TASK COMPLEXITY AND WORKING MEMORY IN PERFORMING LISTEN-TO-SPEAK INTEGRATED TASKS IN A SECOND LANGUAGE

AYŞE GÜL YÜCEL

BOĞAZİÇİ UNIVERSITY

2022

TASK COMPLEXITY AND WORKING MEMORY IN PERFORMING LISTEN-TO-SPEAK INTEGRATED TASKS IN A SECOND LANGUAGE

Thesis submitted to the
Institute for Graduate Studies in Social Sciences
in partial fulfillment of the requirements for the degree of

Master of Arts
in
Foreign Language Education

by
Ayşe Gül Yücel

Boğaziçi University
2022

DECLARATION OF ORIGINALITY

I, Ayşe Gül Yücel, certify that

- I am the sole author of this thesis and that I have fully acknowledged and documented in my thesis all sources of ideas and words, including digital resources, which have been produced or published by another person or institution;
- this thesis contains no material that has been submitted or accepted for a degree or diploma in any other educational institution;
- this is a true copy of the thesis approved by my advisor and thesis committee at Boğaziçi University, including final revisions required by them.

Signature……

Date……

ABSTRACT

Task Complexity and Working Memory in Performing Listen-to-speak Integrated Tasks in a Second Language

This study aimed to explore the effects of task complexity, perceived task difficulty, and working memory capacity (WMC) in listen-to-speak tasks, in second language (L2) English. 40 university students with Turkish as their first language (L1) participated in this study. Adopting the framework offered by the cognition hypothesis, this study manipulated the structural demands of the listen-to-speak tasks using an outline. Listen-to-speak tasks are taken and adapted from the Test of English as a Foreign Language (TOEFL iBT) official website. Perception of task difficulty was measured using a 10-item Likert scale questionnaire. WMC was measured through Turkish translations of operation span (OSpan) and running span (Run Span) tasks. Dimensions of L2 performance i.e., syntactic complexity, lexical complexity, accuracy, fluency, and content were separately measured. Multivariate and univariate repeated measures ANOVA analyses revealed that task complexity had a significant effect on all the above-mentioned dimensions of L2 performance as well as the content. Stepwise multiple regression results demonstrated that task complexity could explain syntactic complexity, accuracy, and content while perceived task difficulty could account for lexical complexity. WMC, on the other hand, could explain fluency. The study concluded that task complexity influences all aspects of task performance; however, its effect is moderated when learner factors (perceived task difficulty and WMC) are considered.

ÖZET

İkinci Dilde Bütünleşik Dinleme-Konuşma Görevlerinin Gerçekleştirilmesinde Görev Zorluğu ve İşler Bellek

Bu çalışma, görev zorluğu ve işler belleğin ikinci dilde bütünleşik dinleme-konuşma görevleri içeren testlerde performans üzerindeki etkisini araştırmaktadır. Araştırmada kullanılan bütünleşik dinleme-konuşma görevleri TOEFL iBT' nin resmi test klavuzundan alınmış ve araştırmacı tarafından uyarlanmıştır. Robinson'nun biliş hipotezi baz alınarak görev zorluğu ve algılanan görev zorluğu kavramları benimsenmiştir. Algılanan işlem zorluğu Robinson'un (2001) anketiyle ölçülmüştür. İşler bellek kapasitesi ise Türkçe'ye uyarlanan işlem uzamı ve akan bellek uzamı testleri ile ölçümlenmiştir. Dinle-konuş test sonuçları sözdizimsel (sentaktik) zorluk, sözcüksel zorluk, dilbilgisel doğruluk, akıcılık ve içerik boyutlarına göre analiz edilmiştir. Çok değişkenli ve tek değişkenli varyans analizi sonuçları görev zorluğunun sözdizimsel (sentaktik) zorluk, sözcüksel zorluk, dilbilgisel doğruluk, akıcılık ve içerik üzerinde etkisi olduğunu göstermiştir. Çoklu regresyon analizi, görev zorluğunun konuşma performansının sözdizimsel (sentaktik) zorluk, dilbilgisel doğruluk ve içerik boyutlarını belirli ölçüde açıklayabildiğini göstermiştir. Algılanan işlem zorluğu ise sözcüksel zorluğu açıklayabilen tek bağımsız değişkendir. İşler bellek kapasitesi, konuşma performasının sadece akıcılık boyutunu bir miktar açıklayabilmiştir ve diğer boyutlarına ölçümlenebilir bir etkisi olmadığı söylenebilir. Sonuç olarak, görev zorluğunun etkisi, algılanan işlem zorluğu ve işler bellek kapasitesi de ele alındığında sınırlı görünmektedir.

ACKNOWLEDGEMENTS

I would first like to express my heartfelt gratitude to my thesis advisor Prof. Gülcan Erçetin whose wealth of knowledge steered this thesis in the right direction. Her expert guidance and invaluable comments enlightened me during the thesis process.

I would like to extend my gratitude to my thesis co-advisor Assist. Prof. Mehmet Akıncı for his invaluable advice, continuous support, and patience during this process.

I would like to offer my sincere thanks to my thesis committee members Assist. Prof. Şebnem Yalçın and Assist. Prof. Fidel Çakmak for their insightful comments and encouraging feedback.

I am gratefully indebted to my friends Henrieta Krupa and Gül Bursa for their invaluable help and support during the data collection and data coding processes of the thesis. I want to thank my dear friend Gözde Aydın for accompanying me on this journey.

Finally, my deepest thanks and gratitude go to my mother Billur, my sister Nurgül and my brother Mustafa for their unwavering support and belief in me.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Integrated tasks, in which test-takers employ at least two language skills (a combination of receptive and productive) to complete a task, have become integral to L2 testing and assessment, including standard tests such as TOEFL iBT, Certificate in Advanced English (CAE), and Pearson Test of English (PTE). These tasks (e.g., listen-to-speak, read-to-write) have been increasingly adopted especially because they are considered more authentic unlike discrete point tests (e.g., multiple-choice tests); thus, they have more potential to draw on the ability to use the language (Alderson, Clapham & Wall, 1995; Asención, 2004; Cumming, Grant, Mulcahy-Ernt, & Powers, 2004). Moreover, integrated tasks were shown to provide positive washback as they motivate classroom teachers to train students in language skills rather than training them in test-taking skills (Weigle, 2004). For instance, listen-to-speak tasks were argued to allow for a more comprehensive representation of the construct of listening and how the language is used in daily communication (Barkaoui, Brooks, Swain & Lapkin, 2013; Frost, Elder & Wigglesworth, 2011). These tasks were also suggested to reflect the educational literacy activities conducted in academic settings (Barkoui et al., 2013). More importantly, they were claimed to reduce the influence of background knowledge as test-takers produce the content, in written or spoken response, from the provided input (Cumming et al., 2004). Considering the above-mentioned points, an integrated skills task was chosen to be investigated in this study.

Two main fields of research that constitute the background of this study are Task-Based Learning and Teaching (TBLT) in Second Language Acquisition (SLA) and

the construct of working memory (WM) in cognitive psychology. One of the most detailed task taxonomies, the Triadic Componential Framework (TCF) by Robinson (2001), forms the basis of task design in this study. It offers the most comprehensive framework and characterizations of task demands under the categories of cognitive, interactive, and learner factors. The TCF addresses the task factors and provides details of task demands, which guides the process for task design and complexity. It also addresses the learner factors (e.g., motivation, cognitive abilities, anxiety) which mediate the task performance. In line with the tenets of the TCF, in this study, the task complexity was operationalized using an outline for note-taking.

Considering the complexity of language processing and production involved in integrated task performance, there is a need for further research to clarify the role of learner constraints (aptitude, anxiety, cognitive abilities) on task accomplishment. Working memory capacity (WMC) referring to an individual's capacity to store, retrieve and process information (Baddeley, 2003) is one of the most widely studied learner-internal factors in SLA literature. However, to the best of our knowledge, it has not been investigated through an integrated speaking task. WMC was chosen as the learner factor to be probed in this study because previous studies reported inconsistent findings. The shared assumption of the previous studies was that high WMC is beneficial for language performance; however, the findings reported inconsistencies concerning the conditions in which WMC effects can be observed. Some of the conditions reported were task complexity (Kormos & Trebits, 2011), planning time (Ahmadian, 2012; Baralt, 2010; Guará-Tavares, 2011), and L2 proficiency (Robinson, 2005). For instance, high WMC benefitted only lower-level learners in morphosyntactic development (e.g., Serafini & Sanz, 2016). In another study, high WMC correlated with enhanced oral production only

in more demanding tasks (e.g., Kim, Payant & Pearson, 2015). WMC seems to interact with L2 performance in general but task complexity and other learner variables in particular.

In line with the purposes of this study, the interaction between test-takers and the task was also explored. As Breen (1987) put forward learners "evaluate, respond to, act on the task and their momentary performance" while engaging in a task, the process of integrated task completion cannot fully be understood without the learners' perspective (p. 24). Previous research also indicated that learner perspective can shed light on the nature of task performance (e.g., Weir, 2005), and their perception can influence their performance (e.g., O'Sullivan & Weir, 2011). Prior studies also measured the perceived task difficulty to ensure that their methodological operationalizations align with the learner perceptions (e.g., Gilabert, 2007; Révész, Ekiert & Torgersen, 2016). Accordingly, in this study, participants' perception of task difficulty was measured through a questionnaire adopted from Robinson (2001) to show the effect of perception on task performance and to ensure the operationalization of task complexity.

Given the increasingly common use of integrated tasks for assessment purposes, the inconclusive findings related to the role of WMC in L2 spoken performance, and the lack of research examining the effect of task complexity and WMC in integrated speaking tasks, there is a clear need for further research. Therefore, this study set out to explore the effect of task complexity, perceived task difficulty, and WMC on L2 performance in listen-to-speak tasks. The research questions probe whether task complexity [+/− outline] affects task performance in listen-to-speak tasks and how much of the variation in task performance can be explained by task complexity, WMC, and perceived task difficulty.

CHAPTER 2

LITERATURE REVIEW

This chapter will provide the theoretical background of the study at hand. First, the adopted theoretical framework of listening comprehension and speech production will be briefly presented. Then, the conceptualization of WM will be discussed. It will be followed by the theoretical framework for task complexity and a review of previous studies exploring task-related factors. After that, an overview of integrated skills assessment and listen-to-speak tasks will be given. A brief conceptualization of CAF (complexity, accuracy, fluency) will be also presented as CAF provides the framework for performance scoring. Finally, the research questions and the hypotheses will be provided.

2.1 Listening comprehension

SLA listening research has relied upon several approaches to investigate the listening construct. The cognitive approach attempted to define the cognitive construct of listening comprehension by focusing on cognitive processes or levels of listening. In line with previous research, Anderson's cognitive model of listening comprehension was adopted in this study.

2.1.1 Anderson's cognitive model of listening comprehension

The model proposes three phases of language comprehension which are perception, parsing, and utilization (Anderson, 1995). In the case of listening comprehension, the perception phase entails decoding acoustic signals, detecting sounds, and grouping them

into meaningful representations i.e., words. In this phase, the role of memory is evident as perceived information should be registered to auditory sensory memory or echoic memory to be analyzed as unregistered information decays without being analyzed. The concept of selective attention becomes relevant as it may be directed to specific aspects of context that may aid the decoding of input such as pauses and acoustic emphases.

In the parsing phase, spoken text is segmented into meaningful units of information, which are called 'propositions' or 'chunks of information'. Parsing relies upon syntactic knowledge i.e., perceived words are partially mapped onto a grammatical structure with the help of syntactic cues such as word order, function words, and so on. It also relies on semantic knowledge, i.e., plausible semantic interpretations of words can guide parsing as well as syntactic cues.

In the utilization phase, parsed propositions are combined with listeners' external knowledge. Two levels of semantic processing are involved at this stage i.e., microstructure and macrostructure. At the microstructure level, individual propositions are conceptually linked with each other while at the macrostructure level propositions are conceptually connected to the theme of a text to form its discourse meaning. Inferencing and linking linguistic information with world knowledge and elaboration are characteristic of this phase.

2.1.2 Listening comprehension in L2

It should be noted here that Anderson's model attempts to explain L1 comprehension, which does not mean that it is irrelevant to L2 comprehension as there are certain similarities postulated between L1 and L2 comprehension. Previous research has shown that cognitive processes involved in L1 and L2 comprehension are fundamentally similar

although L2 learners experience certain linguistic and sociolinguistic constraints (Faerch & Kasper, 1986). L2 listening comprehension is defined as "the ability to process extended samples of realistic spoken language, automatically and in real time, to understand the linguistic information that is unequivocally included in the text, and to make whatever inferences are unambiguously implicated by the content of the passage" (Buck, 2001, p. 114). As its definition also demonstrates, L2 listening comprehension entails the interpretation of decoded input using linguistic knowledge and world knowledge (Vandergrift & Baker, 2018). Comprehension suffers when the relevant knowledge or automaticity is lacking. Vandergrift and Baker (2018) listed "L2 vocabulary knowledge, L1 vocabulary knowledge, working memory, auditory discrimination, background knowledge and metacognition about listening" as the mediating variables in L2 comprehension.

There is evidence to support the presence of three phases i.e., perception, parsing, and utilization in L2 comprehension as well (O'Malley & Chamot, 1990). Anderson's cognitive framework has been adopted in L2 listening studies (e.g., Goh, 2000) as it helps identify the places in cognitive processing where listeners have difficulties with comprehension and employ strategies to successfully deal with them. In sum, Anderson's model provides a comprehensive framework of cognitive processing in both L1 and L2 comprehension.

2.2 Speech production

Several models were developed to account for speech production; however, Levelt's model has become the most influential one in SLA research. The model was developed in 1989 and has gone through some revisions (Levelt, 1993; Levelt, Roelofs & Meyer,

1999). Previous research on task complexity also employed Levelt's model, so using this model enables comparisons with previous studies. Levelt's model is also relevant to this study as it accounts for other processes such as attention and memory mediating language processing.

2.2.1 Levelt's model of L1 production

Levelt's (1993) model viewed speech production as modular and involving several processing components acting autonomously in the system. Despite minor changes in the terms used in the 1999 version of the model, the 1993 version proposed three main systems i.e., conceptualizer, formulator, and articulator involved in speech production. These three systems also overlap with the three phases of listening comprehension i.e., perception, parsing, and utilization proposed by Anderson (1995).

The conceptualizer is the putative place where the propositional content of the message is planned and developed. While planning the message, speakers draw on their knowledge of the discourse, situation, and the external/internal world. The planning involves macro-planning and micro-planning. Macro-planning is used to refer to content preparation, focusing on the speech act and considering the context of speech and its requirements (e.g., level of formality). Micro-planning, on the other hand, refers to linguistic decisions (e.g., tense) made by the speaker considering the available information. The outcome from these stages of planning i.e., the preverbal message is sent to the formulator.

In the formulator stage, speakers transform the preverbal messages into linguistic forms using the mental lexicon. Speakers start grammatical encoding by accessing and selecting lemmas (i.e., knowledge of the word and its syntax) in the mental lexicon. The

outcome of this step is called a ‘surface structure’ and it is passed on to the phonological encoding which entails the use of lexemes (i.e., morphological and phonological properties of the word). In the phonological encoding, surface structures are transformed into ‘internal speech’ also known as a ‘phonetic and articulatory plan’ to be passed on to the articulatory system.

The internal speech is converted into the spoken language in the articulator. The putative place where internal speech is temporarily stored to be executed is called the ‘articulatory buffer’. Internal speech, before articulated as overt speech, is stored and retrieved as chunks in the articulatory buffer.

As Kormos (2006) underlined, the processing components in this model are hypothesized as specialists in specific functions i.e., they do not share their processing functions. Processing is also assumed to be incremental i.e., each component moves to processing the next part of information while the previous is still being processed. For incremental processing to take place, all components work simultaneously, which is also called parallel processing. The great speed of L1 speech production is attributed to the automatized nature of the processing. The formulator and articulator are considered to operate with no conscious awareness and take up little attentional resources whereas the conceptualizer requires attention and conscious effort particularly to generate the message and monitor the whole processing. Speakers can monitor what they want to say and access the mental lexicon to link the information to the preverbal message. Monitoring is available at the conceptualizer stage to check whether the preverbal message matches the communicative intention. In the formulator stage, monitoring is involved in parsing and matching the internal speech against the mental lexicon. Skehan (2014) underlined that morphosyntactic accuracy of speech performance is directly

related to the quality of monitoring. In the articulator stage, speakers listen to the overt speech, parse, and monitor it for meaning and form. In case of an error, the cycle repeats itself resulting in self-repairs.

2.2.2 Speech production in L2

Levelt's (1993) model is developed to account for L1 speech production; however, as Skehan (2014) suggested "it has to be the starting point for a credible analysis of psycholinguistic processes involved in L2 speaking" (p.4). Levelt's model provides a comprehensive framework to understand L2 production. L1 and L2 production mechanisms are fundamentally the same (De Bot, 1992) but still, it is necessary to account for the differences between L1 and L2 speech production, which is often claimed to be 'hesitant', 'less accurate', and 'more fragmented'. Three features of L2 production set it apart from L1 production, which is incomplete knowledge of L2, a lack of automaticity, and the likelihood of code-switching (Poulisse, 1997).

Incomplete knowledge of L2 is often referred to as the reason for lexical or grammatical errors in L2 spoken performance (ibid.). Even though L2 speakers do not lack explicit knowledge of certain linguistic features of L2, they tend to continue making errors which are attributed to the lack of automaticity. As automaticity is lacking especially in the formulator and articulator stages, L2 spoken production may require more conscious attention (Kormos, 1999). Consequently, L2 speech production may take place in a more serial fashion unlike parallel processing involved in L1 speech production (Poulisse, 1997). L2 speech production demands more attentional resources, which are limited, and consumes them in the encoding and articulatory stages leaving fewer resources for monitoring. This results in an erroneous use of linguistic features

which could otherwise be easily corrected (ibid.). Codeswitching during L2 speech production is attributed to one 'shared' mental lexicon where lexical items are stored and retrieved from (De Bot, 1992). It is likely for L2 speakers to borrow word forms from their L1 i.e., codeswitching when they experience problems with finding the relevant words to express a specific concept in L2 due to insufficient knowledge (ibid.).

2.3 Working memory (WM)

The models of listening comprehension and speech production presented above explicitly refer to the role of cognitive resource capacity. They also acknowledge that individuals have different cognitive resource capacities, and this may affect language comprehension and language production. In line with the purposes of this study, the cognitive resource capacity referred to in this study is WMC. The reason for choosing WMC to explore individual differences is that it is often referred to as fundamental in understanding why people perform differently in a variety of real-world tasks (Engle, 2001). Previous research has linked WMC to higher cognitive tasks such as math skills, verbal reasoning skills, and language comprehension (e.g., Baddeley 2003; Conway et al. 2005; Just and Carpenter 1992). More specifically, it has been suggested that WMC is a good predictor of both L1 and L2 comprehension and production (e.g., Daneman & Carpenter, 1980; Daneman & Merikle, 1996; Harrington & Sawyer 1992; Mackey et al. 2010; Sagarra, 2007).

2.3.1 The definition of WM

WM refers to the system involving active long-term memory traces, skills, and controlled attention which is maintained 'active' through certain procedures (Engle,

Tuholski & Conway, 1999). It is also described as a distinct and independent memory store that is responsible for manipulating, managing, and transforming the information taken from short-term or long-term memory (Cowan, 2008). Short-term memory is often referred to as passive storage limited in terms of storage capacity and storage duration. Long-term memory, on the other hand, is considered to be consolidated and a putative place from which information is retrieved. Even though various models of WM attempted to define it, it is basically seen as an interface between perception, short-term memory, and long-term memory which is actively involved in complex cognitive activities such as decision making, problem-solving and language processing (Miyake & Shah, 1999).

2.3.2 Baddeley's model of WMC

The most widely cited model of WM is the multicomponent memory system (Baddeley, 1986; Baddeley & Hitch, 1974). It identified the distinct role of WM, which sets it apart from short term-memory, through the studies with aphasic patients with damaged short-term memory. This research led to a solid and empirically ratified blueprint of the WM hypothesis. The original model was comprised of three components i.e., the central executive, the visuospatial sketchpad, and the phonological loop. In its final version, the episodic buffer is added to the system (Baddeley, 2003).

The phonological loop also called the articulatory loop is one of three subsystems or slave systems supervised by the central executive. It is the subsystem that processes verbal information or auditory stimuli and transforms it into a vocal or sub-vocal speech. It also has two sub-components i.e., a short-term phonological store and an articulatory rehearsal component. The short-term phonological store, as the name

suggests, is a temporary store that keeps the memory items for a couple of seconds which are constantly refreshed by the articulatory rehearsal component. In addition to storing verbal information, the phonological loop also processes visual stimuli which can be stored as verbal information and rehearsed accordingly. The phonological loop is proved to be critical in vocabulary learning in young children; thus, it is also suggested to be critical in L2 vocabulary learning (Baddeley, 2003).

The visuospatial sketchpad is the subsystem involved in receiving visual or spatial information, and temporarily storing and manipulating it. The visual and spatial components are also hypothesized to be separable; thus, two sub-components are proposed i.e., visual cache which processes the information about color and shape, and the inner scribe which processes spatial information on movement (Klauer & Zhao, 2004).

The episodic buffer is the postulated subsystem that is responsible for the interaction between the phonological loop and the visuospatial sketchpad. It basically explains how it is possible to integrate information stored in separate subsystems into a single representation. Integration of this information can also entail access to long-term memory. The episodic buffer is a temporary store as well and its capacity is limited (Alan, Baddeley & Hitch, 2006).

The central executive is postulated as one main component supervising the subcomponents and involved in executive processes and the attentional control of WM. Attentional control is described as the mechanism directing and maintaining the flow of information related to the task at hand and suppressing irrelevant information (Baddeley, 2003). In other words, it controls shifts of attention between different kinds of information regarding the task at hand, monitors the process, and searches for solutions

to problems. The central executive is limited in its attentional resources, which is a cognitive resource that mediates the flow of information in the executive system. Attention has many roles but the most relevant one, considering the context of this thesis, is that it oversees conscious control mechanisms involved in L2 speech processing that are yet to be automated (De Bot, 1992). Its role is critical in learning as attention to input is necessary for input to be processed in the relevant components of memory for hypothesis forming and testing purposes by the learner (e.g., Schmidt, 2001). Individual differences in complex WM span are mostly attributed to the executive control and measured by complex span tasks such as remembering letters/words while performing mathematical calculations or reading out sentences and remembering the last word in each sentence.

2.3.3 Alternative models of WMC

Alternative WMC models view WM as a domain-free and unitary system of processing and storage (e.g., Just & Carpenter, 1992). Cowan (1999) on the other hand provided a two-tier structure for WM which involves the focus of attention embedded within long-term memory processes. The focus of attention is limited in its capacity whereas the activated part of long-term memory is not limited but is prone to interference and decay. In Engle et al.'s (1999) view, WM is basically equivalent to the executive attention which is different from short-term memory and a part of general intelligence. In this alternative view of WMC, WM would equate with the active portion of long-term memory along with executive and attentional control mechanisms (Conway et al., 2009).

Despite the differences between models, three unifying characterizations can be derived from them. First and foremost, WM is limited in its capacity. Research tends to

agree that it is a sub-memory system processing limited cognitive resources (Conway, et al., 2007). This limited capacity is manifested in our ability to maintain a limited amount of information in the focus of attention (Cowan, 1998) or in our immediate consciousness (Baddeley, 1992) and the time course of memory and decay of information (Cowan, 2014). This holding capacity has been speculated to be around seven units of information (e.g., Miller, 1956) and four units or chunks of information (e.g., Cowan, 2010). The information is temporarily stored in the WM and is held there for a few seconds before decaying gradually (Cowan, 2014).

The second unifying characteristic of WM is that it includes multiple mechanisms and executive functions (Miyake & Shah, 1999). Most researchers agree that WM consists of domain-specific storage components such as the phonological short-term memory, the visuospatial store, and the episodic buffer. It also encompasses domain-general executive functions such as updating information, switching between tasks, and inhibiting task-irrelevant information (Williams, 2012). The phonological short-term memory or the phonological WM and the central executive component of WM are particularly important as studies have shown them to be most relevant in L1 and L2 language learning and processing (e.g., Linck et al., 2014; Williams, 2012).

Finally, long-term memory is integral to the WM system and constitutes its underlying foundation. WM is the putative workspace where the bidirectional flow of information takes place between temporary storage components (i.e., the phonological store, the visuospatial sketchpad, and the episodic buffer) and long-term memory (Wen, 2015).

2.3.4 An integrated framework of WM in SLA

An integrated framework of WM specifically developed to guide SLA research is comprised of three key components, which are the definition of WM in SLA, WM components, and their related mechanisms and functions that are related to language and WM assessments in SLA (Wen, 2015).

The working definition of WM adopted by the framework is that it is limited in capacity and consists of multiple mechanisms and processes involved in L2 domains and activities. WM-SLA framework is concerned with the WM components that are shown to be directly involved in language learning and processing i.e., the phonological WM and the executive WM. In other words, the visuospatial WM or the episodic buffer are excluded in the illustration of the model even though some studies showed that they may be involved in language processing. The framework also addresses the issue with the variety of WM tasks and proposes using separate memory span tasks to measure language-related components of WM i.e., the phonological working memory and the executive working memory. The framework suggests the use of simple WM span tasks (e.g., non-word repetition span task) to measure the phonological WM and complex WM span tasks (e.g., operation span task) to measure the executive WM. Finally, it underlines a bidirectional relationship between WM and long-term memory. L2 proficiency/ knowledge residing in the long-term memory acknowledges metalinguistic knowledge along with L2 lexical and grammatical knowledge.

Extant research on WM has shown that it is a significant component of language learning, comprehension, and production. More specifically, individuals with larger WMC are better at reading and listening (e.g., Jiang & Farquharson, 2018; Serafini, 2022). They are also better at learning vocabulary in L1 and L2 (e.g., Atkins &

Baddeley, 1998; Baddeley, 2003; Engle, 2001; Malone, 2018). WMC was found to correlate with syntactic comprehension ability, which is an important part of language aptitude (Miyake & Friedman, 1998). Larger WMC was claimed to help L2 learners inhibit L1 transfer and enhance the accuracy of their production (Trude & Tokowicz, 2011).

The relationship between L2 proficiency and WMC has been a controversial issue due to methodological problems associated with language-dependent WM tasks performed in L2 (Gass & Lee, 2011). It has been suggested that language-dependent WM tasks are measuring L2 proficiency rather than WMC (Wen et al., 2015). Therefore, within this framework of WM, WM measures are carefully evaluated and selected.

2.3.5 The role of WMC in L2 listening comprehension

As for the studies investigating the relationship specifically between listening comprehension and WMC, Miyake and Friedman (1998) demonstrated that WM contributes to L2 listening comprehension as learners with larger WMC were able to use syntactic cues in L2 spoken discourse more effectively and to their advantage. Goh (2000) analyzed learners' self-reports (diaries and semi-structured interviews) and identified the problems learners have while listening. Learners reported they forget what they hear quickly, which was attributed to limited WMC by the researcher. It was basically due to overloading WM as it is required to store the old input and take up the new input simultaneously. Another study by Sakuma (2004) investigated the relationship between WMC and language comprehension. In this study, WMC was measured by a listening span test and language comprehension was measured by a proficiency test. The results showed a positive correlation between listening span test scores and scores in

both listening and reading comprehension sections of the test. In Shanshan and Tongshun's (2007) study, the effect of WMC specifically on listening comprehension was investigated. WMC was measured by a listening span test and a listening test, taken from College English Test (CET 4) was given. The results demonstrated that participants with higher listening spans performed better at the listening test. It should be noted here that the WMC measurement tools in both Sakuma's and Shanshan and Tongshun 's studies are rather domain-specific which may actually account for the positive correlation between WMC and listening comprehension. Similarly, Kormos and Sáfár's (2008) study measured both verbal short-term memory (measured by a nonword span test in L1) and general WMC (measured by a backward digit span task). An L2 proficiency test was used to measure comprehension. The L2 proficiency test results showed almost no correlation between the nonword span tasks and the scores in reading and listening. However, the correlation between general WM scores and proficiency scores was positive and statistically significant (r = .37 for listening). On the other hand, Andringa et al. (2012) examined linguistic and nonlinguistic factors in listening comprehension and showed that WMC is a weak predictor of listening comprehension when other factors are considered. Other factors in question are intelligence, processing speed, and topical knowledge. The findings reported a positive correlation between WMC and listening comprehension (r = .32) for L2 speakers; however, when other factors were included in the analysis, there was no measurable effect of WMC. Thus, they proposed that the relationship between WMC and listening comprehension is not straightforward. Vandergrift and Baker (2015) pointed at the indirect influence of WMC on listening comprehension as it affects the development of L1 and L2 vocabulary knowledge. Brunfaut and Révész (2015) pointed to an issue related to testing listening

comprehension. In their study, they found a significant correlation between WMC (measured by auditory forward and backward digit span tasks) and listening scores in PTE. However, there was no significant relationship between WMC scores and the 30-item passage completion multiple-choice test. These findings were attributed to the differences in the two test types, in other words, PTE listening test requires local comprehension whereas the passage completion test requires global comprehension. This can explain the different findings reported in the studies mentioned above. In a more recent study, Masrai (2019) showed that WMC (measured by a listening span task) is a strong predictor of listening comprehension (measured by the IELTS) combined with vocabulary knowledge (r = .64 for listening). The methodological issues mentioned above are also apparent in this study as well. In sum, the tasks used to measure WMC, and the tests used to measure the listening comprehension can potentially affect the results.

2.3.6 The role of WMC in L2 speech production

Research investigating L2 speech production and WM has demonstrated that WMC plays a role in L2 speech production. Fortkamp (1999) studied the relationship between WMC and L2 speech production in terms of fluency, accuracy, complexity, and lexical density. They used a picture description and a narrative task as speech production tasks and WMC was measured by a speaking span task. Results demonstrated that participants with higher spans produced more fluent, accurate, and grammatically complex speech in both tasks. However, the results did not differ significantly when it comes to lexical density. This was attributed to the role played by WM in controlled processing activities which is grammatical encoding in L2 speech production. Following Levelt's (1989)

terminology, grammatical encoding is the process that follows lexical retrieval, and it qualifies as a controlled processing activity as it entails controlled attention to activate information, maintain, inhibit irrelevant information and monitor for errors (Fortkamp, 1999). In other words, the role of WMC becomes more obvious after the words are retrieved from the mental lexicon. Similarly, Mota (2003) reported significant positive correlations between WMC scores (measured by speaking span test) and fluency in L2 speech (measured by mean length of run). Both studies given above were criticized for their choice of span task used to measure WMC, which is a speaking span task, as it also measures L2 proficiency considering the nature of speaking span tasks (Juffs & Harrington, 2011). Mizera's (2006) work on L2 fluency and WMC (measured by an L1 math span test and an L1 speaking span test) reported a weak correlation, which was attributed to the proficiency level of participants as they were advanced learners in this study. The results were explained in this research suggesting that WM does not play a role in this stage of L2 learning as advanced learners produce speech in more automatic processing (ibid.). More studies reported a significant correlation between L2 speech rate and WMC measured by language-independent (e.g., digit span test given in L1) and L1-based measures (e.g., reading span test in L1) (e.g., Gilabert & Muñoz, 2010; Trebits & Kormos, 2008). The research has focused on the executive WM while few others have also focused on the phonological short-term memory (PSTM). Kormos and Sáfár (2008) measured the PSTM by a non-word recall test and the executive WM by an L1 backward digit span task in two different proficiency levels i.e., pre-intermediate and beginner. For the pre-intermediate group, they reported a significant correlation between the PSTM scores and fluency and vocabulary range in a test which employs subjective marking. There was no significant correlation reported for the beginner group. On the other hand,

the executive WM scores were positively related to the speaking exam scores in the beginner group but not in the pre-intermediate group. It can be suggested that the PSTM and the executive WM contribute to speaking performance in different ways depending on test-takers' proficiency level. A more recent study by Wen (2016) measured PWM (with a speaking span test) and PSTM (with a non-word repetition test) and L2 speaking (with a video-clip retell task). He showed that PWM scores were significantly correlated with lexical and syntactic complexity dimensions of performance whereas PSTM scores were not correlated with any of the aspects of performance indicating to the role of the tasks employed to measure WMC. Task complexity as in procedural demands of the task (e.g., planning time) is also shown to affect the relationship between WMC and speech production (e.g., Baralt, 2010; Kormos & Trebits, 2011;). Guará-Tavares (2011) and Ahmadian (2012) showed that WMC gives an advantage in speech production when planning time is given. It is suggested that planning time allows WM to activate relevant information and suppress irrelevant ones. In sum, the selected WMC tasks, proficiency level of participants, and cognitive and/or procedural demands of the tasks (i.e., task complexity) appear to mediate the effect of WMC on speaking.

2.4 Task complexity

Task complexity, also referred as task difficulty, is an important tenet of TBLT as it is employed as a rationale to sequence pedagogic and testing tasks (Long, 2015). Task complexity or task difficulty in the respective frameworks of TBLT does not refer to linguistic or content related characteristics of the task. Rather it is employed to refer to the complexity of the task itself (ibid.). In the following section, Skehan's (1996) trade-

off hypothesis and Robinson's (2001) cognition hypothesis will be discussed to conceptualize and operationalize task complexity in this thesis.

2.4.1 Skehan's trade-off hypothesis

The trade-off hypothesis assumes that attentional resources are limited and real-time communication constraints these resources, which leads to the prioritization of certain aspects of production over others during language use (Skehan, 1996). CAF is posited as the areas that compete for attentional resources. Fluency is defined as "the capacity to use language in real-time" emphasizing meaning and possibly relying on lexicalized systems (Skehan & Foster, 1999, p.96-97). Accuracy is "the ability to avoid error in performance" which may be reflecting higher levels of control in the language and the ability to avoid complex structures that may induce errors (ibid.). Complexity/range is "the capacity to use more advanced language" and "willingness to use fewer controlled language subsystems" (ibid.).

Drawing on the processing perspectives such as Van Patten's (1990) who argued that learners prioritize meaning over form due to limited attentional resources, Skehan (1996) argued that fluency is more likely to be prioritized over accuracy in language performance. As meaning is associated with fluency, complexity and accuracy are equated with form and emerge as the areas that compete for limited cognitive resources (ibid.). In other words, either accuracy or complexity can accompany fluency, the default priority in task performance, but not both. The trade-off is postulated to be more easily observed in demanding tasks as they put extra demands on limited cognitive resources. Demanding tasks or task demands are also called task difficulty. Skehan's (1998) model of task difficulty (Table 1) provides the framework through which task demands are operationalized.

Table 1. Model of Task Difficulty

Code complexity	Cognitive complexity	Communicative stress	Learner factors
Linguistic complexity and variety Vocabulary load and variety	Cognitive familiarity Familiarity of topic Familiarity of discourse genre Familiarity of task Cognitive processing Information organization Amount of computation Clarity of information Sufficiency of information	Time pressure Scale Number of participants Length of text used Modality Stakes Opportunity for control	Learner's intelligence Breadth of imagination Personal experience

Source: Skehan (1998, p. 107)

As proposed earlier in the limited attentional capacity model by Skehan and Foster (2001), the trade-off hypothesis views attention and memory as limited in capacity. Accordingly, demanding tasks tax the attention and memory sources leading to some decay in certain aspects of performance. Skehan and Foster (2001) suggested that increasing cognitive demands of the tasks would force learners to devote their attentional resources to the content rather than the linguistic forms. Attention, in their view, is not selective and voluntary i.e., one cannot have control over their attention and choose to focus or ignore a stimulus.

Skehan (2009) also explained how limited attentional resources affect spoken performance stating that L2 speakers' mental lexicon is not extensive and organized; therefore, it takes more time for L2 speakers to retrieve accurate linguistic forms while formulating the speech. Thus, they put more effort into the formulator to attend to lemma retrieval and syntax building. Skehan claimed that task characteristics or task demands are more evident in the conceptualization stage. In other words, when the task

requires for development of a complex proposition or dealing with an abstract or large amount of information, this strains the conceptualizer. Task characteristics such as the availability of a pre/post-task activity or the number of participants may have different effects on the stages of speech production. For example, pre-task activity assists both conceptualization and formulation stages as it provides an opportunity to prime lexical and syntactic elements and prepares a pre-verbal message. The number of participants i.e., whether the task is monologic or dialogic is also important as it provides additional time to prepare for conceptual message and linguistic encoding, which helps both the conceptualizer and the formulator. Skehan attempted to link the stages of speech production to complexity, fluency, and accuracy dimensions of performance stating that conceptualization is more involved with lexical and structural complexity whereas accuracy and fluency can be linked to the formulation.

2.4.2 Robinson's cognition hypothesis

Robinson's cognition hypothesis (Robinson, 2001), which is rooted in L1 developmental psychology (e.g., Slobin, 1993), theories of attention (e.g., Wickens, 2002), and SLA research (e.g., Schmidt, 1998), provides an alternative account while attempting to predict the effects of task demands in language learning and production. Adopting the idea of attention that can be voluntarily regulated, the cognition hypothesis makes several predictions concerning output, input, uptake, interaction, automaticity, and individual differences. Regarding output, it is predicted that increasing task demands i.e., cognitive complexity of a task leads to greater complexity and accuracy but mitigates fluency. Learners prioritize complexity and accuracy to meet the high functional demands of the task dictated by its design. As for the other aspects of task

design pertinent to development, Robinson predicted that more effort at this stage leads to the development of relevant L2 linguistic resources to produce more complex conceptual propositions which can be more readily observed during a monologic task performance (Robison, 2011). The underlying idea of the cognition hypothesis is that increased cognitive load requires more complex thinking and increased functional demands entail more complex conceptual representations, for which higher complexity and accuracy ensue. This synchronous increase in complexity and accuracy of output postulated by the cognition hypothesis is rooted in Wicken's (1984) multiple-resource view on attention. In this view, cognitively complex tasks postulate more interaction and negotiation of meaning during task performance. They also induce more attention to language form and meaning to meet the demands of the task which leads to better retention of the given input. Accordingly, Robinson (2001) defines task complexity as "task complexity is the result of attentional, memory, and other information processing demands imposed by the structure of the task on the language learner" (p. 29).

Sequencing tasks from simple to complex, an important tenet of the cognition hypothesis, is assumed to generate greater automaticity. As for individual differences in cognitive abilities such as WM and affective factors such as anxiety, it is hypothesized that their influence on performance and learning will be more evident when the complexity of the tasks increases (Robinson, 2001). TCF is the operational taxonomy of task characteristics proposed by Robinson (2007). This comprehensive framework reviews task characteristics involved in real-world task performance in three dimensions: task complexity, task condition, and task difficulty (Table 2).

Table 2. Triadic Componential Framework (TCF)

Task Complexity (Cognitive Factors)		Task Condition (Interactive Factors)		Task Difficulty (Learner Factors)	
Resource-directing	Resource-dispersing	Participation	Participant	Ability	Affective
+/- Here-and Now	+/- Planning time	+/- Open solution	+/- Same proficiency	H/L Working memory	H/L Openness
+/- Few elements	+/- Prior knowledge	+/- One way flow	+/- Same gender	H/L Reasoning	H/L Control of emotion
+/- Spatial reasoning	+/- Single task	+/-Convergent solution	+/- Familiar	H/L Task-switching	H/L Task Motivation
+/- Causal reasoning	+/- Task structure	+/- Few participants	+/- Shared content knowledge	H/L Aptitude	H/L Anxiety
+/- Intentional reasoning	+/- Few steps	+/- Few contributions needed	+/- Equal status and role	H/L Field independence	H/L Willingness to communicate
+/- Perspective-taking	+/- Independency of steps	+/- Negotiation not needed	+/- Shared cultural knowledge	H/L Mind-reading	H/L Self-efficacy

Source: (Robinson 2011, p. 6)

The task complexity dimension in TCF lists the factors germane to demands of tasks on cognitive resources such as memory, attention, and reasoning. Cognitive demands of tasks are reviewed with a putative distinction between resource-directing (cognitive/conceptual) and resource-dispersing (performative/procedural) variables (Robinson, 2001, 2007; 2011). Manipulation in the cognitive/conceptual dimension of task complexity entails either lower or higher effort at the conceptualization level which makes learners direct and use their attentional resources in line with the purposes of task achievement (Robinson, 2011). In terms of language use, it is hypothesized that the need to convey conceptually complex meaning requires the use of more advanced L2 structures and forms. For instance, if a task requires reference to events that happened in the past, learners' attention and memory resources are directed to the use of relevant

morphology (tense and aspect). Increased task complexity along the resource-directing variables is projected to prompt analysis and development of the theorized 'interlanguage' (ibid.). Tasks with high performative and procedural demands i.e., increased demands along the resource-dispersing dimension, on the other hand, might lead to degraded L2 performance. In other words, complexity, accuracy, and fluency might be negatively affected; however, tasks with such demands create processing conditions of real-time language use. As learners complete these tasks, they practice real-time access to the L2 knowledge which is assumed to result in faster and more automatic access to L2.

On the resource-dispersing dimension of the TCF, the cognition hypothesis makes similar predictions to Skehan's (1998) trade-off hypothesis i.e., increased task demands along with this dimension may lead to lower complexity, accuracy, and fluency. However, the underlying reason for degraded performance is posited differently. Skehan (1998, 2009) suggests that performance decay is related to limited cognitive resources whereas Robinson (2003) refers to loss of control over attention as the reason for degraded performance. As the learners' cognitive resources are dispersed to linguistic and nonlinguistic features of the task at hand, learners lose the control over their attention and they find it more difficult to keep their attention on the relevant linguistic demands of the task, which eventually causes lower complexity, accuracy, and fluency in L2 production.

The task condition dimension of TCF refers to the participant and participation-related variables. Task condition is supposed to be constant while sequencing tasks from simple to complex to ensure the participation and participant conditions are the same and the only variable is the complexity of the task itself (Robinson, 2011).

The task difficulty dimension of TCF concerns learners' individual differences, which are claimed to affect the task performance, and eventually the learning process of L2. Ability variables, which are relatively stable, are listed as WM, reasoning, and aptitude. Task-relevant resource variables, which are less stable, are listed as openness to experience, control of emotion, task motivation, anxiety, willingness to communicate, and self-efficacy. Task difficulty variables may denote variation in L2 task performance among learners (Robinson, 2011). These variables may be a mediating factor between learners' perception of task complexity and the cognitive demands of the task. In other words, learners with higher motivation or higher aptitude may perceive the task as less complex compared to learners with lower motivation or aptitude (ibid.). This may also explain variation in inter-learner performance i.e., learners may have different levels of motivation or anxiety during different tasks. Robinson (2011) postulates a relationship between affective and ability factors i.e., affective variables such as motivation and anxiety may increase or lessen learners' ability resources such as aptitude or WM and eventually affect L2 task performance. Therefore, learners' perception of task difficulty, affective variables, and ability variables are often measured in relevant research mostly through Likert-scale questionnaires to observe interrelationships and also to compare learners' perceptions of task complexity against conceptualized task complexity.

2.5 Integrated skills assessment

L2 testing and assessment had favored discrete-point and indirect testing before the communicative language teaching approach became more commonly implemented and led to the development of skills-based language testing (Weir, 1990). Integrated tasks have been commonly defined as tasks in which test-takers are provided with an input

(written or spoken) and asked to generate their responses (written or spoken) based on this given input (Lewkowicz, 1997). Their counterparts are referred to as independent tasks which involve a single skill being tested in isolation such as independent speaking and writing tests.

Skills-based language testing has brought about other concerns especially because the skills involved have been separately assessed. The authenticity of tests has become a concern in skills assessment as real-life use of language is more interactive and complex (e.g., Frost et al., 2011). As the skills are tested in isolation through independent tasks, test-takers are expected to draw on their personal experience and background knowledge to produce a response. Such tests are therefore criticized for unduly assuming that test-takers' background knowledge would suffice to complete the task (Read, 1990). These tasks do not also reflect the true nature of target constructs in real-life situations (Brown, Iwashita & McNamara, 2005). For instance, assessing listening in isolation might bear validity issues because listening and speaking skills are typically employed together in most oral communication (McNamara, 2000). This is especially true in tertiary education as communication in academic contexts entails comprehension and production with proper employment of higher-order cognitive skills such as summarizing and synthesizing information from source texts (Douglas, 1997). All things considered, integrated-skills assessment has become more widely implemented and emerged as a field of research inquiry in L2 testing and assessment in the past decades (Yu, 2013).

Integrated skills assessment has been claimed and demonstrated to have certain benefits as a method of L2 assessment. First and foremost, they might mediate the effect of background knowledge required for task completion and accordingly minimize the

unfairness pertinent to the issue (Read, 1990; Weigle, 2004). They provide a more realistic context for L2 use thus ensuring a higher level of authenticity (e.g., Plakans, 2015). Especially real-life academic context is well represented in these tasks as students are engaged in writing or speaking based on the given input i.e., books and lectures in their subsequent academic pursuits (Brown et al., 2005). Accordingly, these tasks have a greater predictive capability compared to independent tasks (ibid.). Other listed benefits were positive washback in the classroom and boosting learner motivation (Wesche, 1987). They also garnered positive feedback and preference from test-takers as given input facilitates production and offers an opportunity to learn about the subject matter (Huang, Hung & Plakans, 2018). Finally, such testing tasks are in better alignment with the task-based language teaching and other in-demand L2 teaching approaches that focus on the holistic use of L2 (Plakans, 2013).

On the other hand, Brown et al. (2005) argued the role of input processing and explained that test-takers need to process input material (written or spoken) and integrate the given information into task performance, which makes this process a more cognitively demanding one compared to the processes involved in completion of independent tasks. The complexity of these tasks stems from the fact that test-takers need to employ some cognitive skills such as identifying, selecting, organizing information, and integrating them into language performance, which extends beyond language proficiency. This was echoed by another definition which underlined that test-takers are required to refer to the source materials in integrated tasks whereas in independent tasks they draw on their ideas and knowledge, which is far less complex in nature (Cumming et al., 2006).

2.5.1 Previous research on integrated skills assessment

Previous research on integrated test tasks has compared independent and integrated tasks, examined the underlying construct, explored the effects of task variation, and investigated the relationship between integrated test task performance and language proficiency (Plakans & Gebril, 2012). Most previous research has focused on writing assessment and reading assessment through integrated read-to-write (e.g., Cumming et al., 2006; Plakans, 2009) and read-to-summarize test tasks (e.g., Yu, 2007). A few studies have focused on the use of integrated test tasks to assess speaking ability (Brown et al., 2005; Iwashita, Brown, McNamara & O'Hagan, 2008). The reason for comparatively little interest in integrated listen-to-speak or read-to-speak tasks might be that integrated tasks have been introduced to speaking assessment more recently (Frost et al., 2011; Yu, 2013) and that speaking assessment is more complex in terms of performance scoring compared to other skills (Luoma, 2004).

The research that aimed to understand and describe the construct underlying integrated tasks compared test performance scores from independent tasks against the scores obtained in integrated tasks in an attempt to observe whether the same construct was measured. The findings are rather inconclusive, which is generally attributed to several factors such as the data collection and analysis methods employed, rater behaviors, test-taker characteristics, and behaviors. Most of this research was devoted to read-to-write tasks or listen-to-write tasks (Lee, 2006; Plakans, 2008). They compared integrated task scores against independent task scores and reported that the cognitive processes involved were different. These findings were echoed by Sawaki, Stricker, and Oranjie (2009) who compared task performance scores from independent speaking and integrated speaking and showed they were also different.

Integrated skills assessment research has also compared performance levels and characteristics in different task types. Most research findings indicated that dimensions of performance are different. For instance, Brown et al., (2005) compared four aspects (linguistic resources, phonology, fluency, and content) of both independent and integrated speaking performance and reported different findings. The linguistic characteristics i.e., vocabulary and grammar, as well as content i.e., the ideas were more complex in integrated speaking test tasks compared to independent speaking tasks. This was attributed to the role of input provided to test-takers in integrated tasks. As for fluency and pronunciation, the test-takers were less fluent and experienced more difficulty in pronouncing keywords, which was attributed to the lexical complexity of the input.

Another line of integrated skills assessment research has focused on task design and rating scale development (e.g., Barkaoui et al., 2013; Frost et al., 2011). For instance, Frost et al. (2011) investigated an integrated speaking task developed by Oxford University Press as a part of their language test and suggested that it is "appropriate to consider the accuracy of the content as part of the construct of speaking ability" (p. 366). They also reported that test-takers rely heavily on reproducing idea units in the source text rather than paraphrasing and summarizing them and concluded that rating scale development should consider the appropriate summary of the input i.e., adding the 'input well summarized' to the rating scale to distinguish between different levels of speaking proficiency. In their attempts to operationalize the relationship between input content and reproduced content, Crossley, Clevinger, and Kim (2014) investigated how words from input text are incorporated into speaking performance in the TOEFL iBT listen-to-speak task. They reported a positive correlation between the

integration of words and clauses from input text into oral performance and test-takers' performance scores judged by human raters.

Further research attempted to explore strategy use in integrated tasks. Integrated speaking tasks were shown to elicit a wider range of strategic behaviors compared to independent task performance in TOEFL iBT (e.g., Barkaoui et al., 2013; Swain, Barkaoui, Huang, Brooks & Lapkin, 2009). Rukhthong and Brunfaut (2020) argued that while completing listen-to-summarize tasks, listeners employ different strategies at various levels of task completion. Lower-level processes i.e., acoustic-phonetic decoding, word recognition, and parsing were essential in text comprehension, which is an essential part of integrated tasks; however, higher-level processes i.e., semantic processing and pragmatic processing were activated in listen-to-summarize tasks.

As for the research devoted to test-taker characteristics, which is relatively sparse, Huang and Hung's (2010) research revealed that read-to-speak tasks caused as much anxiety as independent speaking tasks even though they were assumed to reduce anxiety through the provision of source materials. This was attributed to the increasing cognitive demands of integrated speaking tasks. Huang, Hung, and Hong (2016) investigated the role of test-taker characteristics such as anxiety, topical knowledge, and L2 proficiency. Their findings demonstrated that L2 proficiency is the key indicator of task performance in integrated tasks. Topical knowledge was observed to be a significant indicator of task performance as test-takers with prior knowledge of the topic performed better compared to the ones who did not possess relevant knowledge. Considering that integrated skills assessment claims to level the ground for the test-takers with different degrees of topical knowledge, it appears that test-takers with topical knowledge are still more advantageous compared to their counterparts. Anxiety was

observed to affect the performance slightly negatively. Even though the provision of input materials in integrated tasks was assumed to mitigate the potential effects of anxiety (Huang & Hung, 2010), there seem to be other factors provoking anxiety such as cognitive load and difficulty level of input texts (Brown et al., 2005).

2.6 The CAF framework

Aspects of spoken production to be discussed in this part are complexity (syntactic and lexical), accuracy, and fluency, which is often called the CAF framework.

2.6.1 Complexity

Complexity refers to "the extent to which learners produce elaborated language" (Ellis & Barkhuizen, 2005, p.139). There are two types of complexity explored in the previous studies i.e., syntactic complexity and lexical complexity. Variables of syntactic complexity may vary based on the unit of analysis and the amount of subordination. In other words, some measures are based on the unit of analysis such as T-units (terminable units), C-units (communication units), and AS-units (analysis of speech units). Other measures focus on the amount of subordination in the chosen unit of analysis. Subordination refers to the number of dependent, independent, and subordinate clauses. C-units or AS-units are often used as they also include sub-clausal units in the analysis. AS-unit is defined as "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster, Tonkyn & Wigglesworth, 2000). AS-units are claimed to be more reliable than C-units as pausing, and intonation are taken into consideration while it was developed. Foster et al. (2000) explained that pausing often occurs at syntactic unit boundaries

which are evidently units of planning. Pausing and intonation patterns show that speakers plan for multi-clause units; therefore, units of analysis should go beyond a single clause as AS-units do so. In this study, variables of syntactic complexity were selected as words per AS-units referring to the length of the unit of analysis.

Lexical complexity, on the other hand, is often studied with regard to lexical variety (or density) and lexical sophistication. Lexical sophistication is not an aspect of lexical complexity which is investigated in task-based research. The valid and reliable measure for lexical sophistication appears to be the Lexical Frequency Profile (LFP), which basically counts the numbers of the words in a text and compares them against frequency-based word lists. On the other hand, lexical variety has been widely investigated and many variables have been proposed such as the ratio of lexical words (Robinson, 1995), type-token ratio (Robinson, 2007), and mean segmental type-token ratio (Yuan & Ellis, 2003). However, the vocd-D value appears to be a more reliable measure as it employs all the words used in its analysis (Kormos & Denes, 2004). Accordingly, the vocd-D value was used to measure lexical complexity and calculated by Text Inspector in this study. Lexical sophistication was not measured as the previous studies have shown lexical sophistication to be largely affected and restricted by the source texts in integrated tasks (e.g., Kyle & Crossley, 2016).

2.6.2 Accuracy

Housen and Kuiken (2009) defined accuracy as "the extent to which an L2 performance deviates from a norm" (p.4). Accuracy variables that were employed in the previous studies include the percentage of error-free clauses (Yuan & Ellis, 2003), the percentage of error-free C-units (Robinson, 2001), and the number of errors per 100 words

(Mehnert, 1998). There is no research attempting to validate these variables; therefore, the accuracy variable in this study was selected considering its practicality and the nature of integrated speaking tasks. The number of errors per 100 words seems to be an appropriate choice as it does not require the identification of clause-based units. Also, Inoue (2016) showed that it is a good predictor of perceived accuracy as well. Inoue and Lam (2021) employed it as the accuracy variable in their research investigating the effects of extended planning time on the listen-to-speak tasks in the TOEFL iBT test, which is also the type of integrated task investigated in this study. As for the identification of errors Skehan and Foster's (1997) baseline was adopted which views language use that is "nonexistent in English or indisputably inappropriate" as erroneous (p.195). The reason for opting for such a broad baseline is that identifying errors through reference lists of errors is not practical in spoken language which involves incomplete sentences, repetitions, and ellipsis.

2.6.3 Fluency

Fluency is generally associated with 'speed' or 'smoothness' in its definitions. Fluency is also related to whether the utterance is pragmatically acceptable and how listeners perceive it (Lennon, 2000). One widely accepted definition of fluency defines it as "the rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language under the temporal constraints of online processing" (Lennon, 2000, p. 26). There are many fluency variables employed by researchers in the current studies. For instance, Yuan and Ellis (2003) employed 'speech rate' as the fluency variable. Speech rate refers to the number of meaningful syllables per minute. Another fluency variable is the mean length of runs i.e., the number of words per pausally

defined unit (Robinson, 1995; Yuan & Ellis, 2003). Previous research has shown that the speech rate and mean length of runs can best predict perceived fluency (e.g., Towell, Hawkins & Bazergui, 1996). More recent validation studies have also agreed that speech rate and mean length of runs are valid predictors of perceived fluency i.e., fluency ratings (e.g., Kormos & Denes, 2004). In this study, speech rate is chosen as the fluency variable due to its practicality and it was automatically calculated by the online text analysis tool Text Inspector. This web-based tool analyzes the given text and provides detailed information on readability, complexity, lexical diversity, estimated CEFR level, and other key statistics.

2.7 Summary and the goal of this study

The cognition hypothesis has different predictions for complex and simple task performances depending on the resource-directing versus resource-dispersing dimensions of tasks. In the resource-directing dimension, tasks are designed in a way to make conceptual/linguistic demands to direct one's attention to linguistic code and to meet the demands of the complex tasks (Robinson, 2011). In such tasks, greater accuracy, fluency, and complexity have been reported especially in comparison with their simpler counterparts (e.g., Gilabert, 2007; Ishikawa, 2008). However, when manipulation was made along the resource-dispersing dimension of the task i.e., when the task does not direct attention to relevant linguistic aspects that aid task completion but instead divides the cognitive resources with non-linguistic demands, the accuracy, fluency, and complexity of production may decrease.

The cognition hypothesis also acknowledges that individual differences or learner factors such as ability (e.g., aptitude and WMC) and affective (e.g., motivation,

anxiety, and perceived task difficulty) interact with task factors and mediate the above-predicted effects (Robinson, 2011). Robinson and Gilabert (2007) underlined that resource-dispersing variables (e.g., planning time and background knowledge) make "demands on participants' attentional and memory resources but do not direct them to any aspect of the linguistic system which can be of communicative value in performing a task" (p. 165). Studies exploring perceived task difficulty in TCF focused on determining the cognitive demand of the task (e.g., Ellis, 2003) i.e., learners' perception of difficulty and the cognitive demand of the task would be positively correlated, sequencing task for pedagogic purposes to help task designers and reflect learners' perspectives (e.g., Foster & Tavakoli, 2009). Perceived task difficulty in association with other emotional responses such as stress and anxiety, was reported to have detrimental effects on task performance (e.g., Robinson & Gilabert, 2007). Stress and anxiety were shown to increase demands on WM resources as they divide one's attention and deplete limited resources to cope with stress and anxiety (e.g., Ashcraft & Kirk, 2001). However, these studies have not made any references to specific dimensions of language production such as complexity, accuracy and fluency while explaining the effect of perceived task difficulty. The questions guiding this study are;

i. Does task complexity [+/− outline] affect task performance in listen-to-speak tasks?

ii. How much of the variation in task performance can be explained by task complexity, WMC, and perceived task difficulty?

In line with the cognition hypothesis, it is predicted that content, complexity, accuracy, and fluency in participants' spoken performance may degrade when they are not given an outline of the spoken input (complex task) compared to the performance in the simple task in which they are provided with an outline (simple task) (Hypothesis 1).

The ability factor in this study is WMC and the studies investigating it demonstrated that individuals with high WM benefitted while performing demanding tasks (e.g., Kim et al., 2015) but relatively less in the performance of less demanding tasks (e.g., Kormos and Trebits, 2011). A mediating effect of WMC (Hypothesis 2) is predicted i.e., participants' WMC will be able to explain the variation in task performance to a certain extent. As for the dimensions of task performance such as content, complexity, accuracy and fluency, the extant research reported mixed findings. Gilabert and Muñoz (2010) reported a correlation between WMC and complexity (lexical) and fluency in the video narrative task performance. On the other hand, Kormos and Trebits (2011) found that WMC correlated with syntactic complexity in simple task conditions (re-telling a story) rather than complex task conditions (inventing a story). Cho (2018) observed no effect of WMC on either oral task performance in which the 'number of elements' was manipulated in simple and complex tasks. However, it was reported that high WMC gives an advantage in speech production in terms of accuracy and complexity when there is planning involved i.e., simple task conditions (e.g., Ahmadian, 2012; Guará-Tavares, 2011). Considering the inconclusive nature of the results, this study does not make any specific predictions regarding dimensions of task performance. In line with the previous research findings, participants will perceive complex tasks as more difficult than simple tasks (Hypothesis 3). Finally, participants' perception of task complexity will be able to account for the variation in task performance to some extent (Hypothesis 4); however, due to the lack of empirical evidence, this study does not make predictions about which dimensions of task performance would be explained.

CHAPTER 3

METHODOLOGY

3.1 Participants and the context of research

A total of 40 participants between the ages of 18 and 21 took part in this study. They were all English preparatory program students at a foundation university in İstanbul, Turkey. Before the start of data collection, participants had taken The Michigan English Placement Test (MTELP) (Corrigan et al., 1978) which indicated their level as CEFR B2. They were placed in the course to receive English for Academic Purposes (EAP) instruction. At the time of this study, they had received two months of instruction (20 per week and 160 hours of instruction in total) including note-taking listening, an oral summary of written or spoken input, group discussion, speaking (individual long run), essay writing and other relevant academic skills. A careful screening was conducted to ensure participants' language proficiency. Participants who had spent more than six months in an English-speaking country, bilinguals and misplaced (placed in the wrong level) students, and students who had previous TOEFL were excluded. Participants were not offered any material compensation for participating in this study. They were informed that upon completion of this research they would be able to learn their performance scores in speaking and WM tasks. Although 40 participants took part in the study, statistical analyses were conducted on data from 35 participants due to outlier analyses. Also, there was a technical issue with one participant's audio recording which impeded understanding; thus, excluded from the analysis.

3.2 Materials

3.2.1 Pedagogic tasks

The integrated tasks employed in this study were the TOEFL listen-to-summarize tasks in which test-takers listen to an excerpt from a lecture and summarize it. The tasks were chosen from the Official Guide for the TOEFL iBT test available on the ETS website (https://www.ets.org/toefl). The task complexity was operationalized by providing a note-taking paper with a skeleton outline that provides headings corresponding to the excerpt's main points in the simple version and no outline i.e., a blank sheet of paper in the complex version. Based on the TCF proposed by Robinson (2011), providing an outline is assumed to lower the cognitive burden on the listener and aid the note-taking process. It should be noted that outline manipulation has not been explored in task complexity literature before; thus, not listed in the TCF. However, it can be assumed that this manipulation fits in the first column of cognitive factors as it is inherently related to task design not participant or learner factors. Considering the distinction between resource-directing and resource-dispersing factors, outline manipulation is assumed be on the resource- dispersing dimension of the cognitive factors. This is because it does not direct cognitive resources to a specific linguistic form or meaning rather it makes the task performatively complex by dispersing the cognitive resources to listening and note-taking. Topics and outlines used in this study were checked by two English language teachers who also work as speaking examiners for Cambridge exams conducted in Turkey. Examiners conferred on the discrepancies and necessary changes were made by the researcher (Appendix A).

3.2.2 Perception questionnaire

Self-rated questionnaires have been used to measure learners' perception of task difficulty (e.g., Révész, 2014; Révész, Michel, & Gilabert, 2015). The questionnaire used in this study was taken and adapted from Robinson (2001, p.41). Robinson's questionnaire consists of 5 questions addressing task difficulty, anxiety, self-rating of performance, interests, and motivation. It is a short questionnaire, with a 10-point Likert scale (ranging from 0 to 9), which can be given after each task with minimum disruption of the task performance that follows (ibid.). Robinson's questionnaire was adapted to address the needs of this study. Question number four, which is directed at students' interests, was changed into topic familiarity as topical knowledge was shown to affect task performance even in integrated skills tasks (e.g., Huang et al., 2016). Considering that participants are highly proficient in English (B2), the researcher felt no need to translate the questionnaire into Turkish (Appendix B).

3.2.3 WM tasks

The WM tasks employed in this study are operation span (OSpan) (Unsworth et al., 2005) and running span (Run Span) tasks, which are both complex span tasks. OSpan task measures the executive control component of WM which plays a comparatively bigger role in L2 use (Linck et al., 2014). Other complex span tasks such as reading span tasks or listening span tasks were not chosen as they could confound with L1 verbal processing (Conway et al., 2005). Even though the findings are supporting that WM is a domain-general resource as the WM tasks administered in L1 and L2 showed a strong correlation (e.g., Osaka & Osaka, 1992), the L2 proficiency of the participants should be considered before deciding on the language of the WM task. In order to eliminate the

effect of L2 in WM measurement, the Turkish translation of the OSpan task by Çak (2011) was used in this study. The task was taken from the website of the Attention and Working Memory Laboratory of Georgia Institute of Technology (https://englelab.gatech.edu/). It was given and recorded on E-Prime 3 software. In this task, participants are asked to keep track of each letter appearing on the screen while making mathematical calculations. After each letter, a simple mathematical operation [e.g., (2+3) – 1=4] appears on the screen with True or False options. In order to ensure the validity of the results 85% accuracy rate should be maintained throughout the task. This sequence of the letter and mathematical operations ranges from 3 to 7. Once the sequence is over, participants are asked to recall the letters in the order of presentation and tick them in the given matrix. The sum of partially recalled sets was calculated as the OSpan score by the E-Prime.

Along with the OSpan task, a running memory span task, which measures updating, was used. Updating is identified as a significant executive function and it is assumed to correlate strongly with fluid intelligence (Friedman et. al, 2006). Broadway and Engle (2010) demonstrated that running memory span also serves as "a consistent measurement of WMC across widely differing conditions" and suggests strong correlations with other measures of WM and fluid intelligence. (Cowan et al., 2005; Friedman et al., 2006). The Turkish translation of the Run Span task (Cinan, 2001), available on the website of Attention and Working Memory Laboratory of Georgia Institute of Technology, was given. In this task, participants are asked to recall a set of letters in the order of presentation. They are not informed about the number of letters to be presented or the number of letters to be recalled. For instance, participants are given a set consisting of four letters (A-B-C-D) and asked to remember the last three letters.

Each letter appears on the screen for 500 milliseconds and participants mark the letters on the given matrix in the order of presentation. The sum of partially recalled sets was calculated as the Run Span score.

3.3 Design and procedures

In October, 2021 an ethics committee approval was obtained at Boğaziçi University (Appendix C). In November, 2021 a pilot study was conducted to test the materials and fine-tune the procedures. Both practice and actual tasks were randomly used in the piloting stage. Given the sample size of the study, only eight participants took part in the piloting. These participants did not take part in the actual experiment. The tasks were not counterbalanced to control for the topic effect due to the small sample size. Piloting was particularly helpful in clarifying the instructions to be given. It was also helpful to ensure the physical conditions were agreeable to the participants.

Data collection, which started in December 2021, was conducted in two sessions i.e., WM tasks and listen-to-speak tasks were given on different days. The whole process was carried out on a one-on-one basis and the researcher followed the same steps for each participant and task. The first session to be completed was for WM tasks. Each participant was invited to a computer laboratory. First, the researcher went through the background questionnaire (Appendix D), then she gave the instructions in Turkish and clarified vague points if any. The presentation of tasks was counterbalanced. In other words, half of the participants were given OSpan first, and the other half were given Run Span first. Each WM task starts with practice sets, which are embedded in the task and cannot be skipped by the task-taker. There were five practice sets in OSpan and three practice sets in Run Span completed to ensure that the procedures were understood

before moving to the actual test. The session took about 30-40 minutes depending on the participants' pace.

In the second session, each participant was asked to perform four listen-to-speak tasks. They completed two samples (+/- outline) to get familiar with the task format. Practice tasks were followed by actual test tasks. In line with the timing details given in the TOEFL iBT exam, after listening to the excerpt, participants were given 20 seconds to prepare and 60 seconds to speak i.e., respond to the given prompt which basically requires the participants to summarize the main points and the supporting details. A counter-balanced design was used and task types were rotated to ensure an equal number of participants completed each task and its versions to ensure the topic effect was controlled. Participants were asked to complete the perceived task difficulty questionnaire immediately after each task they completed. As explicitly stated in the consent form, participants were audio-recorded for analysis purposes. This session took about 30 minutes.

3.4 .Scoring

Scoring system adopted for pedagogic tasks, perception questionnaire and WM tasks will be discussed in detail in the following section.

3.4.1 Pedagogic tasks

Both task complexity and WM studies tend to use "more precise operationalizations of underlying constructs" (Skehan, 2001, p. 170) and measure complexity, accuracy and fluency as separate units of performance instead of employing global scales to rate the overall performance. Measuring complexity, accuracy, and fluency as dependent

variables also enables comparability of findings across various task demands (Robinson, 2001). For the purposes of this study, the dependent variables measured separately are content, complexity, accuracy and fluency. They also match the descriptors in the TOEFL iBT integrated speaking rubrics to some extent. First the spoken data was transcribed (Appendix E) by the researcher and semantically non-meaningful utterances such as sound fillers (e.g., uh, ah) as well as false starts and repetitions were removed. Transcribed data was segmented into AS- units manually by the researcher. Performance in content was assessed through a content rating scale (out of 5) taken and adapted from Rukthong (2015) (Appendix F). Complexity was divided into syntactic complexity and lexical complexity as suggested in the previous studies (e.g., Frost et al., 2011). Common syntactic complexity variables in the literature were words per AS-unit (e.g., Tavakoli & Foster, 2008), subordinate clauses per AS-unit (e.g., Crookes, 1989; Mehnert,1998), and words per clause (Norris & Ortega, 2009). Considering the time limitations, we chose the mean number of words per AS- unit as the measure because it was the most widely used one. An AS-unit (Analysis of Speech unit) is "a single speaker's utterance consisting of an independent clause or sub clausal unit, together with any subordinate clause(s) associated with it" (Foster et al. 2000, p. 365). Lexical complexity was measured through the vocd-D value as it was the most reliable measurement of lexical variety (e.g., Inoue, 2021). Text Inspector (https://textinspector.com), an online text analysis tool, was used to calculate the vocd-D value. For accuracy, errors per 100 words was chosen as it does not require identification of clause-based units, which can be rather problematic (Inoue, 2021; Mehnert, 1998). Finally, fluency was measured by speech rate (syllables per minute) as it was reported to correlate with perceived fluency (Kormos & Dénes, 2004). It was also

convenient to use speech rate as it can be calculated quickly and reliably on Text Inspector. An independent rater scored 10% of the data to ensure scoring reliability with an overall interrater reliability at around 90%. Variables used in this study are summarized below in Table 3.

Table 3. Task Performance Measures

Dimension	Measure
Content	Content rating scale (out of 5)
Syntactic Complexity	Mean number of words per AS-unit
Lexical complexity	Vocd-D
Accuracy	Mean number of errors per 100 words
Fluency	Syllables per min

3.4.2 WM tasks

Two types of storage scores i.e., absolute score and partial score or partial-credit scores are available for the researchers using complex span tasks. In this study, partial scores, in both tasks, were used as they are reported to have higher internal consistencies (Conway et al., 2005; Friedman & Miyake, 2005). Another reason for using partial scores is that in absolute scoring partially recalled sets are disregarded i.e., absolute scoring does not use all available information, which hinders the detection of individual differences. In sum, partial scores (as provided by E-Prime) from both tasks were used in this study.

3.4.3 Perception questionnaire

After completing each task, participants responded to 5 questions along a 10-point Likert scale. While coding the data, a score of 0 was given to semantically most negative response (e.g., not difficult) and a score of 9 was given to semantically most positive

response (e.g., very difficult). Some items in the questionnaire are reverse coded (e.g., anxiety).

CHAPTER 4

RESULTS

The first research question aimed to explore whether task complexity affected task performance. Table 4 presents the descriptive statistics for the participants' performances in both simple and complex conditions. While scores for syntactic complexity, lexical complexity, fluency, and content were higher in simple task condition compared to those in complex task condition, accuracy scores (i.e., the number of errors per AS-unit) were lower.

Table 4. Descriptive Statistics for Task Performance Scores (N = 70)

Measure	Task Condition	Min.	Max.	Mean	SD	Skewness	Kurtosis
Syntactic Complexity	Simple	10.45	16.20	12.74	1.56	.63	-.47
	Complex	9.96	15	11.86	1.37	.68	.14
Lexical Complexity	Simple	41.38	88.26	61.24	12.05	.34	-.67
	Complex	33.95	67.50	49.92	7.27	.36	.31
Accuracy	Simple	1	3.5	2.08	.73	.04	-1.01
	Complex	1	4.5	2.67	.98	.07	-1.11
Fluency	Simple	100	185	153	20.74	-.18	-.27
	Complex	103	181	141.61	22.23	-.14	-1.14
Content	Simple	3	4.5	3.78	.45	-.46	-.57
	Complex	2.5	4.5	3.42	.59	-.15	-.94

As there were multiple dependent variables, a one-way within-groups multivariate analysis of variance (MANOVA) was performed to investigate the effect of task complexity on task performance. Before the analysis, relevant assumption testing was conducted. Multivariate normality assumption was met considering the Shapiro Wilks' value ($p > .05$). There were no multivariate outliers as indicated by Mahalanobis

distance scores ($p < .001$). Boxplots showed no univariate outliers. Linearity assumption was met based on the scatterplots. There was no multicollinearity as correlations between variables were within the desired range (.20 - .70). Homogeneity of variance-covariance assumption was violated according to the results of Box's M test, which was significant ($M = 18.56$, $p = .313$); therefore, Pillai's trace was reported as it is most robust to violations of assumptions (Bray & Maxwell, 1985).

MANOVA results showed that there was a significant effect of task complexity on the dependent variables combined, $V = .84$, $F(5, 30) = 37.77$, $p < .001$, $\eta^2 = .84$. Univariate tests indicated significant task complexity effects on syntactic complexity, $F(1, 34) = 16.96$, $p < .001$, $\eta^2 = .33$, lexical complexity, $F(1,34) = 45.62$, $p < .001$, $\eta^2 = .57$, accuracy, $F(1,34) = 39.85$, $p < .001$, $\eta^2 = .54$, fluency, $F(1, 34) = 21.51$, $p < .001$, $\eta^2 = .38$ and content $F(1, 34) = 28.71$, $p < .001$, $\eta^2 = .45$.

The second research question attempted to probe how much variance in task performance was explained by task complexity, WM capacity, and perceived task difficulty. Table 5 presents descriptive statistics for WM capacity, measured with Run Span and OSpan tasks, and perceived task difficulty.

Table 5. Descriptive Statistics for WM Measures and Perceived Task Difficulty (N = 35)

	Min.	Max.	Mean	SD	Skewness	Kurtosis
Run Span	9	33	20.74	6.47	.10	-.87
OSpan	47	75	63.97	6.68	-.58	.36
Simple Task Perceived Difficulty	2.5	7	4.48	1.03	.54	.36
Complex Task Perceived Difficulty	5	8	6.98	.74	-1.09	1.31

Prior to conducting regression analyses, correlation coefficients between the independent variables and the dependent variables were checked (Table 6). Separate

Pearson correlations were conducted between each dependent variable (i.e., syntactic complexity, lexical complexity, accuracy, fluency, and content) and predictor variables (i.e., WM task scores and perceived task difficulty). A point-biserial correlation was used for task complexity as it is a dichotomous variable (simple versus complex). Task complexity moderately correlated with lexical complexity, both WM Run Span and OSpan scores moderately correlated with fluency, and perceived task difficulty moderately correlated with lexical complexity.

Table 6. Correlation Matrix

Correlation	Syntactic complexity	Lexical complexity	Accuracy	Fluency	Content
Task Complexity	-.29*	-.49**	.32**	-.26	-.32**
WM Running Span	.10	.04	-.15	-.35**	.10
WM Operation Span	-.03	.07	02	-.32**	.15
Perceived Task Difficulty	-.29*	-.52**	.31**	-.17	-.32**

**. Correlation is significant at the .001 level (2-tailed)
*. Correlation is significant at the .05 level (2-tailed).

In order to check the predictive value of task complexity, WMC, and perceived task difficulty, a stepwise multiple regression was conducted for each dependent variable. Before conducting the analyses, the relevant assumptions were tested. Normality, linearity, and homoscedasticity were checked through residuals and scatter plots, which indicated no significant violations (Field, 2019). The collinearity statistics, tolerance, and VIF were checked to ensure no collinearity and multicollinearity (ibid.). Also, the independent variables were not highly correlated with each other i.e., correlations were within the acceptable range (r = .20 -.70.). Durbin-Watson statistic was within the desired range (1.5 - 2.5) meeting the independence of errors assumption (ibid.). In line with the assumptions of regression, independent variables which were not

significantly correlated with the given dependent variable were not entered into the model.

The first stepwise multiple regression was conducted on syntactic complexity with task complexity and perceived task difficulty scores as the predictors. At step 1 of the analysis, task complexity was entered into the regression equation, F (1,68) = 6.21, p < .001 and explained approximately 0.7 % of the variance in syntactic complexity (Adj. R^2 = .07). Perceived task difficulty did not explain a significant amount of variance in syntactic complexity, t = -.78, p > .05.

The second multiple regression was conducted on lexical complexity with task complexity and perceived task difficulty (both moderately correlated with lexical complexity) as predictors. Perceived task difficulty that entered into the regression equation at Step 1, F (1,68) = 25.26, p < .001 explained 26% of the variance in lexical complexity (Adj. R^2 = .26). Task complexity was not a significant predictor (t = -.1.26, p > .05).

The third regression analysis was conducted on accuracy with task complexity and perceived task difficulty (both weakly correlated with accuracy) as predictors. In Step 1 of the analysis, task complexity was entered into the regression equation and was significantly related to the accuracy of task performance, F (1,68) = 7.9, p < .001. Adjusted R^2 was .09, indicating approximately 10% of the variance of the accuracy of the task performance. Perceived task difficulty did not enter the equation (t =-.71, p > .05).

The fourth regression analysis was conducted on fluency with the Run Span and OSpan scores (both weakly correlated with fluency) as predictors. The Run Span score was entered into the regression equation at Step 1, F (1,68) = 9.76, p <.001 and

explained approximately 11 % of the variance in the fluency of the task performance (Adj. $R^2 = .11$). The OSpan score did not enter the model at Step 2 of the analysis ($t = -1.07$, $p > .05$).

The final regression analysis was conducted on content with task complexity and perceived task difficulty (both weakly correlated with content) as predictors. Task complexity was entered into the regression equation and was a significant predictor, $F(1,68) = 7.9$, $p < .001$, explaining approximately 10 % of the variance in the content of the task performance (Adj. $R^2 = .09$). Perceived task difficulty did not enter the equation at Step 2 of the analysis ($t = -.79$, $p > .05$).

Before calculating the overall perceived task difficulty score for each participant, responses to Question 4 were separately analyzed to check whether participants were equally familiar with the topics given in the tasks. As illustrated in Table 7 below, participants appeared to be at least moderately familiar with the topics, which should be attributed to the fact that topics covered during instruction were considered while choosing and adapting research tasks.

Table 7. Descriptive Statistics for Topic Familiarity

TASK	Task Condition	Min.	Max.	Mean	SD	Skewness	Kurtosis
TASK A Marketing	Simple	6	7	6.94	.23	-4.24	1.8
	Complex	6	8	6.88	.47	-.45	2.15
TASK B Advertising	Simple	5	7	6.82	.52	-3.13	9.79
	Complex	5	7	6.64	.60	-1.59	1.89
TASK C Education	Simple	6	7	6.82	.39	-1.86	1.66
	Complex	5	7	6.58	.61	-1.27	.87
TASK D Psychology	Simple	5	7	6.44	.78	-1.03	-.44
	Complex	6	7	6.72	.46	-1.08	-.94

In order to ensure the operationalization in this study, Mann–Whitney–Wilcoxon test was conducted to compare whether participants' responses for each task changes depending on the condition. The results indicated that participants' responses were significantly different across conditions for TASK A (Marketing) $U = 240$, $p < .05$, for TASK B (Advertising) $U = 213.5$, $p < .05$, for TASK C (Art Education) $U = 210$, $p < .05$ and for TASK D (Psychology) $U = 153$, $p < .05$.

In sum, this study investigated the effect of task complexity, perceived task difficulty, and WMC in listen-to-speak task performance in L2 English. The results revealed a significant relationship between task complexity and task performance. Performance scores (syntactic complexity, lexical complexity, fluency and content) were higher in simple condition compared to the ones in complex condition except for accuracy scores which were lower in simple condition indicating better performance. Task complexity was also a significant predictor of task performance. It accounted for approximately 1% of the variance in syntactic complexity, 10% of the variance in accuracy and nearly 10% of the variance in content. Perceived task difficulty was a significant predictor only in lexical complexity explaining 26% of the variance in performance. Finally, WMC was a significant predictor of fluency accounting for 11% of the variance.

CHAPTER 5

DISCUSSION AND CONCLUSIONS

The main goal of this study was to further explore the role of task complexity and WMC in listen-to-speak task performance in L2. Adopting the framework offered by the cognition hypothesis, task complexity was manipulated by providing an outline to be employed while listening to enhance note-taking making the task procedurally less complex. In line with the adopted framework, perceived task difficulty was measured through a questionnaire. WMC was measured through OSpan and Run Span tasks as suggested by a domain-general perspective of WM (Engle 2002, Engle & Kane, 2004). The data were collected from a small sample with the proficiency level of CEFR, B2 determined by a standard placement test. Adopting a within groups experimental design, all participants took all the TOEFL iBT listen-to-speak tasks and OSpan and Run Span tasks.

5.1 Discussion

The first research question was concerned with the effect of task complexity on task performance, and it was hypothesized that task complexity would have an effect on syntactic complexity, lexical complexity, accuracy, fluency, and content of task performance. In other words, participants' scores would be higher in the simple task condition in which they are given an outline of the spoken input (Hypothesis 1). This was affirmed by the results which revealed higher mean scores for each variable in the simple task condition and showed an effect (with large effect size) of task complexity on each variable. The findings were in line with the predictions of the cognition hypothesis

(Robinson, 2001) i.e., procedurally or performatively complex tasks lead to degradation in L2 performance as they disperse the resources such as attention and memory over different linguistic and nonlinguistic demands of the task. As the demands of the task increase and cognitive resources disperse, it becomes more difficult for the L2 learners to keep control over their attention and keep up with the linguistic demands of the task. The trade-off hypothesis (Skehan, 1998; 2009) also makes similar predictions referring to the limited cognitive sources. Unlike the cognition hypothesis, which does not suggest prioritization of certain aspects of production over others, the trade-off hypothesis assumes that learners would prioritize fluency, associated with meaning, as learners prioritize meaning over form due to limited attentional resources (Van Patten, 1990). In other words, fluency would be least likely to suffer in complex condition whereas complexity and accuracy would emerge as the areas that compete for limited cognitive resources (Skehan, 1998; 2009). However, the findings in this study revealed that fluency would also suffer in complex condition (M = 141.61, SD = 22.23) compared to simple condition (M = 153, SD = 20.74) indicating that all aspects of production would suffer in complex condition to a certain degree. This may be attributed to the complexity manipulation adopted in this study i.e., providing an outline. Previous research has shown that an outline with main ideas in a skeletal form, enhanced both the quality of notes and the note takers' test performance (e.g., Dunkel et al, 1989; Kiewra et al., 1989; Lin, 2006). It should be noted that test performance in these studies refers to the comprehension performance. As the participants keep and review their notes during the comprehension test, it was suggested to facilitate the encoding and decoding processes strengthening the comprehension and internalization of information (ibid.).

As the outline of the spoken input provided participants with the layout of the lecture and key vocabulary, it can be suggested that it enhanced the note-taking process itself garnering its potential benefits to comprehension and for production which is essentially a reproduction of the input. The scoring system and measurement (number of syllables per minute) adopted in this study may have also contributed to the advantages deployed by the outline. As the outline provides the layout, this may have helped participants avoid false starts and repetitions, which were excluded during transcription. In other words, outline manipulation can explain why fluency suffered in complex condition contrary to predictions the of the trade-off hypothesis. While discussing fluency, Skehan and Foster (1999) stated that speakers might rely on lexicalized systems as they use the language in real-time. This supports the assumptions made above regarding the role of outline, which may have helped speakers note down and repeat already existing lexicalized systems in the input. Frost et al. (2012) reported that test-takers rely heavily on reproducing idea units in the source text rather than paraphrasing and summarizing them. In sum, test-takers performed better at all aspects of performance including fluency when they were provided with an outline.

The second research question was concerned with how much of the variance task complexity, perceived task difficulty and WM capacity explain in task performance. Task complexity accounted for variation in syntactic complexity (1%), accuracy (10 %) and content (10%) whereas other variables were not statistically significant predictors in these dimensions of task performance. As mentioned in the discussion of the first research question, syntactic complexity and accuracy are the aspects of language production which are assumed to deteriorate more in demanding tasks (Skehan, 1998; 2009). Accordingly, in this study the only variable that can explain variance in syntactic

complexity, albeit little, and accuracy is task complexity. Task complexity was also claimed to have a measurable effect on listening comprehension (Robinson & Gilabert, 2007) which was confirmed in this study as task complexity was able to predict content performance (10%).

As for the findings related to WMC and dimensions of task performance, it was hypothesized that there would be a mediating effect of WMC (Hypothesis 2). However, no specific predictions were made regarding certain aspects of performance due to the inconclusive nature of the findings. The discussion should start with the content dimension which is directly related to listening comprehension. Listening comprehension in listen-to-speak tasks are mostly measured through the content dimension of language production (e.g., Gebril & Plakans, 2011; Rukhthong, 2015). Accordingly, content was separately measured through a scale adopted from Rukhthong (2015). In this study, WMC (measured through non-linguistic OSpan and Run Span tasks) is shown to have no measurable effect on content aspect of speech production. Previous studies which measured WMC through domain-specific tasks (listening span tasks) reported positive correlations between listening comprehension test scores and listening span task scores (e.g., Sakuma, 2004; Shanshan & Tongshun, 2007). Other studies reported that the relationship between listening comprehension and WMC is not straightforward and there is no measurable effect of WMC especially when other factors such as L2 proficiency, processing speed, and topical knowledge are considered (e.g., Andringa et al., 2012). Considering the above-mentioned points, it can be claimed that WMC has no measurable effect on listening comprehension involved in listen-to-speak tasks especially when task complexity is manipulated and L2 proficiency and topical knowledge are controlled. L2 proficiency is particularly critical here as "the putative

components of WM are all posited to be interacting bidirectionally with LTM… inhabited by learners' L1 mental lexicon and grammar as well as their L2 knowledge/proficiency" (Wen et al., 2015, p. 52).

As for the dimensions of speech production, WMC was reported to have significant effects on fluency, accuracy, and complexity (Ahmadian, 2012; Daneman, 1991; Daneman and Green, 1986; Fortkamp, 2000; Gilabert and Muñoz, 2010; Guarra-Tavares, 2011; Mizera, 2006; Mota, 2003; Trebits and Kormos, 2008). In Fortkamp's (2000) study, there was no measurable effect of WMC on lexical complexity, which was explained in terms of Levelt's (1989) speech production model. In line with the postulations made by this model, the effect of WMC is more evident in controlled processing such as grammatical encoding, which follows lexical retrieval. Controlled processing is the part where attention is controlled to perform several functions such as activating and maintaining information as well as inhibiting irrelevant information and monitoring for errors (Forthkamp, 2000). Accordingly, there was no measurable effect of WMC on lexical complexity in this study.

The extant research has reported inconclusive findings when it comes to the accuracy and syntactic complexity dimensions of spoken task performance (e.g., Gilabert & Muñoz, 2010; Mizera, 2006; Mota, 2003; Kormis & Trebits, 2011; Mitchell et al., 2015). For instance, Mota's (2003) study reported a negative correlation between accuracy and spoken performance; however, it disregarded the role of L2 proficiency, which made its findings less conclusive. WM appeared to have positive correlation with syntactic complexity and negative correlation with accuracy of task performance only in simple task condition indicating to the mediating role of task complexity while

allocating cognitive resources (Kormos & Trebits, 2011). On the other hand, Mizera (2006) reported no correlation between WM and accuracy and structural complexity of task performance. Guara-Tavares (2009) reported an effect of WM on the accuracy of task performance; however, L2 proficiency was not taken into account in this study either. It can be suggested that high WM learners are more likely to produce more grammatically accurate and more complex structures; however, this may also be mediated by the structural demands of the task or the L2 proficiency of learners. In this study, no measurable effect of WMC on accuracy and syntactic complexity was observed. This may be attributed to the performance measures adopted in this study. Syntactic complexity measure (mean number of words per AS-unit) is not a sensitive measure of subordination, which may have affected the scoring. However, accuracy (errors per 100 words) is a standard measure used in cognitive and psycholinguistic research. Another mediating factor may be the time pressure as listed in the communicative stress factors by Skehan (1998). The listen-to-speak tasks employed in this study required test-takers to summarize the given points in 60 seconds. Even though timing, as in planning and speaking time, was not manipulated in this study, the allotted time for the tasks was short and could easily emerge as a communicative stress factor. Time pressure may affect even advanced L2 learners as they are still slower in encoding processes (Ellis, 2003). Previous studies also showed that WMC affects the acquisition of L2 syntactic and lexical knowledge through its role in regulating attention. Attention is at the center of noticing and encoding new information, which is especially important in learning new rules of grammar and vocabulary in L2 (Ellis, 1996; Juffs, 2006; Miyake & Friedman, 1998; Sawyer & Ranta, 2011). Thus, it has been suggested that WMC might have more influence on how L2 learners process new information, how they

regulate their attention to notice linguistic aspects of the input, which eventually affects the learning outcomes in both L2 lexical and grammatical development (e.g., Atkins & Baddeley, 1998; French & O'Brien, 2008; Masoura & Gathercole, 1999, 2005;). In this vein, WMC might be more related to mechanisms involved in input processing rather than being directly related to spoken task performance.

WM was a significant predictor only in the fluency (11%) dimension of task performance, echoing the findings from previous studies (e.g., Daneman, 1991; Fortkamp, 1999; Gilabert & Muñoz, 2010; Kormos, 2006; Kormos & Safar, 2008; Kormos & Trebits, 2011; Trebits & Kormos, 2008). However, it should be noted that the findings are rather inconclusive and there are methodological concerns such as employing domain-specific WM tasks and not controlling for L2 proficiency. For instance, Fortkamp (1999) reported correlation between reading span test and fluency whereas no correlation was reported between speaking span and fluency. This was attributed to the fact that L2 proficiency was not controlled in their study and speaking span in L2 might have actually been measuring L2 speaking proficiency rather than WMC. Mota (2003) reported a positive correlation between WMC scores and L2 fluency, which is in line with the findings of this study; however, WMC was measured by a domain-specific measure i.e., speaking span task. Mizera's (2006) study also reported a correlation, albeit weak, between L2 fluency and WMC. The weak correlation was attributed to the proficiency level of participants who were advanced learners. Gilabert and Muñoz's (2010) study showed a positive correlation between fluency and WMC; however, once L2 proficiency is considered, the correlations disappeared. Trebit and Kormos's (2008) study revealed a positive correlation between WMC (measured by a backward digit task) and fluency in the complex tasks pointing to the mediating

influence of task complexity. Gilabert and Muñoz's (2010) findings showed a weak but positive correlation between WMC and fluency; however, they discussed that L2 proficiency explains fluency better than WMC does. In this study, participants were CEFR B2 level learners, who may also be considered at the stage of speech production where the processing is automatized, hence it could be suggested that there would be no measurable effect of WMC on fluency. However, the findings indicated that WMC could explain fluency to some extent. This may be attributed to the task complexity operationalizations adopted in this study. It could also be attributed to the WM tasks employed. A positive, but moderate, correlation was observed between fluency and OSpan (r = .32) and Run Span (r = .35). However, in stepwise multiple regression analysis, Run Span was a significant predictor while OSpan was not. This may be because OSpan and Run Span scores were covariates as they were strongly correlated. It can also be argued that Run Span measures updating; thus, provide more insight into the individual differences in higher-order cognitive capacities which includes language (e.g., Engle, 2018).

As for perceived task difficulty, it was assumed that participants would perceive complex tasks as more difficult than simple tasks (Hypothesis 3). It was confirmed by the statistical analyses which showed that participants' perceptions were significantly different across conditions. In other words, when the outline was removed, participants reported increased difficulty. It should be noted that observable task manipulations such as providing an outline may easily influence perceptions despite not affecting the amount of cognitive resources to be employed while doing the task (Sasayama et al., 2015). Perceived task difficulty was also assumed to be able to account for variation in task performance (Hypothesis 4) and it was the only significant predictor of lexical

complexity (26%). Even though tenets of TCF have been widely investigated, learners' perception of task difficulty is underexplored, and more empirical evidence is needed to explore its effects on dimensions of task performance. Available research seems to focus on either ascertaining the cognitive demands of the task (e.g., Ellis, 2003), ensuring the operationalization of task complexity in their research (e.g., Robinson, 2005) or simply reporting learners' perspectives (e.g., Tavakoli, 2009). To the best of our knowledge, there have been no studies exploring the effects of perceived task difficulty on specific aspects of task performance i.e., fluency or accuracy. Nevertheless, there are few studies which reported that perceived task difficulty is associated with emotional responses such as anxiety and stress, hence might have detrimental effects on task performance (e.g., Ashcraft & Kirk, 2001; Robinson & Gilabert, 2007). There is also a methodological issue in measuring perceived task difficulty using self-assessment rating questionnaires. Robinson's (2001) questionnaire has been used in seventy percent studies exploring the cognitive load and task complexity (Sasayama et al., 2015). Other measurement tools such as stimulated recall and interviews with open-ended questions might provide more insight into learners' perceptions of task difficulty (Sasayama, 2016).

Considering the limitations presented by self-sating scales and how easy it would be to influence participants' perceptions by observable task manipulations, the findings in this study can be attributed to the outline manipulation and the concept of task complexity itself. Previous research clearly states that perceived task difficulty ratings are almost always in line with task complexity manipulations (e.g., Robinson, 2011). In other words, it can be claimed that perceived task difficulty, unless measured through different tools, is an extension of the task complexity itself. As discussed above, provision of an outline in simple task condition facilitates the note-taking process and

increase the quality of notes taken (e.g., Dunkel et al, 1989; Kiewra et al., 1989; Lin, 2006). Crosley et al. (2014) reported that test-takers who integrate words and clauses from input text into their performance have better scores in TOEFL iBT. Having an outline while listening may help learners to note down more words and clauses from the input text. While speaking, they may refer to the notes and reproduce the same idea units without paraphrasing or summarizing (Frost et al., 2011), which may account for the lexical complexity of the spoken output.

5.2 Conclusions

This study attempts to explore the role of task complexity and WMC in listen-to-speak task performance which, to our knowledge, has not been investigated before. To this end, 40 university students were recruited. They were all students in the preparatory program and had completed two months of EAP instruction at the beginning of data collection. For this study, a standard integrated task i.e., TOEFL iBT listen-to-speak task was adopted. Task complexity was operationalized through an outline. WMC was measured by a complex task (OSpan) measuring executive component and another complex task (Run Span) measuring updating. Task performance assessment was based on its dimensions i.e., syntactic complexity (mean number of words per AS-unit), lexical complexity (Vocd-D) accuracy (mean number of errors per 100 words), fluency (syllables per min), and content (on a scale from 1 to 5).

The findings indicated a relationship between task complexity and dimensions of task performance. Participants performed better in simple task condition as assessed by their scores on all given dimensions. The findings did not suggest that they prioritized one aspect over others (trade-off effects), which could be attributed to the task

complexity manipulation adopted in this study. WMC was observed to be a significant predictor only in the fluency aspect of spoken performance, which may be related to the proficiency level of participants and the type of task used. As Frost et al., (2011) suggested, learners tend to rely on and reproduce the spoken input provided, which affects the performance scores in terms of complexity (structural or lexical) and content. Thus, accuracy and fluency emerge as the areas that would compete for cognitive resources. The reason why WMC was a predictor in fluency but not in accuracy can be attributed to the task design as well. This task is completed under eminent time pressure as task takers are asked to answer prompts in sixty seconds. This design, as a matter of fact, urges task takers to be fast, which is perceived as a component of fluency. Participants may have devoted their cognitive resources to fluency to be able to achieve the task. The other dimensions syntactic complexity, accuracy and content were explained by task complexity, which is in line with the findings from previous studies (Robinson, 2011). Finally, perceived task difficulty was revealed to be a significant predictor in lexical complexity. As argued above, perceived task difficulty works in tandem with task complexity operationalization. Due to methodological shortcomings, it might be suggested that what we observe, by relying on the data from questionnaire, as the effect of perceived task difficulty is a combined effect of task complexity manipulation and learners' perception.

5.3 Implications

There are certain pedagogical implications of this study. First and foremost, the findings suggested that providing an outline made the task performatively complex which indicates that [+/- outline] is a valid complexity manipulation, which can help classroom

practitioners sequence similar pedagogic tasks from simple to complex. This is especially useful while designing and providing strategy training for L2 learners. This study may also guide assessment specialists as it furthers our understanding of cognitive abilities which may influence how L2 knowledge is utilized in addition to facilitating the acquisition of L2 knowledge. Cognitive processes involved in task performance may be closely linked to the task design and characteristics of tasks; therefore, it is important to understand the cognitive demands of tasks to determine the type of tasks to be used for assessment purposes as well as for teaching purposes.

5.4 Limitations and further research

There are several limitations of this study affecting the generalizability of its findings. Sample size is a major limitation as it was relatively small. Participants were all B2 students representing only one level of proficiency. Considering the previous research findings indicating a relationship between L2 proficiency and WMC, future research with participants from different proficiency levels may provide more insight into how WMC works in L2 spoken performance. This study reports scores of integrated tasks; however, further research is required to compare scores from independent listening and speaking tasks to integrated tasks. WM tasks used in this study were both language independent tasks, which did not provide the researcher with an opportunity to compare the effect of language independent tasks with language dependent WM tasks. Finally, dimensions of task performance were scored through one measure (e.g., number of syllables per minutes for fluency). However, there are several measures available in the relevant literature for each aspect and using more than one measure for each aspect would have increased the validity and reliability of performance scores.

APPENDIX A

SAMPLE TASK

Instruction:

You will now listen to part of a lecture. You will then be asked a question about it. After you hear the question, you will have 20 seconds to prepare your response and 60 seconds to speak.

You can take notes using the outline given below.

Marketing strategies and characteristics of target customers Characteristic 1- Age Marketing strategy: Characteristic 2- Geographic location Marketing strategy:

Instruction:

Using points and examples from the lecture, explain how the characteristics of target customers influence marketing strategy for products.

Response Time: 60 Seconds

APPENDIX B

PERCEIVED TASK DIFFICULTY QUESTIONNAIRE

Please circle one of the numbers as appropriate about the task that you have just completed.

1. How easy was this task?

0	1	2	3	4	5	6	7	8	9
Very easy		Easy		Moderately easy	Moderately difficult		Difficult		Very difficult

2. How nervous were you to do this task?

0	1	2	3	4	5	6	7	8	9
Very relaxed		Relaxed		Moderately relaxed	Moderately nervous		Nervous		Very nervous

3. How well do you think you did this task?

0	1	2	3	4	5	6	7	8	9
Didn't do well at all		Didn't do well		Moderately poor	Moderately well		Well		Very well

4. How familiar were you with the topic of listening passage?

0	1	2	3	4	5	6	7	8	9
Not familiar at all		Not familiar		Moderately unfamiliar	Moderately familiar		Familiar		Very familiar

5. Was the task an effective method of testing speaking and listening skills?

0	1	2	3	4	5	6	7	8	9
Not effective at all		Not effective		Moderately not effective	Moderately effective		Effective		Very effective

APPENDIX C

ETHICS COMMITTEE APPROVAL

Evrak Tarih ve Sayısı: 01.11.2021-36423

T.C.
BOĞAZİÇİ ÜNİVERSİTESİ
SOSYAL VE BEŞERİ BİLİMLER YÜKSEK LİSANS VE DOKTORA TEZLERİ ETİK İNCELEME KOMİSYONU
TOPLANTI KARAR TUTANAĞI

Toplantı Sayısı	:	22
Toplantı Tarihi	:	13.10.2021
Toplantı Saati	:	14:00
Toplantı Yeri	:	Zoom Sanal Toplantı
Bulunanlar	:	Prof. Dr. Ebru Kaya, Prof. Dr. Fatma Nevra Seggie, Dr. Öğr. Üyesi Yasemin Sohtorik İlkmen
Bulunmayanlar	:	

Ayşe Gül Yücel
Yabancı Diller Eğitimi Bölümü

Sayın Araştırmacı,

"The role of working memory in performing listening-to-speak integrated tasks" başlıklı projeniz ile ilgili olarak yaptığınız SBB-EAK 2021/55 sayılı başvuru komisyonumuz tarafından 13 Ekim 2021 tarihli toplantıda incelenmiş ve uygun bulunmuştur.

Bu karar tüm üyelerin toplantıya çevrimiçi olarak katılımı ve oybirliği ile alınmıştır. COVID-19 önlemleri kapsamında kurul üyelerinden ıslak imza alınamadığı için bu onay mektubu üye ve raportör olarak Fatma Nevra Seggie tarafından bütün üyeler adına e-imzalanmıştır.

Saygılarımızla, bilgilerinizi rica ederiz.

Prof. Dr. Fatma Nevra SEGGIE
ÜYE

e-imzalıdır
Prof. Dr.Fatma Nevra SEGGIE
Raportör

SOBETİK 22 13.10.2021

APPENDIX D

BACKGROUND QUESTIONNAIRE

Directions: Please provide the following information by writing your response in the space or ticking (□) in the box.

1. First Name: _____

2. Last Name: _____

3. Age: _____ years

4. 1st language: □□Turkish □□Other: _____

5. Current level of study:

□□English Preparatory Programme: _____

□□Undergraduate Year of study: □□1st □□2nd □□3rd □□4th

□□Masters Year of study: □□1st □2nd □□Other: _____

□□PhD Year of study: □□1st □2nd □□Other: _____

6. Faculty: _____

7. Major subject: _____

8. Overseas experience: Have you spent a long period (at least a total of six months) in English speaking countries? □□Yes □□No

If yes, which country? _____

How long did you live there? _____year(s)_____month(s)

Yönerge: Aşağıdaki sorulara yanda verilen kutucukları işretleyerek cevap veriniz.

1. Ad: _____

2. Soyad: _____

3. Yaş: _____

4. Anadil: □Türkçe □Diğer: _____

5. Eğitim Durumunuz:

□İngilizce Hazırlık Programı: _____

□Bölüm: □1. yıl □2. yıl □3. yıl □4. yıl

□Yüksek Lisans: □1. yıl □2. yıl □Diğer: _____

□Doktora: □1. yıl □2. yıl □Diğer: _____

6. Fakülte: _____

7. Bölüm: _____

8. Yurtdışı tecrübesi: İngilizce konuşulan bir ülkede uzun süre (en az toplam altı ay) bulundunuz mu? □Evet □Hayır

Evet ise, hangi ülkede? _____

Bu ülkede ne kadar süre ile bulundunuz? _____yıl_____ay

APPENDIX E

SAMPLE TRANSCIPTION

companies use some marketing strategies to reach their target audience/ there are some characteristics/ these are considered in their target customers which are age and location/ those are the most important ones/ because let's say they are marketing for adults who are of working age/ if they are putting up ads on television they should do so after five or they are more likely at home/ and when it comes to location/ let's say you are selling boats or ships or things that need a water source you should place those ads closer to the water source so people who are considered targeted customers can see them/

most of the ads you see kind of trick you or in other words persuade you into buying products/ they use two main strategies/ the first one being repetition/ just like telling yourself something is right or faking it till you making it/ if you say something a lot of times it would seem likely it would seem trustworthy and true/ and they use that in their ads a lot of the time/ the other one being using celebrities/ because we are starstruck by them meaning we are more trusting in them and we admire them/ when celebrities promote some products we are more likely to buy it/

art impacts a child's development in two main ways/ the first one being expressing complex emotions/ because kids have limited vocabularies when they convey their emotions using their art their drawings they most of the time use body language in their drawings/ so we can find out what they are feeling/ the second one being persistence/ because they will have a clear goal of making let's say a sculpture or a drawing they will work towards that goal/ and when they accomplish it they will now have an exact model of putting in the work and then getting a result/ so this will be an important lesson for them

your locus of control is your belief on where your control comes from/ the first one being internal /and the second one being externals/ internal people are more dependent on themselves/ they will work hard towards a goal/ and if they are not successful they will take it out upon themselves and they will be more hard on themselves / external people they believe most of the things are out of their/ it is not dependent on them/ it is mostly about luck/ so they will take more risks as they will think that they will not be held responsible for it/ but this might also mean they will be lazy and blame everything on other people/

APPENDIX F

CONTENT RATING SCALE

Scale (out of 5)	Descriptors
5	The main point and supporting details are clearly presented.
4	The main point was clearly presented but a few (one or two) supporting details are missed.
3	The main point is presented but details are missed.
2	The main point is not clearly presented and details are missed.
1	Not enough sample to rate

REFERENCES

Ahmadian, M. J. (2012). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly*, *46*(1), 165-175.

Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language test construction and evaluation*. Cambridge: Cambridge University Press.

Anderson, J.R. (1995). *Cognitive Psychology and its Implications, 4th Edition*. Freeman, New York.

Andringa, S., Olsthoorn, N. van Beuningen, C. Schoonen, R. & Hulstijn, J. (2012). Determinants of Success in Native and Non-Native Listening Comprehension: An Individual Differences Approach. *Language Learning*, *62*, 49-78.

Asencion, Y. (2004). *Validation of reading-to-write assessment tasks performed by second language learners*. Northern Arizona University. ProQuest Dissertation & Theses (PQDT).

Ashcraft, M. H., & Kirk, E. P. (2001). The relationships among working memory, math anxiety, and performance. *Journal of Experimental Psychology: General, 130*(2), 224–237.

Atkins, P. & Baddeley, A. (1998). Working memory and distributed vocabulary learning. *Applied Psycholinguistics*, 19, 537-552.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience, 4*(10), 829–839.

Baralt, M. (2010). *Task complexity, the cognition hypothesis, and interaction in CMC and FTF environments*. Unpublished Ph.D. dissertation. Washington, DC: Georgetown University.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2013). Test-takers' strategic behaviours in independent and integrated speaking tasks. *Applied Linguistics, 34*(3), 304-324.

Breen, M. (1987). Learner contributions to task design. In C. Candlin & D. Murphy (Eds.), *Language learning tasks*. (pp. 23-46) Englewood Cliffs, NJ.: Prentice-Hall.

Broadway, J. M. & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*, 563–570.

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 Series* MS-29). ETS.

Brunfaut, T. & Revesz, A. (2015). The role of task and listener characteristics in second language listening. *TESOL Quarterly*, *48* (1), 141-168.

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

Crossley, S., Clevinger, A., & Kim, Y. (2014). The role of lexical properties and cohesive devices in text integration and their effect on human ratings of speaking proficiency. *Language Assessment Quarterly*, *11*(3), 250–270.

Cumming, A., Grant, L., Mulcahy-Ernt, P., & Powers, D. E. (2004). A teacher-verification study of speaking and writing prototype tasks for a new TOEFL. *Language Testing, 21*(2), 107-145.

De Bot, K. (1992). A Bilingual Production Model: Levelt's 'Speaking' Model Adapted. *Applied Linguistics*, *13*, 1-24.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford university press.

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

Engle, R. (2001). What is working memory capacity? In H. L. Roediger, III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 297-314). Washington, DC: American Psychological Association Press.

Engle, R., Tuholski, S., Laughlin, J., & Conway, A. (1999). Working memory, short-term memory, and general fluid intelligence: A latent variable approach. *Journal of Experimental Psychology: General, 128,* 309-31.

Faerch, C., & Kasper, G. (1986). Cognitive Dimensions of Language Transfer. In E. Kellerman, & M. Sharwood Smith (Eds.), *Crosslinguistic Influence in Second Language Acquisition* (pp. 49-65). New York: Pergamon Press.

Fortkamp, M. B. M. (1999). Working memory capacity and aspects of L2 speech production. *Communication & Cognition, 32*(3), 259–295.

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

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: Comparing effects on complexity, fluency, and lexical diversity. *Language Learning*, *59*(4), 866-896.

Friedman, N. P., Miyake, A., Corley, R. P., Young, S. E., DeFries, J. C., & Hewitt, J. K. (2006). Not all executive functions are related to intelligence. *Psychological Science*, *17*(2), 172-179.

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*, 1-25.

Frost, K., Clothier, J., Huisman, A., & Wigglesworth, G. (2020). Responding to a TOEFL iBT integrated speaking task: Mapping task demands and test takers' use of stimulus content. *Language Testing*, *37*(1), 133–155.

Fulcher, G. (2015). Assessing second language speaking. *Language Teaching*, *48*(2), 198-216.

Gebril, A. (2010). Bringing reading-to-write and writing-only assessment tasks together: A generalizability analysis. *Assessing Writing, 15*(2), 100-117.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching, 45,* 215-40.

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

Goh, C. (2000). A Cognitive Perspective on Language Learners' Listening Comprehension Problems. *System*, *28*, 55-75.

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

Huang, H. T. D., & Hung, S. T. A. (2010). Examining the practice of a reading-to-speak test task: Anxiety and experience of EFL students. *Asia Pacific Education Review*, *11*(2), 235-242.

Huang, H. T. D., Hung, S. T. A., & Hong, H. T. V. (2016). Test-taker characteristics and integrated speaking test performance: A path-analytic study. *Language Assessment Quarterly*, *13*(4), 283-301.

Huang, D.H., Hung, A. S., Plakans, L. (2018). Topical knowledge in L2 speaking assessment: Comparing independent and integrated speaking test tasks. *Language Testing*, *35*(1) 27–49.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *Language Learning Journal*, *44*(4), 487–505.

Inoue, C., & Lam, D. M. (2021). The Effects of Extended Planning Time on Candidates' Performance, Processes, and Strategy Use in the Lecture Listening□Into□ Speaking Tasks of the TOEFL iBT® Test. *ETS Research Report Series*, *1*, 1-32.

Iwashita, Noriko, Brown, Annie, McNamara, Tim, & O'Hagan, Sally (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching*, *44*(2), 137-166.

Kim, Y., Payant, C., & Pearson, P. (2015). The intersection of task-based interaction, task complexity, and working memory. *Studies in Second Language Acquisition*, *37*, 549-581.

Kormos, J. (1999). Monitoring and self-repair in a second language. *Language Learning*, 49, 303- 342.

Kormos, J. (2006*). Speech production and second language acquisition*. Mahwah, NJ: Lawrence Erlbaum.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*(2), 145-164.

Kormos, J., & Sáfár, A. (2008). Phonological short-term memory, working memory and foreign language performance in intensive language learning. *Bilingualism: Language and Cognition*, *11*(2), 261-271.

Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance.* Amsterdam: Benjamins.

Kormos, J., & Trebits, A. (2012). The Role of Task Complexity, Modality, and Aptitude in Narrative Task Performance. *Language Learning*, 1–34.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24.

Lee, Y.-W. (2006). Dependability of scores for a new ESL speaking assessment consisting of integrated and independent tasks. *Language Testing, 23*(2), 131-166.

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25-42). Michigan: The University of Michigan Press.

Lewkowicz, J. (1997). The integrated testing of a second language. In C. Clapham & D. Corson (Eds.), *Encyclopedia of language and education* (pp. 121-130). The Netherlands: Kluwer Academic Publishers.

Levelt, W. J. M. (1989). Speaking: *From Intention to Articulation*. Cambridge, MA: MIT Press.

Levelt, W.J.M. (1993). Lexical Access in Speech Production. In: Reuland, E., Abraham, W. (Eds.) *Knowledge and Language*. Springer, Dordrecht.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Linck, J.A. P. Osthus, J. T. Koeth, M. F. Bunting (2014). Working memory and second language comprehension and production: A meta-analysis. *Psychon Bulletin Review 21*, 861–883.

Luoma, S. (2004). *Assessing speaking*. Cambridge: Cambridge University Press.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control*. Cambridge University Press.

Mizera, G. J. (2006). *Working memory and L2 oral fluency* (Unpublished Doctoral dissertation, University of Pittsburgh).

Mota, M. B. (2003). Working memory capacity and fluency, accuracy, complexity, and lexical density in L2 speech production. *Fragmentos: Revista de Língua e Literatura Estrangeiras*, *24*.

Ockey, G. J., Koyama, D., Setoguchi, E., & Sun, A. (2015). The extent to which TOEFL iBT speaking scores are associated with performance on oral language tasks and oral ability components for Japanese university students. *Language Testing*, *32*(1), 39-62.

O'Malley, J. M., & Chamot, A. U. (1990). *Learning strategies in second language acquisition*. Cambridge, UK: Cambridge University Press.

Osaka, M., & Osaka, N. (1992). Language-independent working memory as measured by Japanese and English reading span tests. *Bulletin of the Psychonomic Society*, *30*(4), 287-289.

O'Sullivan, B., & Weir, C. J. (2011). Test development and validation. In B. O'Sullivan (Ed.), *Language testing: Theories and practices* (pp. 13–32). London: Palgrave Macmillan.

Plakans, L. (2008). Comparing composing processes in writing-only and reading-to-write test tasks. *Assessing Writing, 13*(2), 111-129.

Plakans, L. (2009). The role of reading strategies in integrated L2 writing tasks. *Journal of English for Academic Purposes, 8*(4), 252-266.

Plakans, L. (2015). Integrated second language writing assessment: why? what? how? *Language and Linguistics Compass, 9*(4), 159-167.

Plakans, L., & Gebril, A. (2012). A close investigation into source use in integrated second language writing tasks. *Assessing Writing, 17*(1), 18-34.

Poulisse, N. (1997). Some words in defense of the psycholinguistic approach: A response to Firth and Wagner. Modern Language Journal, *81*(3), 324- 328.

Révész, A. Ekiert, M. & Torgersen, E. N. (2016). The Effects of Complexity, Accuracy, and Fluency on Communicative Adequacy in Oral Task Performance. *Applied Linguistics*, *37*(6), 828–848.

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

Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies*, *21*(2), 45-105.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics*, *43*, 1-32.

Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In M. d. P. Garcia Mayo (Ed.), *Investigating Tasks in formal Language Learning* (pp. 7-26). Clevedom Multilingual Matters.

Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL*, *45*, 161-176.

Robinson, P. (2011). Second language task complexity, the Cognition Hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task*

complexity: Researching the cognition hypothesis of language learning and performance. Amsterdam: John Benjamins.

Rukhthong, A. and Brunfaut, T. (2020). 'Is anybody listening? The nature of second language listening in integrated listening-to-summarize tasks. *Language Testing*, *37*(1), pp. 31- 53.

Serafini, E. J. & Sanz, C. (2016). Evidence for the decreasing impact of cognitive ability on second language development as proficiency increases. *Studies in Second Language Acquisition*, *38*, 604-646.

Sakuma, Y. (2004). The characteristics of memory representations in listening span test and EFL abilities. *ARELE: Annual Review of English Language Education in Japan*, *15*, 91-100.

Sawaki, Y., Stricker, L. J., & Oranjie, A. H. (2009). Factor structure of the TOEFL Internet-based test. *Language Testing, 26*(1), 5-30.

Shanshan, G., & Tongshun, W. (2007). Study on the relationship between working memory and EFL listening comprehension. *CELEA Journal*, *30*(6), 46–55.

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

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

Skehan, P. (2014). *Processing Perspectives on Task Performance*. London: John Benjamins.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*, 93-120.

Swain, M., Huang, L.-S., Barkaoui, K., Brooks, L., & Lapkin, S. (2009). *The speaking section of the TOEFL iBTTM (SSTiBT): Test-takers' reported strategic behaviours (TOEFL iBTTM research report)*. Princeton, NJ.

Tabachnick, G. B. & Fidell, S. L. (2007). *Using multivariate statistics*. Boston, MA: Pearson Education, Inc.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied linguistics*, *17*(1), 84-119.

Trebits, A., & Kormos, J. (2008). Working memory capacity and narrative task performance. In Proceedings form the 33rd international LAUD symposium, Landau/Pfalz, Germany.

Unsworth, N., Heitz, R. P., Schrock, J. C., and Engle, R. W. (2005). An automated version of operation span task. *Behavioral Research Methods*, *37*, 498-505.

Weigle, S. C. (2004). Integrated reading and writing in a competency test for non-native speakers of English. *Assessment Writing, 9*, 27-55.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Great Britain: Antony Rowe.