the task content and generate solutions? Is the planning time used to generate 'achievement strategies'

(Fulcher, 2003, p. 32)? Do the achievement strategies that test takers employ differ when planning time is included as part of the test task?

Another related area is the impact of test taker proficiency on the planning process. Do the thought processes that test takers engage in during planning vary according to proficiency? Do low-level test takers concentrate specifically on content while higher-level test takers focus on language? The interaction between language proficiency and planning time has been emphasised throughout this thesis and the evidence (i.e. that the A level group recorded statistically significant improvements in test scores after planning, whereas the B level group did not; see Sections 6.2.2.3 and 7.5) suggests that test takers respond differently to the planning variable depending on their level of proficiency in the second language.

The questions identified above may be answered through a form of verbal protocol such as stimulated recall. For future studies, completing verbal protocol with the test takers might provide key insights in to the planning processes of the test taking population and uncover how the planning variable impacts their approach to the task. Such research would make a valuable contribution to our understanding of the role that planning plays in task-based language assessment, which may be used as validity evidence for language assessment procedures.

8.5.2 Raters and rating scales

A second important area of research is the thought processes raters go through when assessing the test samples. Conformity to the rating scale content has been shown to constitute a key theme in this research. Regardless of the scale content, raters were assumed to have been influenced by the increases in idea units instigated by extra planning time (see Sections 7.2, 7.6, 7.7 and 7.8). However, without evidence in the form of rater accounts, the interpretation of this result is dependent upon analysis of the CAF measures and the test taker transcripts. Once again, the results of this particular aspect of the study raise various questions. Regarding conformity to the scale, a critical question seems to be, do raters base their assessment on criteria made explicit in the scale? If not, what specific features of the speech do raters hone in on during rating?

Another critical question relates to the impact of the EBB, binary choice rating scale on raters' thought processes. How do approaches to assessment differ when raters use the EBB scale in relation to the analytic scale? For example, raters are encouraged to listen for gradations of proficiency in certain aspects of speech when using the analytic scale, whereas the EBB scale promotes a much more holistic, all or nothing assessment. However, it is unclear whether the raters use the EBB scale in the intended way. Do the raters legitimately consider the questions in the correct order or do they treat the scale as a series of numbers indicating general variation in ability rather than binary decisions about specific constructs? Future research might investigate these issues with verbal protocol analysis in which raters verbalise their thought processes when awarding scores to speech samples. Although Lumley (2005) demonstrates that scoring is essentially an idiosyncratic process, the success of the EBB scale in the current study suggests that raters do agree about second language ability to the extent that they can collaborate to develop effective rating scales and apply them with high reliability. Such research has the potential to uncover information that could be used to refine our understanding of the way in which planning impacts raters' assessment of second language ability. More generally, research in this area might be used to develop existing rating scales so that they are more representative of the rating population and hence enhance the validity of assessment scores.

8.5.3 Context and cognitive orientation

An interesting area for potential research that this study has not explored is the impact of the test environment upon the language learners' cognitive orientation (i.e. focussing on and prioritising aspects of their task performance) when planning for the language task. Ellis (2005) has posited a focus on form in the testing context that is absent in low-anxiety, consequence-free classroom environments. One interpretation of Ellis' hypothesis is that the same task would be completed differently in a classroom environment and an assessment environment. In the assessment environment, test takers are concerned with the accuracy of their performance and are unlikely to take risks in their language use. This conservatism would likely influence the test takers' processes during the planning condition, perhaps by forcing them into planning structures with which they feel relatively confident in producing. However, without the threat of any meaningful consequences for language misuse (e.g. not being admitted to a course of undergraduate study), test takers would be free to experiment with language that is not well rehearsed. In non-assessment contexts, the planning variable might foster meaningful improvements in language use, specifically with regards to language complexity (Ellis, 2005). Ultimately, this strand of research would have the potential to account for the discrepancy between accounts of pre-task planning in the TBLT and language testing literature (see Section 2.2). Such an investigation would require quantitative methodology and follow a similar analytical framework as the one employed in this study (i.e. by comparing the effect of pre-task planning on CAF and rater scores between the two contexts). However, qualitative methods such as stimulated recall would also be required to examine the extent to which the pressure to focus on form in the assessment context is a factor.

8.6 Final Comments

This study has presented important findings for the field of second language learning and assessment. It has argued that a period of pre-task planning time is a necessary component to demonstrate validity in a speaking test that assesses the ability to study in an English-medium, tertiary environment. The research underscores the importance of accounting for contextual factors in language testing research. This is demonstrated in the interaction between test taker characteristics (specifically their levels of proficiency in the second language), task type and planning conditions, and the approach to the measurement of test taker performance (specifically the underlying principles supporting the development and use of rating scales).

Contrary to the trends reported in the literature, this study indicates that the complexity, accuracy and fluency of test taker speech does not improve after planning when the test taking population lacks proficiency in the second language and conservative statistical procedures are applied to the analysis of results. However, the planning variable does facilitate the generation of task content and improves the test takers’ chances of obtaining higher scores. This suggests that planning time is a way for test developers to bias for best. The rating scale plays a major role in this. The context specific, rater oriented EBB scale was better able to describe the impact of planning because of explicit reference to task achievement. This was a task feature that raters regarded as particularly salient to their assessment (see Section 5.4.4). In contrast, the general-purpose, intuitively derived analytic scale underrepresented the test construct of second language spoken ability and test takers recorded gains in scores that could not be accounted for by the scale content.

The research has contributed to our understanding of the way that planning impacts test performance while emphasising the importance of context in language testing research. It has offered clear directions for future research and presented implications for researchers and practitioners in the field. More applicably, important recommendations for the university admissions test have been outlined. These recommendations have the potential to increase the validity of the assessment and enhance the likelihood that test takers demonstrate their best possible performance on the test.

References

Alderson, J. (2007). The CEFR and the need for more research. *The Modern Language Journal*, *91*, 659–663.

Audacity Team (2014). Audacity(R): Free Audio Editor and Recorder (Version 2.0.6) [Computer program]. Retrieved from http://audacity.sourceforge.net/

Bachman, L. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453-476.

Bachman, L. (2004). *Statistical Analyses for Language Assessment*. Cambridge: CUP.

Baddeley, A. (2007). *Working Memory, Thought, and Action*. Oxford: OUP.

BBC. (2014, June 14). *Tony Blair: We didn't cause Iraq Crisis*. Retrieved from http://www.bbc.com/news/uk-27852832

Brindley, G., Hood, S., McNaught, C., & Wigglesworth, G. (1997). Issues in test design and delivery. In G. Brindley, & G. Wigglesworth (Eds.), *access: Issues in Language Test Design and Delivery* (pp. 31-65). Sydney: National Centre for English Language Teaching and Research.

Brown, A. (2006). An examination of the rating process in the revised IELTS Speaking Test. *IELTS Research Reports* (Vol. 6, pp. 1-30). Canberra: IELTS and British Council.

Brown, A., Iwashita, N., & McNamara, T. (2005). *An examination of rater orientations and test-taker performance on English-for-academic- purposes speaking tasks* (TOEFL Monograph No. MS-29). Princeton, NJ: Educational Testing Service.

Brown, J. (1990). The use of multiple t tests in language research. *TESOL Quarterly*, *24*(4), 770-773.

Bui, G., & Huang, Z. (2016). L2 fluency as influenced by content familiarity and planning: performance, measurement and pedagogy. *Language Teaching Research*. Advance online publication. doi:10.1177/1362168816656650

BWC EXTRAS. (2012, March 7). *Bubba Gnomes,!?!, & FARCED*. Retrieved from https://bwcdigital.wordpress.com/tag/wordless/

Cepik, S., & Polat, N. (2014). Mandates, needs, equitable resources, and current research in English language teacher education: The case of Turkey. *International Journal of Research Studies in Education*, *3*(2), 83-96.

Cobb, T. *Web Vocabprofile* [Computer program]. Retrieved from http://www.lextutor.ca/vp/

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.

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

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition*, *11*(4), 367-383.

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing*, *33*(1), 117-135.

Dickens, C. (1955). *A Tale of Two Cities*. Letchworth: Aldine Press.

Ducasse, A. (2009). Raters as scale makers for an L2 Spanish speaking test: Using paired discourse to develop a rating scale for communicative interaction. In A. Brown & K. Hill (Eds.), *Tasks and Criteria in Performance Assessment* (pp. 15-39). Frankfurt: Peter Lang.

Elder, C., & Iwashita, N. (2005). Planning for test performance: does it make a difference?. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 219-239). Amsterdam: John Benjamins.

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*(4), 347-368.

Elder, C., & Wigglesworth, G. (2006). An investigation of the effectiveness and validity of planning time in part 2 of the IELTS speaking test. *IELTS Research Reports* (Vol. 6, pp. 1-28). Canberra: IELTS Australia and British Council.

Ellis, R. (1987). Interlanguage variability in narrative discourse: styleshifting in the use of the past tense. *Studies in Second language Acquisition, 9*, 1-20.

Ellis, R. (2005). Planning and task based performance. Theory and research. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 3-36). Amsterdam: John Benjamins.

Ellis, R. (2009). The differential effects of three types of task planning on the fluency, complexity and accuracy in L2 oral production. *Applied Linguistics*, *30*(4), 474-509.

Ellis, R., & Barkhuizen, G. (2005). *Analysing Learner Language*. Oxford: OUP.

Field, J. (2011). Cognitive validity. In L. Taylor (Ed.), *Studies in Language Testing 30 Examining Speaking* (pp. 65-112). Cambridge: CUP.

Foster, P., & Skehan, P. (1996). The influence of planning time on performance in task-based learning. *Studies in Second Language Acquisition, 18*, 299-234.

Foster, P., & Skehan, P. (1999). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, *3,* 215-247.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: a unit for all reasons. *Applied Linguistics*, *21*(3), 354-375.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: the case for a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*, 98-116.

Frost, K., Elder, C., & Wigglesworth, G. (2011). Investigating the validity of an integrated listening-speaking task: A discourse based analysis of test-takers' oral performances. *Language Testing*, *29*(3), 345-369.

Fulcher, G. (2003). *Testing Second Language Speaking*. London: Pearson.

Fulcher, G. (2012). Scoring performance tests. In G. Fulcher, & F. Davidson, (Eds.), *The Routledge Handbook of Language Testing* (pp. 378-392). London: Routledge.

Fulcher, G. (2015). *Re-examining Language Testing A Philosophical and Social Inquiry*. London: Routledge.

Fulcher, G., & Davidson, F. (2007). *Language Testing and Assessment*. London: Routledge.

Fulcher, G., Davidson, F., & Kemp, J. (2011). Effective rating scale development for speaking tests. *Language Testing*, *28*(1), 5-29.

Fulcher, G., & Marquez Reiter, R. (2003). Task difficulty in speaking tasks. *Language Testing, 20*(3), 321-344.

Genc, Z. (2012). Effects of strategic planning on the accuracy of oral and written tasks in the performance of Turkish EFL learners. In A. Shehadeh, & C. Coombe, (Eds.), *Task-based Language Teaching in Foreign Language Contexts Research and Implementation* (pp. 67-89). Amsterdam: John Benjamins.

Geng, X., & Ferguson, G. (2013). Strategic planning in task based language teaching: the effects of participatory structure and task type. *System*, *41*, 982-993.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and (+/- here and now): effects on oral production. In M. Mayo (Ed.), *Investigating Tasks in Formal Language Learning* (pp. 44-68). Bristol: Multilingual Matters.

Green, A. (2014). *Exploring Language Assessment and Testing: Language in Action*. London: Routledge.

Guara-Tavares, M. (2009). The relationship among pre-task planning, working memory capacity, and L2 speech performance: a pilot study. *Linguagem & Ensino*, *12*(1), 165-194.

Harding, L. (2016). What do raters need in a pronunciation scale?: the users' view. In T. Isaacs, & P. Trofimovich (Eds.), *Second Language Pronunciation Assessment: Interdisciplinary Perspectives*, (pp. 12-28). Bristol: Multilingual Matters.

Hirai, A., & Koizumi, R. (2013). Validation of empirically derived rating scales for a story retelling task. *Language Assessment Quarterly*, *10*, 398-422.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics, 34*(4), 1-13.

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: definitions, measurement and research. In A. Housen, F. Kuiken & I. Vedder (Eds.), *Dimensions of L2 Performance and Proficiency* (pp. 1-20). Amsterdam: John Benjamins.

Huang, L. (2013). *Cognitive processes involved in performing the IELTS speaking test: respondents' strategic behaviours in simulated testing and non-testing contexts* (IELTS Research Report Online Series No. 1). Retrieved from https://www.ielts.org/-/media/research-reports/ielts_online_rr_2013-1.ashx

Hulstijn, J. (2007). The shaky ground beneath the CEFR: quantitative and qualitative dimensions of language proficiency. *The Modern Language Journal*, *91*(4), 663-667.

Hymes, D.H. (1972). On communicative competence. In J.B Pride & J. Holmes (Eds.), *Sociolinguistics. Selected Readings* (pp. 269-293). Harmondsworth: Penguin.

Inoue, C. (2013). *Task Equivalence in Speaking Tasks*. Berlin: Peter Lang.

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.

Jaeger, T. (2007). Categorical data analysis: Away from ANOVAs (transformation or not) and toward logit mixed models. *Journal of Memory and Language*, *59*(4), 434-446.

Jefferson, G. (2004). Glossary of transcript symbols with an introduction. In G. H. Lerner (Ed.), *Conversation Analysis: Studies from the First Generation* (pp. 13-31). Amsterdam: John Benjamins.

Kawauchi, C. (2005). The effects of strategic planning on the oral narratives of learners with low and high intermediate L2 proficiency. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 143-165). Amsterdam: John Benjamins.

Khabbazbashi, N. (2017). Topic and background knowledge effects on performance in speaking assessment. *Language Testing*, *34*(1), 23-48.

Kirkgoz, Y. (2009). Students' and lecturers' perceptions of the effectiveness of foreign language instruction in an English-medium university in Turkey. *Teaching in Higher Education*, *14*(1), 81-93.

Kormos, J. (2006). *Speech Production and Second Language Acquisition*. New Jersey: Lawrence Erlbaum.

Kormos, J., & Denes. M. (2004). Exploring measures and perceptions of fluency in speech of second language learners. *System*, *32*, 145-164.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics in Language Teaching*, *45*(3), 261-284.

Kuiken, F., & Vedder, I. (2014). Rating written performance: what do raters do and why? *Language Testing*, *31*(3), 329-348.

Kuiken F., & Vedder, I. (2017). Functional adequacy in L2 writing: towards a new rating scale. *Language Testing*, *34*(3), 321-336.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy and fluency in task-based research: toward more developmentally based measures of second language acquisition. *Applied Linguistics*, *35*(5), 607-614.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307-322.

Laufer, B., Elder, C. Hill, K. & Congdon, P. (2004). Size and strength: do we need both to measure vocabulary knowledge? *Language Testing*, *21*(2), 202-226.

Levelt, M. (1989). *Speaking: from Intention to Articulation*. Cambridge: MIT Press.

Li, L. Chen, J. & Sun, L. (2014). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly*, *49*(1), 38-66.

Linacre, J. (1989). *Multi-faceted Rasch Measurement*. Chicago: MESA Press.

Linacre, J. (1993). Rasch-based generalizability theory. *Rasch Measurement Transactions*, *7*(1), 283-284.

Linacre J. (2010). When to stop removing items and persons in Rasch misfit analysis? *Rasch Measurement Transactions*, *23*(4), 1241.

Linacre, J. (2013). A User's Guide to FACETS Rasch-Model Computer Programs. Retrieved from http://www.winsteps.com/winman/copyright.htm

Linacre, J. (2017). Facets Computer program for many-faceted Rasch measurement (Version 3.71.4) [Computer program]. Beaverton, Oregon: Winsteps.com

Lumley, T. (2005). *Assessing Second Language Writing*. Frankfurt: Peter Lang.

Lumley, T., & O'Sullivan, B. (2005). The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking. *Language Testing*, *22*(4), 415-437.

Lunz, M., & Stahl, J. (1990). Judge consistency and severity across grading periods. *Evaluation and the Health Professions*, *13*(14), 425-444.

Luoma, S. (2004). *Assessing Speaking*. Cambridge: CUP.

May, L. (2006). An examination of rater orientations on a paired candidate discussion task through stimulated verbal recall. *Melbourne Papers in Language Testing*, *11*(1), 29-51.

McCarthy, P., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381-392.

McNamara, T. (1996). *Measuring Second Language Performance*. London: Longman.

Merrylees, B. (2003). An impact study of two IELTS user groups: candidates who sit the test for immigration purposes and candidates who sit the test for secondary education purposes. *IELTS Research Reports* (Vol. 4, pp. 1-58). Canberra: IELTS Australia.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational Measurement* (pp. 13-103). New York: Macmillan.

Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning-level foreign language classrooms: A study of guided planning and relativisation. *Language Teaching Research*, *12*(1), 11-37.

Myford, C., & Wolfe, E. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: part II. *Journal of Applied Measurement*, *5*(2), 189-227.

Nielson, K. (2013). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research*, *18*(3), 272-293.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on oral task performance. *Language Testing*, *31*(2), 147-175.

Norris, J., & 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. (2007). The CEFR illustrative descriptor scales. *The Modern Language Journal*, *91*(4), 656-659.

Orr, M. (2002). The FCE speaking test: using rater reports to help interpret test scores. *System*, *30*(2), 143-154.

O'Sullivan, B. (2000a). *Towards a model of performance in oral language testing*. (Unpublished PhD thesis). University of Reading, UK.

O'Sullivan, B. (2000b). Exploring gender and oral proficiency interview performance. *System, 28*(3), 373-386.

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing*, *19*(3), 277-295.

O'Sullivan, B. (2012). Assessing speaking. In C. Coombe, P. Davidson, B. O'Sullivan & S. Stoynoff (Eds.), *The Cambridge Guide to Second Language Assessment* (pp. 234-247). Cambridge: CUP.

O'Sullivan, B. (2016). Validity: what is it and who is it for?. In Y. Leung (Ed.), *Epoch Making in English Teaching and Learning: Evolution, Innovation, and Revolution* (pp. 201-222). Taipei: Crane Publishing Company Ltd.

O'Sullivan, B., & Green, A. (2011). Test taker characteristics. In L. Taylor (Ed.), *Studies in Language Testing 30 Examining Speaking* (pp. 37-64). Cambridge: CUP.

O'Sullivan, B., & Weir, C. (2011). Language testing and validation. In B. O'Sullivan (Ed.), *Language Testing Theory and Practice* (pp. 13-32). Oxford: Palgrave.

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

Panchin, A., & Tuzhikov, A. (2017). Published GMO studies find no evidence of harm when corrected for multiple comparisons. *Critical Reviews of Biotechnology*, *37*(2), 213-217.

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language reporting in narrative retelling. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (pp. 95-128). Amsterdam: John Benjamins.

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

Porter, D. (1991). Affective factors in the assessment of oral interaction: gender and status. In S. Arnivan (Ed.), *Current Developments in Language Testing* (pp. 92-102). Singapore: SEAMEO Regional Language Centre.

Purpura, J. (2016). Second and foreign language assessment. *The Modern Language Journal*, *100*, 190-208.

Robinson, P. (2005). Cognitive complexity and task sequencing: studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*, 1-32.

Sangarun, J. (2005). The effects of focusing on meaning and form in strategic planning. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 111-143). Amsterdam: John Benjamins.

Sasayama, S., & Izumi, S. (2012). Effects of task complexity and pre-task planning on Japanese EFL learners' oral production. In A. Shehadeh, & C. Coombe (Eds.), *Task-Based Language Teaching in Foreign Language Contexts Research and Implementation* (pp. 23-43). Amsterdam: John Benjamins.

Sato, T. (2012). The contribution of test-takers' speech content to scores on an English oral proficiency test. *Language Testing*, *29*(2), 223-241.

Selvi, A. (2014). The medium-of-instruction debate in Turkey: oscillating between national ideas and bilingual ideals. *Current Issues in Language Planning*, *15*(2), 133-152.

Siegel, F. (1990). Multiple t tests: some practical considerations. *TESOL Quarterly*, *24*(4), 773-775.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: OUP.

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

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (pp. 1-27). Amsterdam: John Benjamins.

Skehan, P. (2016). Tasks versus conditions: two perspectives on task research and their implications for pedagogy. *Annual Review of Applied Linguistics*, *36*, 34-49.

Skehan, P., & Foster, P. (1997). Task type and processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-211.

Skehan, P., & Foster, P. (2005). Strategic and on-line planning: the influence of surprise information and task time on second language performance. In R. Ellis (Ed.), *Planning and Task-Performance in a Second Language* (pp. 193-219). Amsterdam: John Benjamins.

Swain, M. (1985). Large scale communicative testing: A case study. In Y. Lee, C. Fok, R. Lord & G. Low (Eds.), *New Directions in Language Testing* (pp. 35-46). Hong Kong: Pergamon Press.

Tabachnick, B., & Fidell, L. (2007). *Using Multivariate Statistics*. Boston: Pearson/Allyn and Bacon.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis, (Ed.), *Planning and Task Performance in a Second Language* (pp. 239-277). Amsterdam: John Benjamins.

Taylor, L. (2011). Introduction. In L. Taylor (Ed.), *Studies in Language Testing 30 Examining Speaking* (1-36). Cambridge: CUP.

Turner, C., & Upshur, J. (1996). Developing rating scales for the assessment of second language performance. *Australian Review of Applied Linguistics*, *13*, 55-79.

Turner, C., & Upshur, J. (2002). Rating scales derived from student samples: effects of the scale maker and the student sample on scale content and student scores. *TESOL Quarterly*, *36*(1), 49-70.

UCLES. (2001). *Quick Placement Test*. Oxford: OUP.

Upshur, J., & Turner, C. (1995). Constructing rating scales for second language tests. *ELT Journal*, *49*(1), 3-12.

Van Voorhis, C., & Morgan, B. (2007). Understanding Power and Rules of Thumb for Determining Sample Sizes. *Tutorials for Quantitative Methods for Psychology*, *3*(2), 43-50.

Weir, C. (1993). *Understanding and Developing Language Tests*. New York: Prentice Hall.

Weir, C. (2005). *Language Testing and Validation*. Basingstoke: Palgrave Macmillan.

Weir, C., O'Sullivan, B., & Horai, T. (2006). Exploring difficulty in speaking tasks: an intra-task perspective. *IELTS Research Reports* (Vol. 6, pp. 1-42). Canberra: IELTS Australia and British Council.

Weir, C., & Taylor, L. (2011). Conclusions and recommendations. In L. Taylor (Ed.), *Studies in Language Testing 30 Examining Speaking* (pp. 293-313). Cambridge: CUP.

Welch, B. (1951). On the Comparison of Several Mean Values: An Alternative Approach. *Biometrika*, *38*(3/4), 330-336.

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

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly*, *7*(1), 1-24.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing*, *30*(2), 231-252.

Wisniewski, K. (2017). Empirical learner language and the levels of the Common European Framework of Reference. *Language Learning*, *67*, 232-253.

Xi, X. (2005). Do visual chunks and planning impact the overall quality of oral descriptions of graphs?. *Language Testing*, *22*(4), 463-508.

Xi, X. (2010). Aspects of performance on line graph description tasks: influenced by graph familiarity and different task features. *Language Testing*, *27*(1), 73-100.

Yerkes, R. M. (1920). What psychology contributed to the war. In R. M. Yerkes (Ed.), *The New World of Science: its development during the war* (pp. 364-389). New York, NY: The Century Co.

Yuan, F., & Ellis, R. (2003). The effects on pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied Linguistics*, *21*(1), 1-27.

Appendices

Appendix 1. Analytic Rating Scale

Fluency

5. Speaks without hesitation; speech is generally of a speed similar to a native speaker.

4. Speaks fairly fluently with only occasional hesitation, false starts and modification of intended utterance. Speech is only slightly slower than that of a native speaker.

3. Speaks more slowly than a native speaker due to hesitations and word finding delays.

2. A marked degree of hesitation due to word finding delays or inability to phrase utterances easily.

1. Speech is quite disfluent due to frequent and lengthy hesitations or false starts.

Accuracy

5. Errors are barely noticeable.

4. Errors are not unusual, but rarely major.

3. Manages most common forms, with occasional errors; major errors present.

2. Limited linguistic control; major errors frequent.

1. Clear lack of linguistic control even of basic forms.

Complexity

5. Confidently attempts a variety of verb forms (e.g., passives, modals, tense and aspect), even if the use is not always correct. Regularly takes risks grammatically in the service of expressing complex meaning. Routinely attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is occasionally awkward or incorrect.

4. Confidently attempts a variety of verb forms (e.g., passives, modals, tense and aspect), even if the use is not always correct. Takes risks grammatically in the service of expressing complex meaning. Regularly attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is occasionally awkward or incorrect.

3. Mostly relies on simple verb forms, with some attempts to use a greater variety of forms (e.g. passives, modals, more varied tense and aspect). Some attempt to use coordination and subordination to convey ideas that cannot be expressed in a single clause.

2. Produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty.

1. Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect ideas across clauses.

Appendix 2. Consent Letter

University of Bedfordshire
Centre for Research in English Language Learning and Assessment
Putteridge Bury, Hitchin Road
Luton, UK

The purpose of this study is to examine the effects of introducing time planning to a task of second language speaking. The study is part of Stefan O'Grady's MPhil/PhD in English language assessment, under the supervision of Professor Tony Green.

Your identity will be kept completely confidential and your name will not be connected to any of the information include in this study, rather a number will be used for purposes of identification. Information linking your name to the study will never be included in any report or publication. The data you provide will be accessible to people working on the study only.

Statement of Consent:

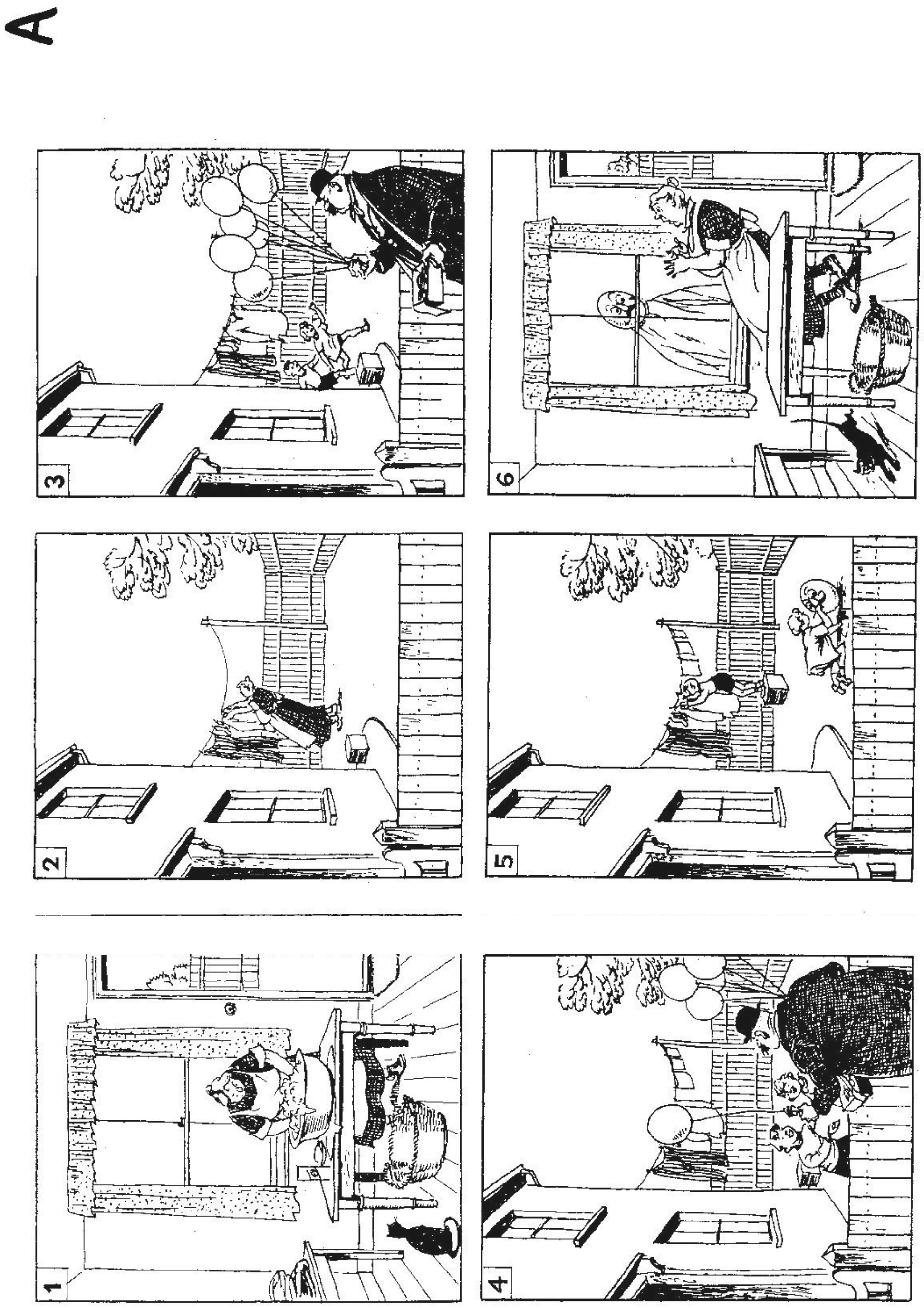
I have read the above information. My questions about the study have been satisfactorily answered and I agree to participate in the study.

Name of participant (PRINT) ______________________________ Date: ________________

Signature of participant____

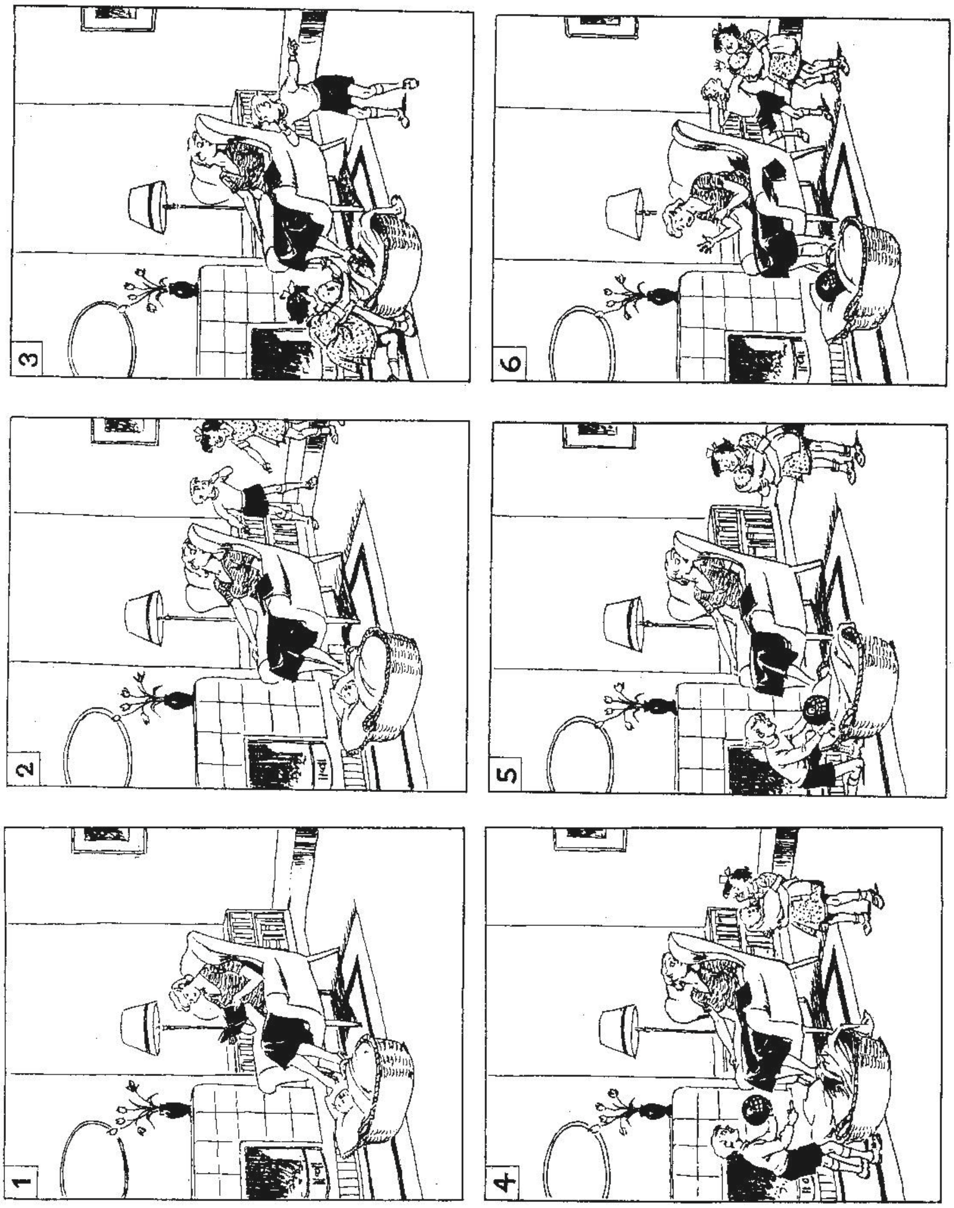
Age: **(Participants must be 18 years of age or older. If you are under the age of 18, please let the researcher know at this point)**

Appendix 3. Task 3.



Appendix 4. Task 4.

B



Appendix 5. CAF Descriptive Statistics Main Study

Table 64 Descriptive statistics of CAF measures according to planning time

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	4.79	1.44	76.21	4.82	.00	18.81	60.46	75.05	91.33	74.47	6.42	1.14	82.30	1.53	66.47	62.12	81.70	3.50	62.08	9.45
	(.14)	(.07)	(1.27)	(.44)	(.20)	(1.26)	(6.17)	(5.60)	(1.75)	(3.30)	(1.35)	(.09)	(1.89)	(.15)	(3.48)	(4.61)	(5.09)	(.21)	(3.24)	(.54)
1	4.6	1.50	77.41	5.59	.17	17.75	61.50	79.91	87.51	75.10	4.51	1.14	80.73	1.22	70.22	67.74	64.77	3.95	61.45	7.82
	(.15)	(.08)	(1.36)	(.47)	(.17)	(1.35)	(6.64)	(6.03)	(1.89)	(3.55)	(1.45)	(.10)	(2.03)	(.16)	(3.74)	(5.00)	(5.48)	(.23)	(3.49)	(.58)
5	5.03	1.58	75.55	4.75	.16	18.87	66.29	70.45	89.55	78.12	4.69	1.07	90.83	.93	82.10	76.33	72.88	4.23	74.12	10.58
	(.17)	(.09)	(1.55)	(.53)	(.21)	(1.53)	(7.53)	(6.83)	(2.14)	(4.03)	(1.65)	(.11)	(2.3)	(.18)	(4.24)	(5.63)	(6.21)	(.26)	(4.00)	(.66)
10	5.24	1.59	76.89	5.32	.24	17.43	60.33	69.36	92.20	77.48	3.42	1.10	86.57	1.24	74.53	64.72	88.10	4.07	68.36	10.14
	(.14)	(.07)	(1.33)	(.46)	(.13)	(1.32)	(6.49)	(5.90)	(1.84)	(3.48)	(1.42)	(.10)	(2.00)	(.15)	(3.66)	(4.85)	(5.36)	(.22)	(3.41)	(.57)

Standard deviation appears in parenthesis

Table 65 Descriptive statistics of CAF measures according to planning time: Task 1

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	4.89	1.37	76.42	3.23	.35	20.36	53.25	74.40	95.83	78.08	5.55	1.15	85.47	1.67	69.83	69.23	63.17	3.17	60.78	8.17
	(.36)	(.18)	(3.32)	(1.14)	(.21)	(3.29)	(16.19)	(14.69)	(4.60)	(8.66)	(3.54)	(.24)	(4.95)	(.38)	(9.11)	(12.09)	(13.36)	(.56)	(8.50)	(1.42)
1	4.58	1.32	79.45	8.18	.07	16.01	51.16	84.17	81.07	85.36	10.92	1.07	79.22	1.09	74.69	74.30	51.11	3.54	64.28	6.67
	(.31)	(.16)	(28.3)	(.97)	(.15)	(2.8)	(13.76)	(12.49)	(3.91)	(7.36)	(3.01)	(.20)	(4.21)	(.32)	(7.74)	(10.28)	(11.35)	(.47)	(7.22)	(1.20)
5	5.98	1.32	79.40	4.07	.4	16.38	70.80	68.93	92.86	90.25	9.60	.77	92.50	.72	79.33	74.65	78.75	5.12	76.70	12.00
	(.38)	(.2)	(3.53)	(1.21)	(.24)	(3.49)	(17.17)	(15.58)	(4.88)	(9.19)	(3.76)	(.25)	(5.25)	(.40)	(9.66)	(12.83)	(14.17)	(.59)	(9.02)	(1.50)
10	5.19	1.38	74.76	2.59	.70	22.41	66.74	89.30	91.36	79.97	2.80	.97	93.01	1.54	86.80	72.78	64.63	3.68	66.94	7.30
	(.23)	(.11)	(2.01)	(.72)	(.15)	(2.08)	(10.24)	(9.29)	(2.91)	(5.48)	(2.24)	(.15)	(3.13)	(.24)	(5.76)	(7.65)	(8.45)	(.35)	(5.38)	(8.96)

Standard deviation appears in parenthesis

Table 66 Descriptive statistics of CAF measures according to planning time: Task 2

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	5.35	1.60	76.54	3.05	.13	20.07	63.10	90.25	95.35	85.00	9.18	1.00	91.35	1.06	77.05	74.70	73.25	4.46	80.88	9.75
	(.38)	(.20)	(3.53)	(1.21)	(.11)	(3.49)	(17.17)	(15.58)	(4.88)	(9.19)	(3.76)	(.25)	(5.25)	(.4)	(9.66)	(12.83)	(14.17)	(.59)	(9.02)	(1.50)
1	4.65	1.70	77.82	5.37	.00	19.69	47.93	85.79	94.82	87.41	1.11	1.13	80.55	1.82	75.13	72.22	60.17	3.58	54.27	5.67
	(.27)	(.14)	(2.50)	(.86)	(.18)	(2.48)	(12.18)	(11.10)	(3.46)	(6.52)	(2.66)	(.18)	(3.73)	(.29)	(6.86)	(9.10)	(10.05)	(.42)	(6.40)	(1.07)
5	5.23	1.99	83.21	.47	.40	15.92	80.00	59.60	82.25	81.60	4.00	1.25	96.95	.50	83.30	78.50	67.50	4.81	85.30	12.00
	(.44)	(.23)	(4.07)	(1.40)	(.24)	(4.04)	(19.82)	(18.00)	(5.63)	(10.6)	(4.33)	(.29)	(6.07)	(.47)	(11.16)	(14.81)	(16.36)	(.68)	(10.41)	(1.74)
10	5.79	1.48	79.38	5.28	.69	14.64	71.28	70.24	92.54	94.22	6.29	.75	79.74	1.26	73.42	53.39	106.78	4.00	66.72	9.56
	(.27)	(.12)	(2.48)	(.85)	(.15)	(2.46)	(12.06)	(11.00)	(3.43)	(6.46)	(2.64)	(.18)	(3.69)	(.28)	(6.79)	(9.01)	(9.96)	(4.14)	(6.34)	(1.06)

Standard deviation appears in parenthesis

Table 67 Descriptive statistics of CAF measures according to planning time: Task 3

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	4.75	1.44	77.94	5.11	.13	16.82	76.41	75.89	88.62	76.48	8.13	1.04	76.50	1.69	60.74	48.23	96.08	3.67	58.28	10.08
	(.2)	(.1)	(1.84)	(.63)	(.11)	(1.82)	(8.95)	(8.12)	(2.54)	(4.79)	(1.96)	(.13)	(2.74)	(.21)	(5.04)	(6.68)	(7.38)	(.31)	(4.70)	(.78)
1	4.57	1.56	75.73	4.95	.00	17.36	73.63	77.24	85.56	57.80	1.18	1.13	82.33	.78	67.29	59.04	79.56	4.63	64.30	10.11
	(.32)	(.16)	(2.93)	(1.01)	(.18)	(2.91)	(14.28)	(12.96)	(4.05)	(7.64)	(3.12)	(.21)	(4.37)	(.34)	(8.03)	(1.07)	(11.8)	(.49)	(7.50)	(1.25)
5	4.83	1.46	65.43	7.66	.59	26.32	64.15	90.07	92.83	65.19	2.95	1.10	89.28	1.05	93.14	77.33	71.70	3.31	67.48	9.90
	(.34)	(.18)	(3.15)	(1.08)	(.19)	(3.13)	(15.36)	(13.94)	(4.36)	(8.23)	(3.36)	(.23)	(4.7)	(.36)	(8.64)	(11.5)	(12.70)	(.53)	(8.07)	(1.35)
10	4.72	1.55	70.89	6.51	.08	22.52	56.43	60.05	94.07	80.65	4.10	1.13	92.50	1.27	89.17	79.87	84.43	3.93	70.98	11.83
	(.26)	(.11)	(2.35)	(.81)	(.14)	(2.33)	(11.45)	(10.39)	(3.25)	(6.13)	(2.50)	(.17)	(3.5)	(.27)	(6.44)	(8.55)	(9.44)	(.39)	(6.01)	(1.00)

Standard deviation appears in parenthesis

Table 68 Descriptive statistics of CAF measures according to planning time: Task 4

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	4.39	1.38	74.14	6.77	1.00	18.92	47.57	64.50	88.35	63.03	3.46	1.32	79.97	1.67	62.91	62.56	85.23	2.93	54.20	9.53
	(.22)	(.11)	(2.01)	(.69)	(.12)	(2.00)	(9.80)	(8.90)	(2.78)	(5.25)	(2.14)	(.14)	(3.00)	(.23)	(5.52)	(7.32)	(8.09)	(.34)	(5.15)	(.86)
1	4.62	1.38	76.27	2.97	.10	18.08	79.15	68.72	89.15	67.22	4.10	1.34	80.85	1.19	60.54	64.20	70.00	4.11	63.68	9.33
	(.24)	(.12)	(2.20)	(.76)	(.13)	(2.18)	(10.71)	(9.72)	(3.04)	(5.73)	(2.34)	(.16)	(3.28)	(.25)	(6.03)	(8.00)	(8.83)	(.37)	(5.62)	(.94)
5	4.37	1.55	74.62	6.12	.04	17.55	55.58	65.66	90.01	76.22	3.03	1.13	86.66	1.28	75.78	75.35	73.33	4.04	69.39	9.17
	(.22)	(.11)	(2.05)	(.71)	(.12)	(2.03)	(9.99)	(9.07)	(2.84)	(5.35)	(2.18)	(.15)	(3.06)	(.24)	(5.62)	(7.46)	(8.24)	(.34)	(5.25)	(.88)
10	5.08	1.72	79.80	6.38	.32	13.50	47.68	61.40	91.17	56.95	.50	1.54	85.17	1.00	57.72	60.57	87.50	4.50	69.21	11.50
	(.33)	(.17)	(3.03)	(1.04)	(.18)	(3.01)	(14.8)	(13.41)	(4.20)	(7.91)	(3.23)	(.22)	(4.52)	(.35)	(8.32)	(11.04)	(12.2)	(.51)	(7.76)	(1.29)

Standard deviation appears in parenthesis

Table 69 Descriptive statistics of CAF measures according to planning time: proficiency group A

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	4.88	1.38	74.01	5.24	.00	20.63	58.57	69.96	93.87	77.13	5.03	1.08	83.03	1.78	71.00	65.93	67.38	3.25	62.46	8.11
	(.16)	(.08)	(1.51)	(.52)	(.09)	(1.50)	(7.35)	(6.67)	(2.09)	(3.93)	(1.61)	(.11)	(2.25)	(.17)	(4.14)	(5.49)	(6.07)	(.25)	(3.86)	(.64)
1	4.85	1.51	73.95	5.05	.18	21.91	68.59	80.68	87.65	74.20	3.25	1.12	82.35	1.45	73.32	64.27	68.25	3.61	62.48	7.85
	(.16)	(.08)	(1.46)	(.50)	(.09)	(1.45)	(7.13)	(6.47)	(2.02)	(3.81)	(1.56)	(.11)	(2.18)	(.17)	(4.01)	(5.32)	(5.88)	(.25)	(3.74)	(.62)
5	5.16	1.45	75.10	4.77	.06	20.07	66.38	71.65	91.98	78.00	8.05	1.02	92.24	1.11	85.63	73.22	72.85	4.15	73.48	10.95
	(.24)	(.12)	(2.20)	(.76)	(.13)	(2.18)	(10.73)	(9.74)	(3.05)	(5.74)	(2.35)	(.16)	(3.28)	(.25)	(6.04)	(8.02)	(8.86)	(.37)	(5.63)	(.94)
10	5.03	1.46	73.30	4.09	.25	22.36	46.60	70.54	87.77	76.75	5.06	1.33	88.96	1.73	82.42	74.73	74.87	3.35	64.49	9.03
	(.17)	(.08)	(1.58)	(.54)	(.10)	(1.56)	(7.68)	(6.97)	(2.18)	(4.11)	(1.68)	(.11)	(2.35)	(.18)	(4.32)	(5.74)	(6.34)	(.26)	(4.03)	(.67)

Standard deviation appears in parenthesis

Table 70 Descriptive statistics of CAF measures according to planning time: proficiency group B

	G. Index	CAS	K1	K2	AWL	NONE	ART	PREP	TENSE	VERBS	SELF	M.N.E	P.T.R	M.N.H	F.HES	F.PA	T.S.T	M.L.U	SP.R	IDEAS
30	4.86	1.54	75.41	4.39	.17	20.03	61.68	93.23	92.51	77.63	7.16	1.10	86.44	1.05	71.13	64.99	78.68	3.87	66.93	9.94
	(.26)	(.13)	(2.39)	(.82)	(.14)	(2.37)	(11.62)	(10.55)	(3.30)	(6.22)	(2.54)	(.17)	(3.56)	(.27)	(6.54)	(8.68)	(9.59)	(.40)	(6.10)	(1.02)
1	4.61	1.52	76.80	5.31	.05	17.06	82.68	73.91	89.56	77.25	6.38	1.24	82.00	.97	80.84	61.33	73.13	4.32	64.58	9.00
	(.26)	(.13)	(2.39)	(.82)	(.14)	(2.37)	(11.62)	(10.55)	(3.30)	(6.22)	(2.54)	(.17)	(3.56)	(.27)	(6.54)	(8.68)	(9.59)	(.40)	(6.10)	(.10)
5	5.17	1.8	75.37	4.53	.54	19.56	71.66	76.45	88.93	80.80	.63	1.07	91.73	.74	85.81	77.90	69.88	4.49	77.19	9.88
	(.27)	(.14)	(2.49)	(.86)	(.15)	(2.47)	(12.14)	(11.02)	(3.45)	(6.50)	(2.65)	(.18)	(3.72)	(.29)	(6.83)	(9.07)	(10.02)	(.42)	(6.38)	(1.06)
10	5.36	1.65	75.89	6.25	.58	17.30	76.18	60.64	95.12	82.43	.40	.89	88.52	.99	79.20	65.60	107.00	4.25	71.23	11.33
	(.26)	(.13)	(2.35)	(.81)	(.14)	(2.33)	(11.45)	(10.39)	(3.25)	(6.13)	(2.50)	(.17)	(3.50)	(.27)	(6.44)	(8.55)	(9.44)	(.39)	(6.01)	(1.00)

Standard deviation appears in parenthesis