THE EFFECTS OF TASK COMPLEXITY ON L2 ORAL PRODUCTION AS MEDIATED BY DIFFERENCES IN WORKING MEMORY CAPACITY

Mary Recio Crespo

A Master's thesis for a Degree in Applied Linguistics and Language Acquisition in Multilingual Contexts (LAALCM)

Barcelona

2011

Supervised by: Dr. Roger Gilabert

Department of English and German Philology
Faculty of Philology
University of Barcelona

ABSTRACT

This paper analyzes the effects of increasing task complexity along reasoning demands on L2 oral performance, factoring in individual differences in working memory capacity (WMC) and affective factors. Existing research in manipulation of oral task cognitive demands has until only recently overlooked the role of individual differences. In this experiment, thirty subjects performed both a simple and complex version of a decision-making task (counterbalanced), three different working memory (WM) tests, and an affective variables questionnaire. A novelty of this study is the use of subjective time estimation, together with time on task and affective difficulty, as an autonomous measure of the cognitive load of the task. The three measures display a significant difference in complexity between the two versions of the task. The analysis showed that increased task complexity caused a decrease in fluency and an increase in accuracy with no overall effects of WMC or affective factors. Only attention control displayed a weak negative correlation with lexical complexity and a positive one with structural complexity. Results are discussed in the light of cognition theories, WM models of processing and attention and previous findings in task based research.

Keywords: task complexity, reasoning demands, L2 oral performance, working memory capacity, affective factors, subjective time estimation.

ACKNOWLEDGEMENTS

First of all I am sincerely thankful to Roger Gilabert, my supervisor, for his continuous support, patience, availability and enthusiasm in every step of this process. This paper would definitely not have been possible without his continuous help and encouragement. I would also like to thank Mayya Levkina for her valuable contribution with the results of the span tasks. Moreover, I must give special thanks to the 48 participants who volunteered for the study. Furthermore, I would like to extend my thanks to my family and all my master's colleagues for their support and group contributions; together, objectives seemed more reachable. I want to give special thanks to Eve Conway for correcting my paper and to Lidia Montero for interrating the data. Finally thanks to Tamara Aguilera and Elena Guardiola for their continuous encouragement.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF FIGURES AND TABLES

• Figures

• Tables

LIST OF ABBREVIATIONS AND SYMBOLS

CAF: Measures of fluency, accuracy and lexical and structural complexity

L2: Foreign language

LST: Letter span test

RST: Reading span test

STE: Subjective time estimation

STM: Short-term memory

TCF: Triadic Componential Framework

TMT: Trail Making Test

WM: Working memory

WMC: Working memory capacity

Counter-intuitive measure

* Significant result

1. INTRODUCTION

In a global world where communicative competence in foreign languages has become a necessity, one of the main concerns of teachers and researchers is to find ways to help learners to improve their speech production. In this vein, a number of research studies have investigated possible variables affecting L2 oral performance. Evidence proves that manipulating different features of task design influences learners' output in terms of lexical and structural complexity, fluency and/or accuracy. The main areas of research in task design have been: planning time (see Ellis, 2005; Foster & Skehan, 2008 for recent reviews)[1]; task familiarity studies (Bygate, 2008)[1]; task complexity studies (Robinson & Gilabert, 2007)[1]; and interaction studies (Pica T., Kang H. & Sauro S., 2006; Mackey A. & Goo J., 2007)[1]. In other areas of task design there has been little or no research. This is the case with individual differences and task performance, specifically the relationship between task complexity (i.e. cognitive factors) and task difficulty (i.e. individual differences)[2] with oral production. Still, both groups of factors, together with task condition[3], simultaneously influence oral performance. This is the gap which the present study aims to fill.

Thus, the objective of this study is twofold. On one hand, we are going to partially replicate Gilabert et al.'s study (in press) in order to measure the effects that manipulating task reasoning demands (i.e. task complexity) can have on learners' oral fluency, accuracy and lexical and structural complexity. On the other hand, we are going to analyze how these results are mediated by learners' affective and ability differences (i.e. task difficulty). Regarding ability factors we are going to focus on WM as a central construct in psychological studies about individual abilities as it is one of the most intensively studied areas in contemporary cognitive psychology (Miyake & Shah, 1999: op.cit. Mota, 2003). Moreover, WM is at the heart of complex behaviour and evidence shows that it is a source of individual differences in both learning and the performance of complex cognitive tasks (Baddeley, 1999; Daneman & Carpenter, 1980: op.cit. Mota, 2003). Therefore this study will try to bring together these intrinsically related task cognitive and individual factors affecting L2 speech production.

[1] op.cit. Gilabert et al. (in press)
[2] Coming from Robinson's (2007) taxonomy of tasks demands
[3] The present study only takes into account task design under monologic conditions.

2. LIT REVIEW

Bearing in mind the objectives presented in the introduction, this review, primarily aims to contextualize the factors we are going to deal with in the experiment within the theoretical framework of the Cognition Hypothesis. Secondly, the characteristics and implications of each factor will be explained more in depth and, finally, the major findings in research in the related fields will be summarized.

The central theoretical framework for analyzing the influence of task demands on L2 production is the Cognition Hypothesis of adult second language acquisition and its associated Triadic Componential Framework (TCF) (Robinson, 2011, 2007; Robinson & Gilabert, 2007). Robinson (2011) refers to Candlin (1987: op.cit. Robinson, 2011), who raises the issue that tasks can be used as constructs for theoretical hypothesis of SLA since effects of different tasks on production can be measured and compared. In order to guide research analyzing these effects, Robinson provides a taxonomy of task characteristics (the TCF) to define tasks and explain their means of operationalization. The TCF differentiates three main categories of task demands: (1) task condition, which refers to interactional demands; (2) task complexity, which is concerned with cognitive factors which are intrinsic to the task (i.e. reasoning demands); and (3) task difficulty, which deals with individual differences in learners' factors which make the same task more or less difficult for different subjects (i.e. affective factors and working memory capacity). Regarding the first group, we are going to deal with only monologic conditions; therefore the categories we want to describe and define for the purpose of the study are the two latter ones.

2.1. Task complexity

To begin with, task complexity is defined as: "the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner." (Robinson, 2001:28).

From this definition and from other studies on task complexity, two important assumptions can be inferred that would lead to the basic grounds of the study to be carried out. The first one is that tasks differ in their degree of complexity, which in turn

affects L2 production. The second is that the internal features of a task can be manipulated so that the effects of different factors on L2 production can be measured and later predicted.

According to Robinson, in the TCF, features affecting the cognitive complexity of the tasks can essentially be manipulated along two types of variables that affect resource allocation differently during L2 task performance:

- *Resource-dispersing variables*: related to performative and procedural demands (e.g. less planning time or familiarity of task or topic). Increasing these variables makes great demands on learners' attentional and memory resources and, consequently, disperses them.

- *Resource-directing variables*: related to cognitive and conceptual demands (e.g. number of elements, reasoning demands). It draws learners' attention to vocabulary and syntax encoding.

Resource-dispersing variables should encourage faster and more automatic L2 access and use (i.e. therefore approximating real-life demands), but they do not direct resources to features of language code, whereas resource-directing variables direct learners' attention to forms needed to meet task demands, and therefore, they will use a wider lexical variety, more complex grammatical structures and more accurate speech, usually at the expense of fluency. This point, however, has been shown to depend on the degree of proficiency (as suggested by Gilabert et al., in press).

The Cognition Hypothesis rationale differs in some aspects from another model of task demands. Skehan's limited attention capacity model (2003: op.cit. Ellis, 2010) and his Trade-off Hypothesis (2009) suggest that it is not task complexity but particular combinations of task characteristics and conditions that predict correlations between different dimensions of performance. Due to attentional and memory limitations competition for attention exists and it leads to trade-off effects, typically between complexity and accuracy. When task conditions are simplified, mainly by giving planning-time to students, this competition is diminished.

The fundamental difference between Skehan and Robinson appears to be how they view attentional resources. “While Skehan sees attention as a single mechanism with all cognitive demands competing for the same finite resource, Robinson sees it as comprising multiple resources that can operate separately and/or simultaneously through a central executive (Baddeley, 1986, 1996)” (Ellis, 2010:4). Thus, the distinction between resource-directing and resource-dispersing factors is not heeded by Skehan’s claims. Robinson would predict trade-off effects for resource-dispersing variables, whereas, along resource-directing factors, both linguistic accuracy and complexity can increase simultaneously without conflict.

In any case, both authors share the idea that manipulation of task complexity should lead to different results in oral performance. Notwithstanding, to date there is a dearth of evidence to support this prediction. The few existing studies have used measures of fluency, lexical and structural complexity and accuracy (CAF) for analysis. Findings point in the direction of task complexity manipulations affecting CAF, nevertheless evidence is not conclusive.

Referring only to studies dealing with reasoning demands, in terms of fluency, Niwa found out that, under complex reasoning demands, fluency was significantly reduced (Niwa, 2000: op.cit. Gilabert et al., in press). With regard to accuracy, the results show a positive impact of reasoning demands on the number of instances of self-repair used as a measure of accuracy (Gilabert, 2007). As far as complexity (lexical and structural) is concerned, there is no evidence of a significant effect of task complexity manipulation. Gilabert et al. (in press) obtained only a strong trend for lexical complexity in the fire chief task.

Typically, experiments that aim to measure the effects of complexity on production establish two levels of complexity (e.g. more or less reasoning demands) in task performance, and then production is measured to assess the effects of the manipulation. In this study we are going to deal with two of the weaknesses of this way of operationalization. First, as Gilabert et al. (in press) pointed out, research “has not factored in individual differences (e.g. differences in WM capacity) which have been shown to affect production (Kormos & Trebits, in press) and which may provide a much richer picture of L2 performance as mediated by cognitive complexity”. The

second limitation often brought up is that there is no autonomous way to measure cognitive complexity itself. To date most studies operationalize task complexity as simple/complex, make predictions about effects on performance, and use results to confirm or dismiss the differences in task complexity, hence falling into a circular argument.

It is precisely in this vein, that the present study is to be contextualized as it is replicating Gilabert et al.'s (in press) study while factoring in individual differences; and using subjective time estimation (STE) as an independent measure of complexity. In the same light, Robinson (2001) maintains that interactions between the three categories of task to be perceived as demands (i.e. complexity, difficulty and condition) may be expected, and so task difficulty is going to be reviewed below.

2.2. Task difficulty

According to Robinson (2001), cognitive factors contribute to intrinsic task complexity. However, the demands of the task are also dependent on learner individual differences which will make the task more or less difficult (as opposed to complex). Learners' factors in TCF are divided into two subcategories: Affective and ability variables.

- Affective factors

On the one hand, affective factors (e.g. confidence, motivation, anxiety…) are temporary as they may change and affect task production to different degrees. Not much research has been carried out on this field. Robinson (2001) and Gilabert et al. (2009) administered an affective variable questionnaire and found similar results. A significant main effect was shown for perceived difficulty, stress and confidence and no significant main effects for interest and motivation was found. Robinson also concludes that findings for affective factors on production are weaker than those for cognitive complexity. Regarding the perception of the simple and the complex versions of the task, Robinson observes that sequencing tasks from simple to complex or vice versa does not significantly affect perceived difficulty ratings, so typically more complex tasks are perceived as more difficult.

- Working memory capacity

On the other hand, ability variables (e.g. aptitude, reasoning, WM…) are more stable, so their effects on task performance should be more predictable. From the array of factors categorized under this label, this study is focusing on WM because, as Conway et al. (2005) state, it is a central psychological construct that has been widely used scientifically; it is involved in a wide range of complex cognitive behavior, such as comprehension, reasoning and problem solving; and, furthermore, WMC "is an important individual-differences variable and accounts for a significant portion of variance in general intellectual ability" (pg. 769). In fact, Miyake & Friedman (1998: op.cit. Gilabert & Muñoz, 2010:22) suggested that WMC should be equated with foreign language aptitude as it "can capture the essence of the three important components of the language aptitude suggested by Skehan (1989) – a language analytic capacity, memory ability, and phonetic coding ability". Moreover, many other researchers in the field have appraised the importance of WM as a potential measure of aptitude (e.g. Robinson, 2002[4]; Hummel, 2009).

From the various attempts in the literature at defining WM (for a review see Miyake & Shah, 1999)[4], WM could be defined as a limited-resource multi-component memory system in charge of temporary active maintenance and accessibility of task-relevant information during the ongoing processing of complex cognitive tasks (e.g. Baddeley, 1981, 1990, 1992, 1999; Baddeley & Hitch, 1974; Carpenter & Just, 1989; Carpenter, Miyake, & Just, 1994; Daneman, 1991; Engle, Cantor, & Carullo, 1992; Miyake & Shah, 1999)[5]; (Conway et al., 2005; Mota, 2003).

The most widely accepted conceptualization of WM today is the model developed by Baddeley and Hitch (1974)[4] and Baddeley (1984, 1986, 2000)[4]. Its architecture model consists of four limited-resource components: (1) a central executive system, responsible for switching attentional focus and control of three other subcomponents; (2) a phonological loop, in charge of temporarily storing verbal information (for around 2 seconds) and maintaining it through rehearsal for ongoing processing; (3) a visuo-spatial sketchpad, which stores and manipulates visuo-spatial information; and (4) an

[4] op.cit. Gilabert & Muñoz, 2010
[5] op.cit. Mota, 2003

episodic buffer, controlling the integration of the material in the other subcomponents with the information in the long-term memory and the creation of episodes (Gilabert & Muñoz, 2010; Kormos & Sáfár, 2008).

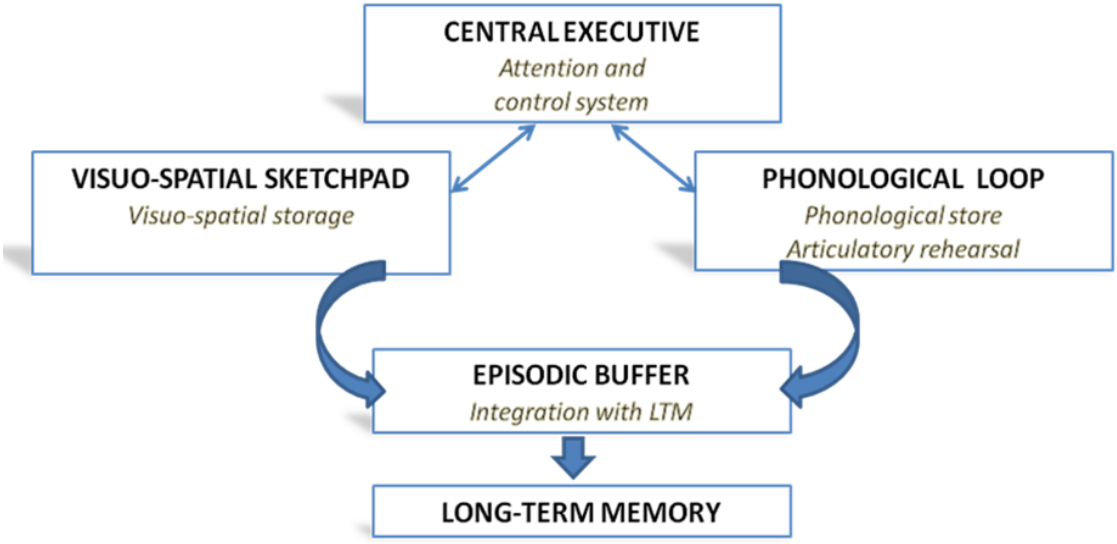


Figure 1: Architecture of WM according to Baddeley and Hitch (1974) & Baddeley (1984, 1986, 2000). Graph created by Gilabert & Muñoz (2010)

As WMC is limited, mental processes involved in the performance of complex tasks compete for attention. Therefore, it is reasonable to assume that differences in WMC should affect performance on more complex tasks. As Mota, (2003) posits, the interpretation of this model is that individuals with a higher WMC tend to demonstrate better performance on complex tasks than individuals with a lower working memory capacity and she reports that research findings in the area support the hypothesis. Also Trebits and Kormos (2008) confirmed an impact of WMC on a complex version of a narrative task on fluency (measured as speech rate) and lexical complexity (scored with D-value).

However, not all studies uphold her claim, especially if we focus on the effects of WMC on oral production. Even Fortkamp's 1999 study confirmed the initial hypothesis only for the reading task and not for the oral one and, in her experiment of 2000, she found positive evidence only with an L2 span task. It can be argued that span tasks in the L2 measure, at least partially, L2 proficiency; consequently, it comes as no surprise that proficiency and the L2 span task were correlated. Another two studies with the same drawback are Mota's (2003) and Guara-Tavares' (2009) studies. Both experiments found positive correlations of WMC –measured with a span test in the participants' L2– and fluency, structural complexity and accuracy. Once more, the L2 span task is

likely to be measuring proficiency or at least it is a confounding variable with WMC. An independent measure of proficiency is needed to disentangle the effect of each measure.

Results in other experiments have been, at least, inconclusive. Fortkamp (1998) found no significant correlations between measures of L2 fluency and a WM span task. Mizera (2006) looked for a relationship between working memory capacity and L2 fluency, lexical access speech and monitoring abilities. Against her predictions, only the two latter factors correlated. Finally, Gilabert and Muñoz (2010) found no correlation between overall proficiency and WMC and a weak correlation between fluency and lexical complexity, and WMC. It is interesting that, when they split the participants into two groups, the correlation with lexical complexity increased to moderate for the high-proficiency group suggesting that it is possible that WMC is more influential for learners in the later stages of acquisition.

In light of this review, our goal is to investigate how task complexity affects performance and how WMC and affective variables mediate the effects of task complexity in performance. Accordingly, our research questions are stated as follows.

3. RESEARCH QUESTIONS

1. How does the manipulation of task complexity along reasoning demands affect learners' fluency, accuracy and lexical and structural complexity under monologic conditions?

2. Does WM mediate the effects of task complexity on performance?

3. Do affective variables have an effect on performance when manipulating task reasoning demands?

The corresponding hypothesis in the light of previous research and frameworks:

1. Following Robinson's Cognition Hypothesis, we predict that the more cognitively complex the task, the more complex and accurate the speech will be, but also the less fluent. However, we also predict that, based on previous research outcomes, proficiency will play a significant role and it will be a factor to take into account while analyzing the results.

2. Although theory points to the direction that the higher the WMC, the better the performance on cognitively complex tasks; it is in language comprehension that this theory has been proven (Fortkamp, 1998). The varied results in research comparing L2 oral production and WMC lead us to favor the hypothesis that WMC will not affect the results significantly.

3. According to the previous findings referred to before, the hypothesis is that perceived difficulty, stress and confidence will have an effect on performance but not as significant as cognitive demands. Interest and motivation will not affect CAF.

4. METHOD

4.1. Participants

The participants who volunteered for this study were 30 native speakers of Spanish and/or Catalan with English as a foreign language (L2) (8 males and 22 females). The only previous requirement of the study regarding proficiency was being able to perform the fire chief task. Their L2 proficiency ranged from lower intermediate to advanced (although most of the subjects had a high level of proficiency) and was controlled for with two tests that will be described in the following section. Regarding age, following our requirements, subjects were between 18 (before WMC is not fully developed) and 40 (start point of a decline in WMC), with a large majority in their 30s.

4.2. Materials

As for the instruments used for data collection, our participants completed a battery of tests and tasks as described below:

- Proficiency tests:

Participants had to complete three tests to control for their proficiency. The first two are the X_Lex and Y_Lex tests (as described in Meara and Milton, 2003: op.cit. Gilabert et al., in press), measuring vocabulary size. Results in these tests have been proven to correlate with general proficiency (Meara, 2005). The other one is the C-Test (Gilabert, 2005; Wesche and Paribakht, 1996) that measures other language knowledge aspects (e.g. morphological, syntactic and textual competence) besides vocabulary.

- Decision-making task:

In order to elicit speech under the two levels of reasoning demands, we used a simple and a complex version of the “Fire chief” task as described in Gilabert (2007) and Gilabert et al., (2009) (appendix 1). In this task, adapted from cognitive psychology, a building has caught fire and several people need to be saved. The subjects have to act as if they were the fire chief and they have to decide which strategy to follow, the order of the rescue and then they have to

justify their actions. To make the complex version of the task, reasoning demands were raised by making the relationships between the elements more complex, increasing the number of factors to be taken into account, and diminishing the resources available. In the simple version, the fire is located on one side of the building; the people that need to be saved have no particular needs; there is no smoke getting into the building and the subjects have plenty of resources to fight the fire (i.e. two fire engines and a helicopter). Conversely, in the complex version, the fire has various focus so that access to the building and the evacuation are more difficult; the victims have specific roles (i.e. an old man, a pregnant woman with two children, a severely injured person and a hero); the wind is blowing into the building and the smoke is spreading inside; and the only resource they have available is a fire engine. All of these factors force the subjects to prioritize and justify their decisions more, thus increasing the reasoning demands of the task.

- Affective variables:

After each version of the fire chief task, subjects answered the Affective Variable Questionnaire (Robinson, 2001), in which they rated the perceived difficulty, stress, confidence, interest, and motivation while performing the tasks. Once the second questionnaire was completed, participants had the option to answer three open questions about the difficulties they encountered during the fire chief task and their feelings (appendix 2).

- Trail-making Test (TMT):

The TMT is the first of the three tests used to measure WMC. As described in Bialystok (2010), the TMT "is a neuropsychological test that involves motor speed and attentional control" (pg. 94). TMT has probably been the most widely used instrument to assess the executive function of WM. Arbuthnott & Frank (2000) found empirical evidence for the validity of this test. The TMT consists of two timed subtasks called Trail A and Trail B. In Trail A, the participants have to draw a line, as quickly as possible, through a sequence of numbers from 1 to 25 scattered over a page. The instruction is to connect the

numbers in order beginning with 1 without lifting the pen from the paper. In Trail B, the page contains numbers from 1 to 13 and letters from A to L. Subjects again have to draw a line beginning with 1 but they have to alternate between numbers and letters (1 A 2 B 3 C…) (appendix 3).

- Letter span test (LST):

The LST was designed to measure STM factor, measuring only the storage capacity of WM (Conway et al, 2005; Kane et al., 2004). Participants had to recall sequences of letters presented in 14 increasingly larger series from 3 to 9 items. The words appeared on a blank screen, one after the other, and at the end of the set participants were to recall the letters in the correct order by clicking on a screen displaying 12 different letters. After each set, accuracy feedback was given.

- Reading span test (RST):

Finally, the reading span is a dual task incorporating STM span demands (i.e. storage) with a secondary task that engages the central executive (i.e. processing) (Conway et al., 2005). The reading span task was firstly developed by Daneman & Carpenter (1980). Since then different versions have been created and tested (e.g. Turner and Engle, 1989: op.cit. Conway et al. 2005; Kane et al., 2004), including some attempts to adapt it to Spanish (Sagarra, 2002: op.cit. Gilabert & Muñoz, 2010). WM span tasks have been proven to be both reliable and valid measures of general WMC (Conway et al., 2005; Unsworth et al., 2005). The procedure followed in this case was very similar to the letter span task. However, before each to-be-remembered letter, participants were presented with a sentence and they had to decide whether it made sense or not. Right after the answer the letter was prompted followed by a new sentence. This time the test consisted of 15 sets of 3 to 7 items (sentence and letter), 75 sentences in total plus 15 trial sentences – used to measure participants' reaction times. The sentences were presented in the dominant mother tongue of the participant (i.e. Spanish or Catalan), ranged from 8 to 12 words and 50% made sense while 50% did not. The presentation

order of the sets was randomized to avoid the use of strategies that come from knowing the size of the memory set.

Both span tests were performed as a single 30-minute-long task on a computer. The version of the span tasks used in this study was designed by the GRAL research group (UB) with e-Prime software. Sagarra's (2002) Spanish versions of the reading span were used as a basis to create this version in Spanish and Catalan. This computer-based version of the test is entirely mouse driven, paced depending on individuals' reaction times and automatically yields scores upon completion. It is basically divided into 3 parts. The first one is the letter span, which also serves as a practice for the reading span. It was followed by 15 trial sentences, from which mean RTs were calculated to adapt individually the last part, that is the reading span task. The program has already been used by the GRAL group and "internal consistency reached a Cronbach's alpha of 0.872" (Gilabert & Muñoz, 2010: pg. 31).

The reason for using three different tasks to measure WMC is that, as suggested by research findings in the area (Conway et al., 2005; Kane et al., 2004) no single task is a perfect measure of the WM construct. Hence different tests were used, tapping into the different subcomponents of the WM construct (as explained above) in order to have a more comprehensive approach towards WMC. Figure 2 shows the application of every test within Baddeley's model of WM.

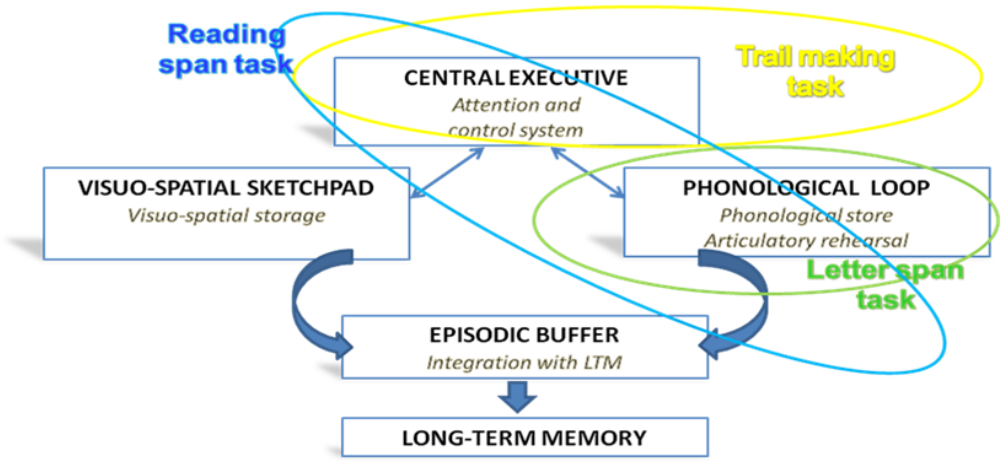


Figure2: subcomponents tapped by each WM test administered in the study

- *Personal information and questionnaires:*

All participants completed a language background and personal details questionnaire (appendix 4) and also signed an informed consent sheet (appendix 5).

4.3. Data collection procedures

First of all, a pilot study was carried out with a baseline of 18 native speakers just to discard the use of L2 in the fire chief task as the cause for higher task complexity. The participants in the pilot were also from 18 to 40 years of age and only had to undergo a part of the whole experiment: the fire chief task (simple and complex version counterbalanced) with the STE question, the affective variables test, the three open questions on the task, and the personal details and language background questionnaire. They performed the experiment in their dominant L1 and the session lasted around 10-15 minutes. The results confirmed the higher cognitive demands of the more complex version. Although the distance between the time judgments did not yield significant differences, it followed a strong trend and it would have probably reached significant levels with a higher population (table 1).

Table 1

Pilot study: task complexity measures and correlations

	Task	N	Min.	Max.	Mean	SD	Wilcoxon Signed Rank Test
Affective Difficulty	*Simple*	15	1	6	3.33	1.718	.002* r = -.460
	Complex	15	2	9	5.27	2.017	
TOT	*Simple*	15	17	89	46.07	23.206	.001* r =-.398
	Complex	15	17	188	76.33	43.567	
Distance TOT/STE	*Simple*	15	2	43	18.00	12.778	.069 r = -.290
	Complex	15	2	68	28.07	19.746	

SD= standard deviation

Valid N= 15 because 3 subjects were discarded from the results for invalid answers in the time judgment

*significance at p<.05

r= effect size low 0.1-0.3; medium 0.3-0.5: large >0.5

For the general study, data collection took place in a single individual one-and-a-half hour session. Due to participants' location and time constraints, the meeting point for the experiment changed depending on their availability. We tried to meet in a quiet place to facilitate concentration and yield clear recording grounds. The instructions to

follow were the same for each subject. The sequence of the tasks was structured as follows:

- TMT (Trails A and B)
- Letter and reading span tasks.
- X-Lex & Y-Lex (computer format)
- C-test (paper version)
- Fire chief task: simple or complex version (the order was counterbalanced to avoid task effects). The narrative was recorded and timed with a voice-recorder
- STE: participants were asked to note down their estimate of the time needed to solve the task (see appendix 2)
- Affective variable questionnaire
- Fire chief task (the version missing)
- Subjective time estimation
- Affective variable questionnaire
- The three open questions about the difficulties of the task
- Personal details and language background questionnaire
- Informed consent and comments or questions about the experiment

4.4. Transcription, measures and coding

Once all the data was collected, the analysis obtained the scores for each factor.

First of all, we transcribed the recordings using Sound Scriber software and the CLAN mode and conventions of the CHILDES database (see guidelines in appendix 6). Once the transcriptions were coded, we proceeded to measure fluency, lexical and structural complexity and accuracy.

- Measures for oral production (dependent variables)

Fluency was measured as speech rate A (unpruned) and rate B (pruned). The speed was calculated in syllables per minute without pauses for both rates and also without repetitions, self-repairs retracements or false starts for rate B. According to Mora and Valls (2006), “speech rate has been found a robust measure and a very reliable predictor of perceived oral fluency in a wide variety of studies”.

In order to measure *lexical complexity* we used Guiraud's Index of Lexical Richness, which calculates the variety of vocabulary dividing the number of types by the square root of the number of tokens in the narrative. Type-Token Ratio score was rejected because it has proven to be very sensitive to production length (Gilabert et al., in press) and the recordings in this study are quite short.

Structural complexity is measured by length of clauses and nodes per AS-unit. The first measure was calculated by the rate of words per clause and the second by the number of clauses per AS-unit. The reason for using two different measures is that subordination is not always a sign of higher complexity. In fact, coordination has been proven to be an indicative source for beginners; subordination for intermediate levels (measured as nodes per AS-units in this study); while high level students make use of phrasal-level complexification (measured here by length of clauses) (Norris and Ortega, 2009).

Finally, *accuracy* was measured as the number of errors per 100 words. This is a standard measurement in psycholinguistics and it is more reliable than other commonly used measures, such as error-free AS-units, which do not take into account the number of errors in the unit (Gilabert et al. in press).

For the transcription and coding of CAF scores, interrater measures were used on 5% of the data with overall interrater reliability reaching 97%.

- *Losses and gains in performance*

In order to run some of the statistical analyses (i.e. linear regressions) new scores had to be created for each dimension of performance reflecting losses or gains, (according to predictions in hypothesis 1). For instance, as fluency was thought to decrease, the fluency *loss* was calculated by subtracting fluency in the complex task from the simple one. By the same criteria, lexical complexity, structural complexity, and accuracy *gains* were determined by subtracting the performance of the simple task from the complex one (table 2).

Table 2

Gains or losses in performance (operationalization of the new measures)

Hypothesis 1	New measure	Calculation
Fluency decrease	*Fluency loss*	= Fluency simple-fluency complex
Accuracy increase	*Accuracy gains*	=Accuracy simple-Accuracy complex
Lexical complexity increase	*Lexical gains*	=Lexical complexity complex-Lexical complexity simple
Structural complexity increase	*Structural gains*	=Structural complexity complex-Structural complexity simple

- Operationalization of task complexity (independent variable)

As explained in the materials section, the complex version of the fire chief task was designed to have higher reasoning demands based on more complex relationships between the elements. Besides the manipulation of task complexity, two other indexes of the cognitive load of the task were used in order to have empirical evidence on the different demands of the tasks. The first measure was time on task (TOT). Given that both settings had the same number of elements; it can be inferred that time on task for each version should be relatively similar, unless one of the tasks demands longer processing. Therefore, the more time spent solving the task, the more complex it is. The second measure was perceived complexity, rated in the affective variable questionnaire. In this respect, Robinson (2001) asserts that "It is also possible that stable relationships may exist between increases in task complexity and learner perceptions of difficulty, assessed via affective variables, with more complex tasks also being judged to be more difficult". Furthermore, the qualitative data withdrawn from the open questions can also help to confirm the different levels of reasoning demands of the tasks.

However, as these indicators are subjective or dependent on other factors, a possible solution to obtain an independent objective measure of complexity could lay on time judgements. STE has been used in previous research as a reliable index of cognitive load in task performance (Macar et al., 1994; Chastain & Ferraro, 1997; Casini & Macar, 1997; Fink et al., 2001). In these studies, it has been found that STE becomes more inaccurate as nontemporal processing demands are increased. This finding supports the attentional model of time perception proposed by Thomas and Weaver (1975: op.cit. Fink et al. 2001). According to this model, information is analyzed by two

processors: One processing temporal information –a timer– and a nontemporal processor. Attention is shared between these two processors, therefore the more attention devoted to nontemporal processing (e.g. with increased reasoning demands), the more imprecise the temporal information would be (Fink et al., 2001; Macar et al., 1994). Overestimation and underestimation can be expected, especially when time judgments are retrospective[6], as they "are not appropriate to activate a cumulative timer" (Casini & Macar, 1997: pg. 817).

Hence, our added independent measure of complexity is the distance between the actual TOT and the STE for the simple and the complex task. It was calculated by subtracting TOT and STE and using it as a positive value for the importance lays on the inaccuracy of the time judgment regardless of whether it is by overestimation or underestimation.

As we can see in the table below (table 3) the three indexes yielded significant differences. Thus, while they are indirect measures of task complexity, it can be asserted that the higher cognitive complexity of the complex version of the fire chief task was confirmed by three robust independent measurements.

Table 3

Task complexity measures and correlations

	Task	N	Min.	Max.	Mean	SD	Wilcoxon Signed Rank Test
Affective Difficulty	*Simple*	30	1	9	4.57	2.542	0.014* r = -.228
	Complex	30	1	9	5.83	2.829	
TOT	*Simple*	30	30	158	75.03	37.931	0.000* r = -.352
	Complex	30	45	332	114.53	63.850	
Distance TOT/STE	*Simple*	30	0	205	47.20	54.450	0.004* r = -.179
	Complex	30	4	334	71.53	77.214	

SD= standard deviation
*significance at p<.05
r = effect size low 0.1-0.3; medium 0.3-0.5: large >0.5

[6] A retrospective time judgment is required unexpectedly after a given interval has already passed by.

- Individual differences (mediating variables)

    - *Affective factors*

The affective variables are measured on a 9-point likert scale. Also the three open questions at the end of the fire chief tasks yielded useful comments on subjects' feelings during the task.

    - *Working memory capacity*

In the *TMT (Trail Making Task)*, the score is the ratio between the B/A sections of the test. The most commonly used measure for this test was the subtraction of the seconds spent on Trail B from A. Nevertheless, Arbuthnott & Frank (2000) proved in their validation study that "B/A ratio score may provide the best indicator of executive control function" (pg. 527). Consequently the B/A ratio was used in this experiment.

The *LST* and the *RST* generated a single document per subject with the results following a partial credit unit scoring procedure. It means that every element (letter) remembered within the same set is proportionally calculated. Therefore, every letter remembered is computed for the final score but a letter in a longer set has higher value than a letter in a shorter one – on the grounds that it is more difficult to remember. This scoring method is backed up by Conway et al. (2005), based on empirical results and established psychometrics' procedures. These authors (refuting the scores used by Unsworth et al., 2005 for their automated Ospan) claimed an absolute scoring procedure –in which the span score is the sum of all perfectly recalled sets – to be inappropriate in dual tasks for individual-differences research. The reason is that the difficulty of a span item may vary in different dimensions and, as absolute scoring procedure discards the information of the other trials, it might threaten test reliability across different tasks.

For the RST, it is important to add that a percentage of 85% in sentence accuracy was required for the results to be computed, as explicitly stated in the instructions. Lower scores do not guarantee that attention was focused on the processing component of the task, allowing for rehearsal or other strategies, undermining the validity of the result as a WMC measure.

5. RESULTS

Once all the scores from the data were obtained, the SPSS statistical package was used to analyze the results. Firstly, descriptive statistics (tables 2, 3, 6 and 9) and analysis of normality were carried out. The Kolmogorov-Smirnov test showed that the following variables were normally distributed: X-Lex and Y-Lex test; fluency rate A & B, lexical complexity and subordination in both the simple and the complex version; structural complexity only in the simple version; difficulty simple; stress complex; confidence simple and complex; LST; and RST. The rest of the variables were not normally distributed. Therefore, as around half of the variables are non-normally distributed, and also considering that the sample size is relatively small (i.e. N=30), non-parametric statistical tests were run as they are more restrictive and increase the strength of the results.

Regarding proficiency (table 4), our population was skewed towards the right showing a general high level of L2 English with some subjects being low intermediate. The minimum value for the vocabulary size was 3450 tokens and the maximum 7850 tokens (mean 6106.67 tokens) and for the C-test, the lowest score was 39% and the highest 99% (mean 81.60%). The tests displayed a significant correlation p= 0.000 at r= .742 with Rho Spearman.

Table 4

Descriptive statistics of proficiency

Test	*N*	*Minimum*	*Maximum*	*Mean*	*SD*
X-Lex + Y-Lex	30	3450	7850	6106.67	1227.819
C-Test	30	39	99	81.60	15.240

N= sample size; *SD*= standard deviation

Before looking at the rest of the results, it is important to point out that some of the measures used are counter intuitive (i.e. accuracy, stress and TMT ratio B/A). That means that the higher the score, the worse the result (e.g. the higher the number of errors, the higher the score, but it means less accuracy). This will be taken into account for the interpretation of the outcomes, but the figures will be reported as calculated by SPSS directly. These measures will be marked ($^{\#}$) to draw attention to the change.

5.1. Research Question 1

In order to answer research question 1, oral performance in the simple and the complex version of the fire chief task was compared for each measure of speech production by means of the Wilcoxon Signed Rank Test. CAF scores followed the trend predicted in the hypothesis (table 5), except for structural complexity measured as subordination ratio, which was quite similar in both tasks but slightly lower in the complex one. In any case, the only significant differences (appendix 7: table 6) were found in fluency rate B and accuracy (p= .041 and .011 respectively), which confirms partially our hypothesis as task complexity did not have an impact on structural complexity.

Table 5

Descriptive statistics CAF measures: simple and complex measures

	Dependent Variable	Simple					Complex				
		N	*Min.*	*Max.*	*Mean*	*SD*	*N*	*Min.*	*Max.*	*Mean*	*SD*
Fluency	Unpruned Speech Rate A	30	53.46	209.09	135.98	38.95	30	56.47	197.59	126.78	31.94
	Pruned Speech Rate B	30	46.54	208.00	123.83	41.34	30	42.79	171.72	112.46	33.75
Lexical Complexity	Guiraud's Index	30	3.58	6.53	5.17	.69	30	3.69	6.78	5.33	.84
Accuracy#	Number of errors X 100 words	30	.00	20.91	6.51	5.60	30	.00	15.27	4.42	3.61
Structural Complexity	Number of words x clause	30	3.47	8.70	5.90	1.21	30	4.41	10.53	6.12	1.46
	Subordination Ratio	30	.33	1.83	.98	.38	30	.19	3.20	.9069	.58

N= Sample size; *Min.*= Minimum; *Max.*= Maximum; *M*= Mean; *SD*= Standard deviation.
#counter-intuitive measure

As the order of the simple and the complex version was counterbalanced to avoid task effects in performance, sequence might have played a role in participants' production. In order to control for this factor, Mann-Whitney U tests (appendix 7: table 7) were run finding significant differences (p<.005) for fluency and lexical complexity in the simple task. It means that, in the complex-simple sequence, subjects were faster and had a higher lexical variety.

5.2. Research Question 2

Regarding the second research question, correlational analyses were performed in order to compare WM scores (table 8) with CAF scores in the simple and the complex versions (appendix 7: table 9). The outcomes for the TMT# and the RST display almost no correlations with performance, except for TMT ratio B/A# in the complex task with lexical complexity. Gilabert & Muñoz (2010) and Mizera (2006: op.cit. Gilabert & Muñoz, 2010) also found similar trends but measured using WM span tasks. However, the RST in this study only showed a weak correlation with improved accuracy# in the complex version. These findings go in line with the original expectations of obtaining not very significant relationships between WM and production. Nonetheless, hypothesis 2 was only partially confirmed as several correlations were found with the LST. In the simple task, subjects with a higher score on the LST performed better in fluency, lexical complexity and accuracy#. The results were similar in the complex version, except for accuracy# where only a strong trend was visible.

Another interesting finding from the analysis was that RST and LST were strongly correlated.

Table 8

Descriptive statistics of WM measures

Mediating factors	*N*	*Minimum*	*Maximum*	*Mean*	*SD*
TMT ratio B/A#	30	1.00	2.92	1.64	.48
Letter Span Test	30	42.24	91.21	69.54	12.98
Reading Span Test	30	25	75	52.30	12.45

N= sample size; *SD*= standard deviation
#counter-intuitive measure

As mentioned in the lit review, proficiency has been proven to be a very influential factor on results; therefore linear regressions were run with the measures of performance loss or gain, factoring out proficiency and weighting the load of WM on oral production results. Notwithstanding, the only significant impact that emerged from the results was that the TMT ratio B/A explained a significant percentage (16.6%) of the variance in lexical complexity gains, in line with the aforementioned correlation (appendix 7: table 10).

5.3. Research Question 3

To answer the third question, we checked first for significant differences between the participants' affective variables in the simple and the complex task (tables 11). As expected, in line with previous research (Robinson, 2001; Gilabert, 2007; Gilabert et al., 2009), Wilcoxon tests (appendix 7: table 12) indicated significant differences in difficulty, stress[#] and confidence (at p=.014; .000; .003 respectively) but yielded non-significant results for interest and motivation.

Table 11

Descriptive statistics affective variables: simple and complex measures

Dependent Variable	Simple					Complex				
	N	*Min.*	*Max.*	*Mean*	*SD*	*N*	*Min.*	*Max.*	*Mean*	*SD*
Difficulty	30	1	9	4.57	2.54	30	1	9	5.83	2.83
Stress#	30	1	9	6.17	2.45	30	1	9	4.33	2.52
Confidence	30	1	9	5.47	2.19	30	1	9	4.17	2.26
Interest	30	3	9	6.70	1.82	30	2	9	6.93	1.78
Motivation	30	1	9	6.03	2.47	30	1	9	5.83	2.32

N= Sample size; *Min.*= Minimum; *Max.*= Maximum; *M*= Mean; *SD*= Standard deviation.
#counter-intuitive measure

With regard to the relationship between affective variables and performance, different patterns were noticed for the simple and the complex task (appendix 7: table 13). In the simple one, fluency, lexical complexity and accuracy[#] correlated positively with confidence and negatively with stress[#]; whereas in the complex version, only fluency maintained the correlation with stress[#], although to a lesser degree (simple: r=.652; complex r=.368), and had a moderate correlation with perceived difficulty.

Affective factors also presented interesting correlations among them, and this time, results were very similar in the complex and the simple version (appendix 7: table 14). Stress[#] had a strong negative correlation with confidence and a positive one with difficulty. Motivation and interest showed a moderate correlation.

The role of sequence was also analyzed for affective factors. The Mann-Whitney U test (appendix 7: table 15) found significant differences in the perceived difficulty, stress

and interest of the simple task (p=.011; .007; .035). Contrary to Robinson's (2001) findings, the subjects in the complex-simple sequence found the simple task significantly less difficult and stressful but more interesting.

In sum, hypothesis 1 was partially confirmed for fluency decreased and accuracy increased but lexical and structural complexity did not increase significantly. The second hypothesis was also confirmed to some extent because there was a significant correlation between lexical complexity and attention control. Finally, the third hypothesis was supported by the results as there was a significant difference in stress, perceived difficulty and confidence under higher cognitive demands.

6. DISCUSSION

6.1. Research Question 1

As shown in the results section, increased demands of task design on the participants triggered significant results only for fluency loss and accuracy gains.

In everyday life when we are faced with a complex task, we slow down to complete it successfully. For instance, if it starts to rain heavily while we are driving we would typically slow down. Therefore, to fulfill the demands of complex tasks we reduce speed to concentrate on other aspects of performance. The decrease in fluency in this study proves that this pattern is also followed in speech production. As stated by Robinson's Cognition hypothesis and in line with Niwa's (2000: op.cit. Gilabert et al. in press) and Robinson's (1995, 2001) outcomes, speakers slow down, probably, due to the extra attention paid to the linguistic and conceptual encoding required by high reasoning demands.

Supporting this argument, evidence was found for a positive impact of task complexity on accuracy as in previous research (Gilabert, 2007 and Kormos & Trebits, in press). This outcome suggests that, as more complex ideas require more precise language, speakers monitor the oral production process more, being more accurate and, probably, self-repairing more as claimed by Gilabert (2007). As in this paper self-repairs were not taken into account, further analysis of the data would be needed to confirm this hypothesis.

Contrary to the Cognition Hypothesis, lexical complexity did not reveal a significant difference in the complex task, but a trend. Only Robinson (2001) found more lexical variety and Gilabert et al. (in press) a strong trend. It could be argued that the demands of the task were too high and no resources were left for vocabulary complexity. Evidence can be found in the trade-off effect yielded in our data between lexical complexity and accuracy gains. This behavior supports Skehan's (2009) model of limited attention in which accuracy is in competition with fluency and complexity.

Structural complexity has not shown effects of task complexity in other studies either (e.g. Robinson 1995, Gilabert, 2005, 2006). Further specific research should be carried

out to reveal the causes of this contradictory finding. Perhaps, it could be argued that either the hypothesis has to be reviewed for this dimension of performance; or the task was not conducive to more complex structures; or the measures are not sensitive enough to capture the structural complexity used. In fact, this is the main reason argued for the unexpected results shown in this study for structural complexity measured by the index of subordination. As the population in this experiment has mainly a high proficiency in English, subordination might not be a sensitive indicator of structural complexity (Norris & Ortega, 2009). Indeed, an issue for future research is to look for more sensitive measures of structural complexity (Gilabert et al., in press).

All in all, even if the complex task was confirmed as being more complex, these differences were not significantly reflected in oral performance as hypothesized. Some explanations are possible relying on Robinson's Cognition Hypothesis and its TCF. First, the open nature of the task, the lack of familiarity, and the absence of pre-task planning-time might have increased task demands also along resource-dispersing variables producing, as a consequence, mixed results. Another possibility is that, although fluency decreased, subjects still needed more resources to tackle the development of a plan of action with the aforementioned constraints. Hence, trade-off effects occurred. It would be interesting to carry out further research regarding the effects of the combination of more or fewer demands on resource-directing and resource-dispersing variables, especially because both dimensions are usually present simultaneously in everyday tasks.

In the same vein, but in relation to Levelt's model of oral production (1989: op.cit. Kormos and Trebits, in press), Kormos and Trebits (in press) point out that it is important to bear in mind that the tasks analyzed are oral. It means that participants have to conceptualize and encode at the same time and this cognitive load can act as a resource-dispersing factor. As a result, subjects might not have sufficient attentional resources for producing syntactically complex language. It is also feasible that, as it is an open-answer task, subjects can rely on their own linguistic resources to create the message. Thus, they employ the vocabulary that is the most easily accessible from their mental lexicon and, accordingly, they might free attention control for accuracy.

Finally, and opposite to Robinson's (2001) results, we found an impact of sequence on performance. The subjects performing complex-simple were more fluent and displayed higher lexical complexity during the simple task. This finding is supported by the results obtained from the affective variables as when the simple task was performed in second place, it was perceived as less difficult than in the opposite sequence and more difficult tasks caused subjects to slow down. As regards lexical complexity, a higher lexical variety was activated for the complex task and obviously it remained active for the second one. Therefore, lexical retrieval was quicker and it could have also freed resources for fluency.

6.2. Research Question 2

There is evidence in many areas that WMC capacity is predictor of success in cognitively demanding tasks. Nevertheless, research relating L2 oral performance to WMC, such as the present study, has not shown overall significant effects (Fortkamp, 1998; Mizera 2006, Gilabert & Muñoz 2010). A lack of correlation might be due to the intrinsic complex nature of L2 speech production. The process of understanding and conveying messages in a foreign language is itself so demanding that perhaps, adding more complexity on top does not affect remarkably the results. It might be because task complexity does not significantly increase the challenge for WM due to the high demands of L2 oral production; or because even the simple version of the task is so complex that it has already reached a threshold where no more resources are available. A baseline of native speakers to compare patterns of behavior between L1 and L2 would help to understand the reasons why. Another possible explanation is that task completion has not been controlled for. Their performance as fire chiefs is not assessed, only the language used. Yet, participants' involvement in task completion must have played a role in the amount of resources available for oral production.

In any case, the results of this experiment, display a weak correlation, in the complex task, of the Trail Making Task (TMT) with lexical complexity, and also for the reading span task (RST) with accuracy. This partially supports the theory that the higher WMC is, the better performance is. A relevant finding is that, when looking at the impact of WM tasks on gains in performance, TMT$^{\#}$ was correlated with lexical complexity and

accounted for 16.6% of its variance. This could mean that subjects with higher attentional control would allocate more resources to lexical encoding. This reasoning is consistent with Levelt's (1989: op.cit. Kormos & Trebits, in press) idea that formulation processes are lexically driven. Thus, higher attentional resources – administered during demanding cognitive tasks by the central executive (Baddeley, 1986) – would direct attention to lexis available in order to perform better on complex tasks.

Against predictions, the letter span task (LST) yielded multiple positive correlations with CAF measures. Accuracy, in the complex version, did not show a significant correlation, but followed a strong trend (p=.057). Only structural complexity displayed no correlation with the LST, leading us to favor again the previously commented argument that more sensitive measures are needed to assess structural complexity. Hence, the overall correlation of LST (measuring short-term memory) would point to the idea that STM is a different construct from WMC and, as O'Brien et al. (2007: op.cit. Hummel, 2009) concluded, it could be considered a better predictor of oral performance. Furthermore, for Kormos and Sáfár (2008,) the assertion that STM is a key element in oral production would be reasonable as "L2 learners have to store already processed bits of their message in memory while planning or linguistically encoding the next segment of their utterance" (pg.267)

The strong correlation between RST and LST sustains the fact that RST is also tapping into short-term memory. On the same grounds, LST and TMT were not correlated. The lack of correlation between TMT and RST did not comply with our predictions. Further research should look into the matter to confirm whether the attention control measured in the TMT and reasoning in RST load on the same mental component and processes.

It is also feasible to believe that differences in working memory alone, cannot explain the differences in oral production and a combination of variables may be the key (Gilabert and Muñoz, 2010). In this study, the role of proficiency was analyzed displaying no significant relationships. Future research should factor in other intervening variables and individual differences (i.e. self-efficacy, control of emotion, familiarity, planning-time…).

6.3. Research question 3

In line with previous findings (Robinson 2001, Gilabert, 2007 and Gilabert et al., 2009), task complexity had an effect on perceived difficulty, stress[#] and confidence but not on interest and motivation. As Robinson states, the lack of relation between the two latter factors is encouraging for language learning since it implies no loss of interest or motivation when tasks are more cognitively complex, as they approach authenticity of target task demands.

Regarding the effects of affective factors on production (or vice versa as we do not know the direction of the correlation), the most influencing variables appear to be stress[#] and confidence with an overall negative and positive impact on performance respectively. However, in the complex task, only fluency seemed to decrease with higher stress[#] or perceived difficulty.

Previous research encourages us to look further into the affective dimension. Sparks (2009) found evidence that a combination of non cognitive variables (i.e. L1 skills, L2 aptitude and L2 affect explained) explained 66% of the variance in L2 proficiency. Although L2 aptitude alone accounted for 56% of the variance, there was a strong impact of the combination of factors.

Motivation and interest did not have a relevant impact in our study. A reason could be that a decontextualized task in an experiment might not offer the best environment to generate motivation or interest. Yet, longitudinal studies in more adequate contexts might yield different results. Sparks (2009) suggested that motivation is a "driving force" for L2 learning and confirmed a positive correlation between L2 achievement and motivation. Moreover, Robinson (2001) points out that motivation and interest, among all affective factors, might be the most stable, controllable and influencing in syllabus design, thus, further studies should be carried out with more specific instruments of analysis to investigate the effects of these affective variables in L2 acquisition.

7. CONCLUSIONS LIMITATIONS AND IMPLICATIONS

The aim of the present study was to investigate the role of working memory and affective variables on oral performance under increased reasoning demands. In order to achieve this objective, some of the suggestions in previous studies were taken into consideration to try to control for some limitations and improve the operationalization of task complexity.

Firstly, differences in cognitive load (i.e. task complexity) between the simple and the complex task were confirmed by three robust independent measures. To date, previous research has used TOT and perceived difficulty to test the level of complexity of the tasks. STE has been used for this purpose in other fields and this study has incorporated this measure of task complexity in oral production with positive results. Secondly, the complexity of the tasks was also confirmed in a pilot study carried out in the L1 of the participants to exclude L2 being the reason why the task was more complex. Thirdly, two tests were administered to control for proficiency. Moreover, three different WM tests were carried out by the subjects, which tapped into the different components of WM to compare the effect of each one. Finally, as the order of the tasks was counterbalanced to avoid task-effects, the influence of sequence was also analyzed.

The manipulation of reasoning demands in the fire chief task affected oral performance as predicted by Robinson's Cognition Hypothesis but only with significant differences for fluency loss and accuracy gains. It is argued, in the line of Gilabert (2007) and Kormos and Trebits (in press) that the combination of demanding resource-dispersing factors and task conditions together with the limitations of attention capacity might cause trade-off effects in speech production. As these factors indeed happen in combination in real life, further research focusing on the effects of their interaction in performance would help to clarify their impact and the implications of task manipulation for syllabus design, sequence and L2 teaching.

Neither WMC nor proficiency had a relevant impact as mediating factors for speech production. However, it is worth noticing the impact (16.6%) of attention control on lexical complexity gains. As argued in the discussion, this finding, in line with previous

studies (Gilabert & Muñoz, 2010; Mizera, 2006), sustains the hypothesis that, under higher cognitive demands, the executive control directs more attentional resources to lexical encoding in order to achieve better performance as formulation processes are lexically driven. Another promising finding was the overall correlation of STM (measured by LST) with CAF scores, suggesting that STM might play a central role in WM for oral performance and should be further analyzed. As for task complexity, future research combining WMC with other individual differences (e.g. aptitude, intelligence, L2 affect…) would shed more light on the effects of task difficulty.

Finally, the affective variables that seemed to be the most influential for speech production were stress, confidence and perceived difficulty. However, under increased complexity, only fluency was affected by perceived difficulty and stress. More research is needed using more accurate instruments and including other variables, such as anxiety, which was a missing factor in our study but showed an impact in previous SLA research (e.g. Sparks, 2009). Once again, it is probably the combination of L2 affect with other learners' factors that will show a clearer picture of its effects.

Despite the efforts with operationalization, the study still has a number of shortcomings and limitations to be considered for future experiments. As already mentioned above, there were some intervening factors that were not controlled for. Resource-dispersing and directing variables occur simultaneously and the first ones were not factored in. The same applies for other individual differences (i.e. anxiety). Due to time constraints, it was not possible to analyze the transcriptions in the L1 to compare the patterns in performance. It would have helped to distinguish the linguistic behaviors inherent to the use of the L1 from the characteristics of the pure L2 performance patterns regardless of individual behavior in L1. Equally, the qualitative data was not deeply analyzed although we confirmed that it contained relevant information regarding task complexity and affective factors. As for STE, a retrospective[7] protocol was applied, however, it is likely that after the first task, subjects expected that they were going to be asked for a time judgment in the second task, affecting their attention to time estimation and yielding more accurate results. In future studies, subjects should be warned in advance that they will have to give a STE

[7] Participants are not informed beforehand that they will be asked to estimate their time on task

to even the conditions in both tasks. Finally, the spaces where the participants performed the experiment were, in some cases, not the most adequate and it might have affected their concentration.

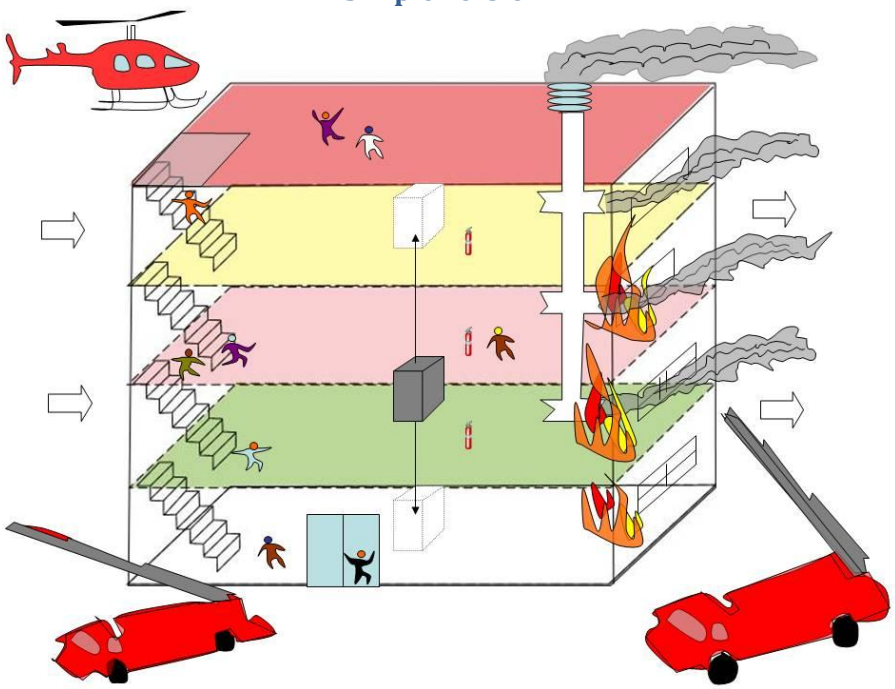
As for the potential implications of the findings in this study, the most direct one is for syllabus design and task sequencing. Robinson (2001) argues that sequencing decisions can be based on task complexity for its "robust and manipulable influence on learners' production" (pg. 51). He also argues that task complexity is preferred over task difficulty because it can be diagnosed in advance (before a language program starts) and it is more stable. Nevertheless task difficulty should be taken into account by teachers for they can assess learners' factors in the classroom. Furthermore, task complexity also has implications for exams since different outcomes can be predicted depending on the type of task.

In addition to the suggestions already commented for future research, longitudinal studies could show the effects of task complexity and difficulty on L2 learning.

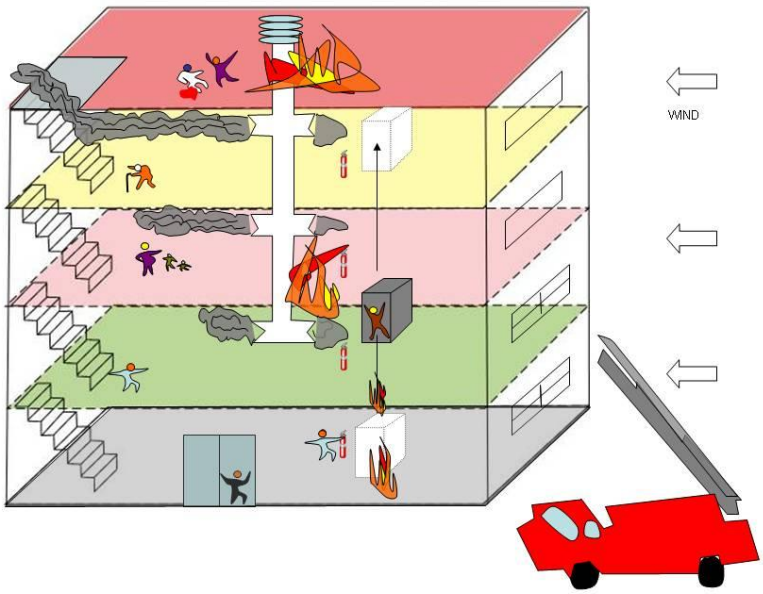
APPENDICES

Appendix 1: Fire chief task

Simple version



Complex version



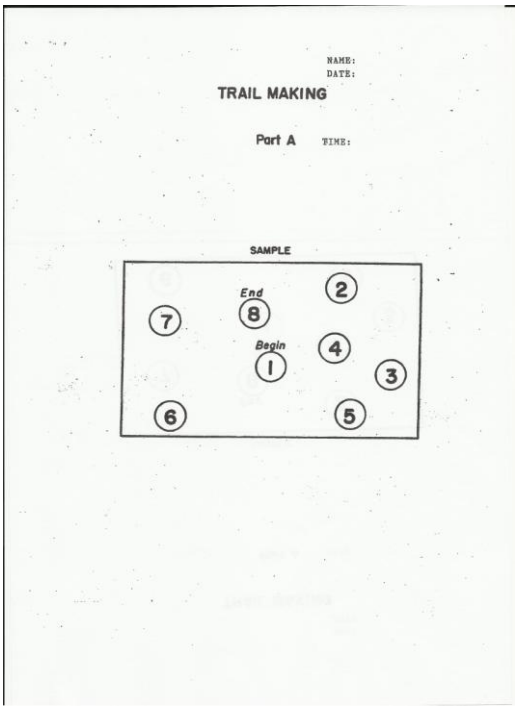
Appendix 2: STE and Affective variables questionnaire

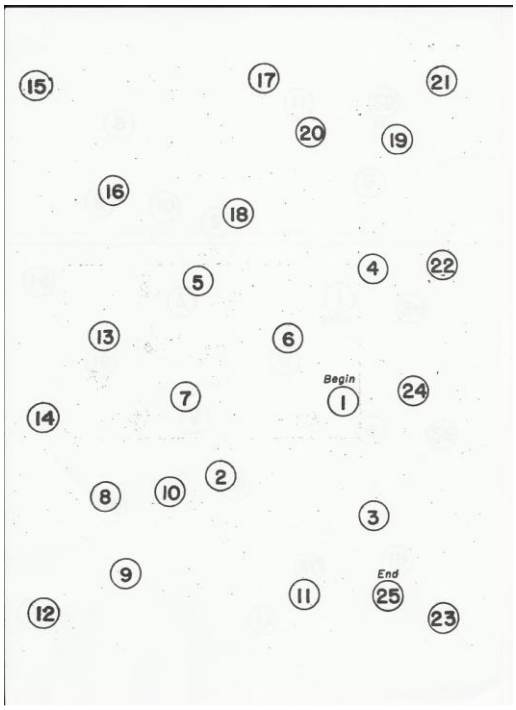
Name: _____ S☐ C☐

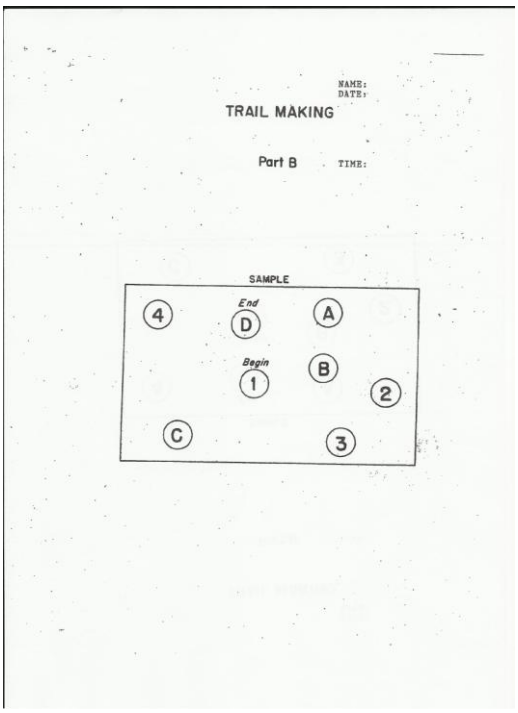
- How long do you think it has taken you to complete the task? _____ (in seconds)
- Now evaluate each task by circling the appropriate answer in each case:

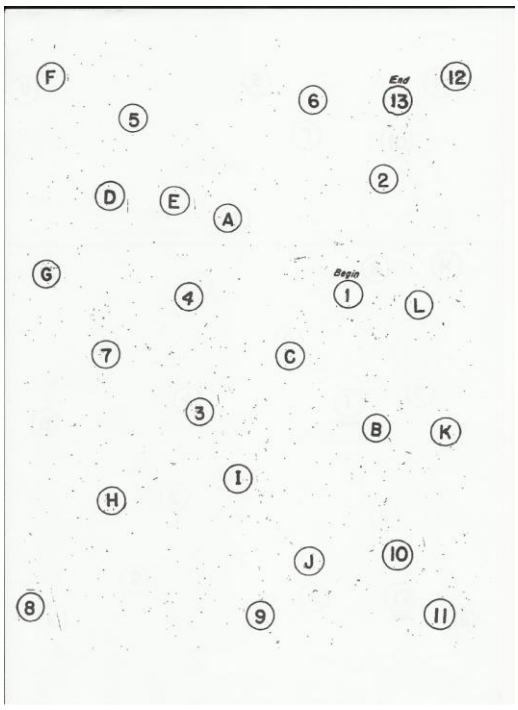
I thought this task was **easy**	1	2	3	4	5	6	7	8	9	I thought this task was **difficult**
I felt **frustrated** doing this task	1	2	3	4	5	6	7	8	9	I felt **relaxed** doing this task
I did **not** do this task **well**	1	2	3	4	5	6	7	8	9	I did this task **well**
This task was **not interesting**	1	2	3	4	5	6	7	8	9	This task was **interesting**
I **don't want** to do **more** tasks	1	2	3	4	5	6	7	8	9	I **want** to do **more** tasks like this like this

Appendix 3: Trail Making Task (TMT)









Appendix 4: Open questions and language background and personal details questionnaire

- In general, did you find it hard to carry out the tasks? Why?
- Was one of the two tasks more complex? Which one? Why?
- Do you want to make any comments about the complexity of the task or about your feelings while performing it?

Age: _____ Telephone or e-mail:_____

Nationality:_____

		Dominant
Mother tongue(s):	1. _____	☐
	2. _____	☐

Foreign language(s) and proficiency: ORAL / WRITTEN

- _____ HIGH MEDIUM LOW / HIGH MEDIUM LOW
- _____ HIGH MEDIUM LOW / HIGH MEDIUM LOW
- _____ HIGH MEDIUM LOW / HIGH MEDIUM LOW
- _____ HIGH MEDIUM LOW / HIGH MEDIUM LOW
- _____ HIGH MEDIUM LOW / HIGH MEDIUM LOW

At what age did you start learning English?_____

How many years have you studied English?_____

Do you use English usually? In which contexts?_____

_____

Have you been living abroad? Where? How long?_____

_____

Did you use English as the main language of communication? For how long?_____

_____

Do you feel comfortable speaking English?_____

Do you like speaking English?_____

Highest level of education attained: _____

Profession:_____

Appendix 5: Informed consent form

Consent Form for Participation in Research

Purpose of this Study
The purpose of the study is to gather data for a study analyzing the effects of task-manipulation in foreign language oral performance and relate it to working memory capacity.

Procedures
You will be asked to complete some tasks, tests and questionnaires. The researcher will give you detailed instructions for each of them.

Rights
Your participation is voluntary. You are free to stop participation at any point.
Your data will be confidential and your identity will be protected. All research data will be assigned a participant code to record them and only the experimenter will have access to the corresponding names.

Optional Permission
1. I understand that the researcher may want to use a short portion of the data for illustrative reasons in presentations of this work, for scientific or for educational purposes. I give my permission to do so provided that my name will not be used.

□ YES □ NO

Right to Ask Questions and Contact Information
If you have any questions about this study, you should feel free to ask them now. If you have any questions later, you should contact the researcher: marienkarecio@gmail.com

Voluntary Consent
By signing below, you agree that the above information has been explained to you and all your current questions have been answered. You understand that you may ask questions about any aspect of this research study during the course of the study and in the future. By signing this form, you agree to participate in this research study.

Name and surname: ……..

……..
Signature

Appendix 6: Transcription Guidelines[8]

- **Obligatory Headers**
  - An Example: This is how it would look like:

@Begin
@Languages: en
@Participants: Subject16-Ana
@Transcriber: Mary Recio
@Sequence: Simple-Complex
@Condition: COMPLEX
@Time on task: 66 seconds
@Content: FIRECHIEF. only one truck

- **The end should be always indicated:**

@End

- **Special Form Markers**

Letters	Example	Meaning
@s	casa@s	for Spanish words
@i	uhhuh@i	for interjections

- **Unidentifiable Material**
- *INV: what ?
  *MAR: **xxx** .

- **The transcriber's best guess** whether this guess is correct or not. *Example:*
  *INV: what ?
  *MAR: milk and butter **[?]** .

- **Exclamations**

Only the ones frequent in our data.

Exclamation	Meaning
Ah	Relief, joy
Ahhah	Discovery
Heehee	Amusement
Mmm	tasty, good
Ugh	disgust, effort
Wow	Amazement

Marker	Function
Hmm	thinking, waiting
:	lengthening
Whoops	Blunder

We will mark interactional markers in the following way: **uhhuh@i**.

[8] We adapted these guidelines from CHAT criteria. We have excluded the codes related to interaction, as our tasks were monologic, as well as the ones that referred events that did not happen in our narratives.

- **Simple Events**

&=cough
&=laugh
&=gasp

- **Complex Local Events**
  # - For pauses

- **Special Utterance Terminators**

- **Trailing Off +…** : utterances which are left incomplete but not interrupted. *Example:*
  *MAR: smells good enough for +…

- **Self-Interruption +//** : a speaker breaks off an utterance and starts up another. *Example:*
  *MAR: smells good enough for +//. what is that ?

- **Paralinguistic Scoping**

- **Stressing (!)** : When the preceding word or string of words are stressed. If what we have is a string of words we will use angle brackets. *Example:*
  *MAR: I said <this book was mine> !

- **Questioning:**
  *MAR: he comes slowly, no ?

- **Explanations and Alternatives**

- **Retracing without correction [/]** : repeated but not self-corrected, *Example:*
  *MAR: <I want to> [/] I want to break free .
  Or repetitions:
  *MAR: <flower flower flower> [/] flower .

- **Retracing with correction [//]** : corrected by the same speaker. *Example:*
  *MAR: I thought [//] I wanted to break free .

- **Retracing with Reformulation [///]**: involve full and complete reformulation. *Example:*
  *MAR: all of my friends [///] we had decided to go home for lunch.

- **False Start without Retracing [/-]** : finishes an incomplete utterance and starts another completely different. *Example:*
  *MAR: <I wanted> [/-] uh when is she coming ?
  Or when we don't know if it was a new idea or not:
  *MAR: then he took <t>[/-] the hammer

- **Compound words (+)**
- *MAR: I would use the fire+extinguisher

Appendix 7: Tables

Table 6

Comparison of simple and complex CAF measures with Wilcoxon Signed Rank Tests

Dependent Variable	*Fluency Rate A*	*Fluency Rate B*	*Accuracy*	*Lexical complexity*	*Structural complexity*	
					Words x clause	*Subordination ratio*
p	.072	.041*	.011*	.116	.428	.120
r	.128	.149	.216	-.103	-.081	.074

* significant at p<.005
r= effect size low 0.1-0.3; medium 0.3-0.5: large >0.5

Table 7

Mann-Whitney U-test. Significance of sequence on CAF measures

Dependent Variable		*Fluency Rate A*	*Fluency Rate B*	*Accuracy*	*Lexical complexity*	*Structural complexity*	
						Words x clause	*Subordination ratio*
p	simple	.026*	.033*	.272	.044*	.351	.054
	complex	.071	.351	.443	.093	.694	.384

* significant at p<.005

Table 9

Spearman's Rank Correlation Coefficient: WM & CAF measures Simple and Complex tasks.

N= 30

Dependent Variable			Simple			Complex		
			TMT#	*LST*	*RST*	*TMT#*	*LST*	*RST*
Fluency	Unpruned Speech Rate A	r	.074	**.560****	.186	.133	**.502****	.118
		p	.349	**.001**	.163	.241	**.002**	.267
	Pruned Speech Rate B	r	-.075	**.583****	.287	-.035	**.458****	.148
		p	.348	**.000**	.062	.427	**.005**	.217
Lexical Complexity	Guiraud's Index	r	-.119	**.563****	.244	**-.356***	**.402***	.172
		p	.266	**.001**	.097	**.027**	**.014**	.182
Accuracy#	Number of errors x 100 words	r	-.001	**-.332***	-.144	-.019	-.295	**-.388***
		p	.497	**.036**	.224	.460	.057	**.017**
Structural Complexity	Number of words x clause	r	.061	-.056	-.149	.178	-.025	.093
		p	.374	.384	.217	.174	.448	.312
	Subordination Ratio	r	-.010	.004	-.106	.022	**.436****	.120
		p	.479	.491	.288	.453	**.008**	.263

r= correlation coefficient; p= significance at p< .05
#counter-intuitive measure

Table 10

Linear regressions: Impact of WM in losses or gains in performance factoring out proficiency

N= 30

	Dependent Variable	*Proficiency*		*TMT*		*LST*		*RST*	
		Signif.	*Impact*	*Signif.*	*Impact*	*Signif.*	*Impact*	*Signif.*	*Impact*
Fluency Loss	Unpruned Speech Rate A	.185	.062	.405	.024	.570	.011	.689	.006
	Pruned Speech Rate B	.151	.072	.423	.022	.514	.015	.814	.002
Lexical Complexity Gains	Guiraud's Index	.917	.000	**.028***	**.166**	.672	.007	.469	.020
Accuracy Gains	Number of errors x 100 words	.225	.052	.995	.000	.175	.064	.271	.042
Structural Complexity Gains	Number of words x clause	.757	.003	.133	.081	.829	.002	.100	.097

*significance at p< .05

Table 12

Comparison of simple and complex affective variables with Wilcoxon Signed Rank Tests

Dependent Variable	*Difficulty*	*Stress*	*Confidence*	*Interest*	*Motivation*
p	.014*	.000*	.003*	.120	.428
r	-.228	.347	.280	-.064	.042

* significant at p<.005

r = effect size low 0.1-0.3; medium 0.3-0.5: large >0.5

Table 13

Spearman's Rank Correlation Coefficient: CAF & affective variables Simple and Complex tasks.

N=30

	Dependent Variable		Simple					Complex				
			D	*S#*	*C*	*I*	*M*	*D*	*S#*	*C*	*I*	*M*
Fluency	Unpruned Speech Rate A	r	-.270	**.734****	**.588****	.176	.147	-.353	.316	.214	-.100	.092
		p	.149	**.000**	**.001**	.351	.437	.056	.089	.257	.598	.630
	Pruned Speech Rate B	r	-.204	**.652****	**.511****	.128	.074	-.412*	.368*	.281	-.262	-.035
		p	.281	**.000**	**.004**	.500	.699	.024	.046	.132	.161	.854
Lexical Complexity	Guiraud's Index	r	-.229	**.408***	**.405***	.103	.056	-.325	.238	.194	-.122	-.118
		p	.223	**.025**	**.026**	.588	.771	.080	.206	.303	.520	.535
Accuracy#	Number of errors x 100 words	r	.095	**-.574****	**-.493****	-.018	-.015	.178	-.325	-.341	.058	-.032
		p	.618	**.001**	**.006**	.924	.939	.346	.080	.065	.761	.867
Structural Complexity	Number of words x clause	r	-.300	.231	.018	.238	.086	-.042	.061	.082	.105	-.131
		p	.108	.219	.927	.205	.650	.825	.749	.666	.580	.491
	Subordination Ratio	r	.104	-.214	.096	**-.372***	**-.509****	-.178	.148	.185	.013	.127
		p	.584	.256	.614	**.043**	**.004**	.348	.435	.329	.946	.503

D= difficulty; *S*= stress; *C*= confidence; *I*= interest; *M*= motivation
r= correlation coefficient; p= significance at p< .05
#counter-intuitive measure

Table 14

Spearman's Rank Correlation Coefficient: Affective factors in Simple and Complex tasks.

N=30

Dependent Variable		Simple					Complex				
		D	*S#*	*C*	*I*	*M*	*D*	*S#*	*C*	*I*	*M*
D	r	1	**-.392***	-.300	-.145	-.087	1	**-.435****	-.315	.108	-.104
	p		**.032**	.107	.444	.648		**.016**	.090	.569	.585
S#	r	**-.392***	1	**-706****	.355	.355	**-.435****	1	**.885****	-.174	-.058
	p	**.032**		**.000**	.054	.054	**.016**		**.000**	.359	.761
C	r	-.300	**-706****	1	.344	.214	-.315	**.885****	1	-.204	-.059
	p	.107	**.000**		.063	.257	.090	**.000**		.279	.756
I	r	-145	.355	.344	1	**.720****	.108	-.174	-.204	1	**.590****
	p	.444	.054	.063		**.000**	.569	.359	.279		**.001**
M	r	-.087	.355	.214	**.720****	1	-.104	-.058	-.059	**.590****	1
	p	.648	.054	.257	**.000**		.585	.761	.756	**.001**	

D= difficulty; *S*= stress; *C*= confidence; *I*= interest; *M*= motivation
r= correlation coefficient; p= significance at p< .05
#counter-intuitive measure

Table 15

Mann-Whitney U-test. Significance of sequence on affective variables

Dependent Variable		*Difficulty*	*Stress*	*Confidence*	*Interest*	*Motivation*
p	**simple**	**.011***	**.007***	.097	**.035***	.084
	complex	.154	.259	.615	.457	.258

* significant at p<.005

LIST OF REFERENCES

Arbuthnott, K., & Frank, J. (2000). Trail making test, part B as a measure of executive control: validation using a set-switching paradigm. *Journal of Clinical and Experimental Neuropsychology , 22* (4), 518-528.

Baddeley, A. (1999). *Essentials of human memory.* East Sussex: Psychology Press.

Baddeley, A. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology , 49A*, 5-28.

Baddeley, A. (1990). *Human Memory: Theory and Practice.* Hove, UK: Lawrence Erlbaum.

Baddeley, A. (1981). The concept of working memory: A view of its current state and probable future development. *Cognition , 10*, 17-23.

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Science , 4* (11), 417-423.

Baddeley, A. (1984). The fractionation of human memory. *Psychological Medicine , 14*, 259-264.

Baddeley, A. (1986). *Working memory.* Oxford: Oxford University Press.

Baddeley, A. (1992). Working memory. *Science , 255*, 556-559.

Baddeley, A., & Hitch, G. (1974). Working memory. In G. Bower, *The Psychology of Learning and Motivation: Advances in Research and Theory* (Vol. 8, pp. 47-90). New York: Academic Press.

Bialystok, E. (2010). Global-local and trail making tasks by monolingual and bilingual children: beyond inhibition. *Developmental Psychology , 46* (1), 93-105.

Bygate. (2008). Quality of language and purpose of task: patterns of learners' language on two oral communication tasks. *Language Teaching Research , 3* (3), 185-214.

Candlin, C. (1987). Towards task-based language learning. In C. Candlin, & D. Murphy, *Language Learning Tasks* (pp. 5-22). London: Prentice Hall.

Carpenter, P., & Just, M. (1989). The role of working memory in language comprehension. In D. Klahr, & K. Kotovsky, *Complex Information Processing: The Impact of Herbert A. Simon* (pp. 31-68). Hillsdale, N.J.: Lawrence Erlbaum.

Carpenter, P., Miyake, A., & Just, M. (1994). Working memory constraints in comprehension: Evidence from individual differences, aphasia, and aging. In M. Gernsbacher, *The Handbook of Psycholinguistics* (pp. 1075-1122). San Diego: Academic Press.

Casini, L., & Macar, F. (1997). Effects of attention manipulation on judgments of duration and of intensity in the visual modality. *Memory & Cognition , 25* (6), 812-818.

Chastain, G., & Ferraro, F. R. (1997). Duration ratings as an index of processing resources required for cognitive tasks. *The Journal of General Psychology , 124* (1), 49-76.

Conway, A. R., Kane, M. J., Bunting, M. F., Hambrick, D. Z., Wilhelm, O., & Engle, R. W. (2005). Working memory span tasks: A methodological review and user's guide. *Psychonomic Bulletin & Review , 12* (5), 769-786.

Daneman, M. (1991). Working memory as a predictor of verbal fluency. *Journal of Psycholinguistic Research , 20*, 445-464.

Daneman, M., & Carpenter, P. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior , 19*, 450-466.

Ellis, D. P. (2010). *The role of task complexity in the linguistic complexity of native speaker output.* Retrieved from http://es.scribd.com/doc/46883995/QP1-Ellis-Task-Complexity

Ellis, R. (2005). *Planning and Task Performance in a Second Language.* Amsterdam: John Benjamins.

Engle, R., Cantor, J., & Carullo, J. (1992). Individual differences in working memory and comprehension: A test of four hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition , 18*, 972-992.

Fink, A., & Neubauer, A. C. (2001). Speed of information processing, psychometric intelligence: and time estimation as an index of cognitive load. *Personality and Individual Differences , 30*, 1009-1021.

Fortkamp, M. B. (1999). Working memory capacity and aspects of L2 speech production. *Communication and Cognition , 32*, 259-296.

Fortkamp, M. (1998). Measures of working memory capacity and L2 oral fluency. *Ilha do Desterro , 35*, 201-238.

Fortkamp, M. (2000). Working memory capacity and L2 speech production: An exploratory study. *Unpublished doctoral dissertation* . Universidade Federal de Santa Catarina.

Foster, P., & Skehan, P. (2008). The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research , 3* (3), 215-247.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *IRAL , 45*, 215-240.

Gilabert, R. (2005). Task complexity and L2 oral narrative production. *Unpublished Ph.D dissertation* . University of Barcelona, Deparment of Applied Linguistics, Spain.

Gilabert, R., & Muñoz, C. (2010). Differences in attainment and performance in a foreign language: the role of working memory capacity. *International Journal of English Studies , 10* (1), 19-42.

Gilabert, R., Barón, J., & Levkina, M. (in press). Manipulating task complexity across task types and modes.

Gilabert, R., Barón, J., & Llanes, À. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *IRAL , 47*, 367-395.

Guará-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.

Hummel, K. M. (2009). Aptitude, phonological memory, ans second language proficiency in nonnovice adult learners. *Applied Psycholinguistics , 30*, 225-249.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A lantent-variable approach to verbal an visuospatial memory span and reasoning. *Journal of Experimental Psychology: General , 133* (2), 189-217.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition , 11*, 261-271.

Kormos, J., & Trebits, A. (in press). *The role of task complexity, modality and aptitude in narrative task performance.* Retrieved from http://eprints.lancs.ac.uk/34247/

Macar, F., Grondin, S., & Casini, L. (1994). Controlled attention sharing influences time estimation. *Memory & Cognition , 22* (6), 673-686.

Mackey, A., & Goo, J. (2007). Interaction and research in SLA: a meta-analysis and research synthesis. In A. Mackey, *Conversational Interaction in Second Language Acquisition* (pp. 407-452). Oxford: Oxford University Press.

Meara, P. M. (2005). Designing vocabulary tests for English, Spanish and other languages. In C. Butler, M. Gómez González, & S. Doval Suárez, *The dynamics of language use: functional and contrastive perspectives* (pp. 271-285). Amsterdam: John Benjamins.

Meara, P., & Milton, J. (2003). *X_Lex: The Swansea Vocabulary Levels Test.* Express, Newbury.

Miyake, A., & Friedman, N. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy, & L. Bourne, *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339-364). Erlbaum: Mahwah, N.J.

Miyake, A., & Shah, P. (1999). Models of working memory: An introduction. In A. Miyake, & P. Shah, *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control* (pp. 1-27). Cambridge, U.K.: Cambridge University Press.

Mizera, G. J. (2006). Working memory and L2 oral fluency. *PhD Dissertation* . University of Pittsburg.

Mota, M. B. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos , 25*, 69-104.

Niwa, Y. (2000). Reasoning demands of L2 task and L2 narrative production: Effects of individual differences in working memory, intelligence, and aptitude. *Unpublished M.A. dissertation* . Aoyama Gakuin University, Tokyo.

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

Pica, T., Kang, H., & Sauro, S. (2006). Information gap tasks: Their multiple roles and contributions to interaction research methodology. *Studies in Second Language Acquisition , 28* (2), 301-338.

Robinson, P. (2001). Task complexity, task difficulty, and task production: exploring interactions in a componential framework. *Applied Linguistics , 22* (1), 27-57.

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics , 45*, 195-215.

Robinson, P. (2011). Task-Based Language Learning: A Review of Issues. *Language learning , 61* (Issue supplement s1), 1-36.

Robinson, P. (2002). The Cognition Hypothesis of task-based L2 development: Theory and research. *Journal of the Korean English Education Society , 2*, 1-26.

Robinson, P., & Gilabert, R. (2007). Task complexity, the Cognition Hypothesis and second language learning and performance. *International Review of Applied Linguistics , 45* (3), 161-176.

Sagarra, N. (2002). The role of syntactic modifications on L2 oral comprehension. In C. Wiltshire, & J. Camps, *Romance Phonology and Variation* (pp. 197-210). Amsterdam: John Benjamins.

Skehan, P. (1989). *Individual Differences in Second Language Learning.* London: Edward Arnold.

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

Skehan, P. (2003). Task-based instruction. *Language Teaching , 36*, 1-14.

Sparks, R. L. (2009). Long-term relationships among early first language skills, second language aptitude, second language affect, and later second language proficiency. *Applied Psycholinguistics , 30*, 725-755.

Thomas, E., & Weaver, W. (1975). Cognitive processing and time perception. *Perception & Psycholinguistics , 17*, 363-367.

Trebits, A., & Kormos, J. (2008). Working memory capacity and narrative task performance. *Proceedings from the 33rd International LAUD Symposium.* Landau/Pfalz, Germany.

Turner, M., & Engle, R. (1989). Is working memory capacity task dependent? *Journal of Memory & Language , 28*, 127-154.

Unsworh, N., & Heitz, R. P. (2005). An automated version of the operation span task. *Behavior Research Methods , 37* (3), 498-505.

Wesche, M., & Paribakt, T. (1996). Assessing second language vocabulary knowledge: Depth vs. breadth. *The Canadian Modern Language Review , 53* (1), 13-40.