EFFECTIVE PLANNING IN REAL-TIME SPEAKING TEST TASKS

By

Shinhye Lee

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirement
for the degree of

Second Language Studies—Doctor of Philosophy

2018

ABSTRACT

EFFECTIVE PLANNING IN REAL-TIME SPEAKING TEST TASKS

By

Shinhye Lee

In this dissertation, I documented the effectiveness of a particular test-taking condition in the current task-based performance testing (e.g., TOEFL iBT, IELTS, OPI); namely, planning time (the time given for test takers to plan their responses before actually performing). In assessment contexts, the construct of planning addresses both language and test-related theories; in terms of the latter, it is associated with components constituting test qualities (test validity, authenticity, and fairness; Wigglesworth & Elder, 2010). This is because planning time is already a critical test accommodation, or a task implementation condition, as termed from the task-based research paradigm. Indeed, researchers and test developers suggest planning time as a major component in determining task difficulty in that the varying planning conditions are appropriate to represent the items' different cognitive-demand levels (Norris, 2009; Robinson, 2001). Therefore, I explored whether test takers' efficient use of varying planning times is contingent upon test-task characteristics in the context of the TOEFL iBT Speaking test, in which which varying degrees of task conditions such as planning, and test-task types co-exist.

Ninety-nine Korean university students took three speaking tests, which each consisted of one independent task (impromptu task) and two integrated tasks (reading-listening and listening-only tasks). As in operational testing, independent tasks were given 15 seconds to plan while the two integrated tasks were given 30 and 20 seconds, respectively. For each test set, participants performed under a specific planning condition: namely, Unguided planning (planning without specific instructions), Guided planning with Writing (planning with instructions given as to write

to plan), and Guided planning with silently thinking (planning with instructions given as to think or outline silently to plan). After each test, test takers partook in a series of surveys and interviews to reflect on the appropriateness of each task's planning times. Subsequently, I undertook multiple methods (quantitative and qualitative) on the collected data points. Three independent raters scored a total of 891 speech samples according to the TOEFL iBT speaking rubric. Two trained coders coded the speech samples for the three discourse quality measures pertaining to *complexity*, *accuracy*, and *fluency*. I thematically analyzed survey and interview responses through NVivo.

Findings indicated that participants' performance and perceptions were directed by the influence of test-task characteristics regardless of the planning activities they made use of. Their test performance and speech quality varied from independent tasks to integrated tasks; they generally scored lower, and demonstrated slower speech rate, increased lexical errors, and simplified language in independent tasks. In addition, test takers believed that extended planning time was unnecessary for integrated tasks, for the reading and listening sources were readily applicable to actual responses; yet 15 seconds did not suffice for them for the independent tasks to familiarize themselves with the given prompt.

I discuss study results and conclude the dissertation by making connections between speech-production planning theories in task-based research (Robinson, 2001; Skehan, 1998) and planning time practices in second language assessment (Elder & Iwashita, 2005; Wigglesworth & Elder, 2010).

ACKNOWLEDGMENTS

Completion of this dissertation would not have been possible if it were not from the tremendous support of a number of people. I would like to take this opportunity to extend my thanks and appreciation to them.

First of all, I want to express my sincerest gratitude to my advisor and dissertation chair Dr. Paula Winke, who took me under her wing from the very beginning to the end of my Ph.D. journey in the Second Language Studies program. I am grateful for her support and encouragement, which helped me believe in myself and try out more. Thanks to her, I was able to grasp a number of exciting opportunities during my Ph.D. studies that immensely helped me shape my career path early on. I give my deepest thanks to my dissertation committee members, Drs. Susan Gass, Daniel Reed, and Koen Van Gorp, for their critical and insightful comments from the earlier phase of my dissertation project to the final manuscript. Special thanks go to Dr. Patti Spinner for devoting her time to support my job search this past academic year. I am also grateful for my advisor at Ewha Womens University, Dr. Sang-Keun Shin, for his continuous support until this day from my master's studies in Korea.

Special thanks also go to Hima Rawal, Joshua Smith, Chad Bousley, Chris Bartoluzzi, Laura Bowman, Erin Degerman, Brandon Jung, and Aaron Ohlrogge, for putting substantial amount of time and efforts to help me out in coding, rating, and organizing my dataset. If it were not for their help, I would not have been able to complete any part of the main data analysis for my dissertation.

I want to thank deeply my friends and colleagues at Michigan State University: Myeongeun Son, Jongbong Lee, Xiaowan Zhang, Melody Ma, and Michael Wang. They have

demonstrated true friendship and have stood by me in both good and bad times. I feel very lucky to call them my friends and cannot wait to see them thrive as the academics that they hope to become.

Last but not least, I am forever grateful for my Mom and Dad, Sun-Rye Lim and Yoon-Jae Lee, for their unconditional love and patience. I want to congratulate them on finally witnessing their daughter graduate and start off a career, after all these seemingly endless years of schooling.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

In performance-based testing, L2 oral test performance refers to one's production of language on a series of test tasks. As opposed to discrete-point item types, on these assessments, test takers are expected to *act upon* a given test task (Davis, Brown, Elder, Hill, Lumley, & McNamara, 1999); that is, they are to provide their oral responses on the basis of how they process and undertake the presented test tasks. Within such contexts, test developers wish to have test takers perform on a variety of test tasks that when taken as a whole, they broadly represent and tap onto the target language ability, or the test construct (Bachman & Palmer, 1996). More precisely, test tasks are designed in the way that closely resemble the identified language outcome yet differ from one another in terms of the conditions of task implementation that make up each test task uniquely reflect an array of their real-world counterparts (Elder, Iwashita, & McNamara, 2002). For instance, TOEFL iBT Speaking sections take up the format of which test takers advance through varied task types, each comprised of varied parameters of task conditions underlying each task type. Test takers may first take an impromptu test task, then make their way to subsequent task types, which require the channeling of the integration of other language skills than speaking for task completion. Yet taken altogether, the different task types are designed to make up a body of *academic speaking*, which is the core construct of the TOEFL iBT Speaking test.

The elements differentiating each test task are allegedly known to have theoretical underpinning, particularly drawn from the task-based research (Robinson, 2001; Skehan, 1998). The premise of this line of work is that certain task characteristics and performance conditions have an additive function to *complexity* and/or the *difficulty* of task types. For instance, whether given the option of planning for subsequent responses in advance constitutes a major component

of the inherent complex nature of tasks. The consensus in the field is that the provision of planning lessens one's online processing load during real-time production, and leads to better performance (e.g., Crookes, 1989; Ellis, 2005, 2009). The option of planning may also contribute to how speakers perceive certain tasks to be much difficult from others (Robinson, 2001); tasks inherent with planning may nurture higher confidence or motivation of task completion amongst speakers (Tavakoli & Skehan, 2005). When applying this thinking to performance-based, task-based language testing setting, further understanding of different underlying factors of the dimensions of task *complexity* and *difficulty* can underscore both (a) test-takers' levels of test performance and (b) perceptions toward corresponding test-task conditions. As a consequence, it would contribute to informed decisions of selecting suitable range of tasks for assessment purposes of oral language and see how they meet with the test developers' hunches in task design (Elder et al., 2002).

In this dissertation, I apply insights from the task-based research in exploring the test-task characteristics of the TOEFL iBT Speaking test. Particularly, I aim to examine the quantitative and qualitative difference in test performance amongst the two TOEFL iBT task type, namely, independent and integrated tasks in light of their inherent planning conditions. I specifically attend to the planning conditions in the TOEFL iBT oral tasks as its differentiated employment per test-task type are based on *ad hoc* decisions (Elder & Iwashita, 2005), which lacks clear articulation of (1) the link between a precise planning condition and a test-task type (Butler, Eignor, McNamara, Jones, & Suomi, 2000) and (2) how the theoretical underpinning of speech production (that has guided the task-based research) might shed light on the precise test-task design (Elder et al., 2002; Mislevy, Steinberg, & Almond, 2003; Wigglesworth & Frost, 2017). Unanswered questions regarding the TOEFL iBT task conditions are: (1) Why are the planning

times given differently across test-task types? and (2) Would such difference in allotment per test-task type matter in how test takers perform and react to certain test-task types? Exploring the very nature of a specific task condition inherent in test-task types affect subsequent test-taker' performance and perceptions could render insights as to the appropriateness of the test-task design, and further inferring information on related test qualities (e.g., construct representativeness, authenticity, and practicality; Elder & Iwashita, 2005; Wigglesworth & Elder, 2010).

In addition, this study was set to promote more ecological validity in language testing research on pre-task planning. To date, there has been a lack of contextualized language-testing studies on planning time that adopt the actual planning conditions in language tests (Xi, 2010). Few existing studies follow the steps of laboratory-based SLA studies by providing a longer stretch of planning time that is not plausible in testing settings, thereby improperly reflecting the timed nature of the testing situation (Ellis, 2009). Another parameter of interest is the provision of guidance in planning. Current test directions often give little or no information on how planning should be done while test publishers require test candidates to use their preparation time as effectively as possible (Elder & Iwashita, 2005). Considering that tests should be biased for the best (Swain, 1984), it is of great interest whether guiding test takers to engage in certain planning activities lead to meaningful test taking.

The central aim of my dissertation, therefore, is to conduct a *de facto* documentation of the interplay between planning-related variables and test-task types in the TOEFL iBT. Given that planning time is pervasive in real testing, the ultimate goal of this research inquiry is not to argue for the removal of such planning practice; the overall goal is to seek for and suggest ways to promote meaningful planning conditions within this particular testing environment.

CHAPTER 1: Literature Review

In this chapter, I first outline the theoretical models of speech production taken by the task-based research and make connections to language testing research on the effects of task characteristics, especially, pre-task planning. Then, I present both testing and non-testing research studies on task characteristics that highlight the role of pre-task planning on oral performance. To illuminate the effects of planning-related variables on oral performance, I take the order in this sub-section of first delineating how oral performance has been defined and commonly measured in conventional task-based research (e.g., *complexity*, *accuracy*, and *fluency* dimensions) followed by what needs to be adapted in relevant language testing research. In the subsequent sub-sections, I then introduce research studies on the effectiveness of varied planning parameters on oral performance (which are the core interest of the present study) such as the length of planning time, planning types (pre-task and within-task planning), and speaker-directed planning activities. I also touch on research studies on the TOEFL iBT Speaking test-task types, but only briefly, given the scarcity of relevant research accounts. At the end of the chapter, I provide the overall purpose and rationale of the study by building upon the previous line of research and addressing the gap in the literature.

1.1 Task-based research and theoretical underpinning

1.1.1 Modeling and researching speech production and the effects of pre-task planning

While task-based research on oral performance has drawn from a number of theoretical perspectives, Levelt's (1989) model of speech production has served as the major basis of accounting for the effects of pre-task planning on oral performance (Ellis, 2009). The model specifies a three-step procedure of speech production: (1) conceptualization; (2) formulation; and (3) articulation. In the conceptualizing stage, speakers go through three sub-phases of selecting

the information subscribing to their intended communication goals. They first set their goals for articulation (i.e., macro-planning of speech), which is then followed by retrieving information necessary for meeting those goals (e.g., determining relevant speech acts), and finally making use of the selected information for achieving the set goals (i.e., micro-planning of speech, conceptualizing a pre-verbal message). Based on the general plan of what information they would draw onto, speakers go through a formulation stage in which they establish linguistic representation of the pre-verbal message. At the very least, they retrieve relevant lexical items from their mental lexicon, which serve as the building blocks of further grammatical encoding. Finally, such linguistic information gets processed by a phonological encoder, which further shapes internal speech (Levelt, 1989), which is the speakers' internal representation of how the message should be articulated. In the final articulation stage, the speakers transform their internal representation into overt speech. The entire process is very much driven by the self-monitoring practices of speakers; Levelt claims that they particularly attend to three sub-processes of (1) matching the internal speech to the identified communication goal, (2) scrutinizing the internal speech before being articulated, and (3) inspecting the over speech that has been generated.

Yet Levelt also asserted that some sub-components of the model are operated automatically, not requiring conscious efforts or controlled processing. This particularly pertains to the formulation and articulation of speech; that is, once the blueprint has established, speakers are able to generate speech largely in an automatic and smooth fashion (Kormos, 2011). Because of this point, SLA researchers made subsequent adaptations to the model to better account for the discrepancy in L1 and L2 speech production (De Bot, 1992; Ellis, 2009; Kormos, 2011). While the L2 speakers also go through a similar order of the three-stages speech production (Kormos, 2011), for them, many of the linguistic properties in L2 are not automatized/limited and thus

benefit more from close inspection or monitoring within and across the stages. The advantageous nature of pre-task planning on L2 speech production precisely lies here; the facilitation of the limited resources L2 speakers might have in retrieving linguistic information, forming internal speech, and finally articulating overt speech.

However, other accounts (Batsone, 2005; Ellis, 2009) have extended Levelt's model further in incorporating the role played by the individual differences on speech production. That is, how learners perform will result from not only their on-going monitoring of speech, but also from how they orient to a task (Batsone, 2005; Tajima, 2003). Ellis (2009) supported this view in explaining why language testing research on planning has demonstrated null effects in planning, which is discrepant from the findings in the conventional laboratory-based research. It seems to be critical, therefore, to inspect the effect of planning from multiple data points, including not only the quantitative accounts, but also the qualitative orientations of speakers to better ascertain the effects of pre-task planning.

1.1.2 Modeling and researching task-based oral test performance: Skehan's (1998) expanded model on Kenyon-McNamara (McNamara, 1995)

While language testing research on task characteristics is rooted in task-based research, how different properties of test-tasks influence subsequent test performance has not been precisely streamlined through an organized framework. Skehan (1998), in this sense, proposed a working model that specifically pertains to explaining how test performance exert differences based on different test-task parameters.

Skehan (1998) devised a framework drawing onto the initial model of Kenyon-McNamara (McNamara, 1995), which essentially illustrated the intertwined relationship amongst different factors influencing task-based test performance (e.g., underlying competence of the test

candidate, interlocutor to whom the test candidate speaks with, etc.). Skehan built upon Kenyon-McNamara's model as it takes into account that it is the task functioning as a central unit within a testing context. As shown in Figure 1, Skehan included three new factors to the initial model, namely, *ability for use (dual-coding)*, *task characteristics*, and *task conditions.* The premise of including an element such as *ability for use (dual coding)* is to account for the way processing is adapted to performance conditions. Skehan claimed that previous models of language competence (e.g., Bachman's strategic competence; Canale and Swain's communicative competence framework) conceptualized language competence as a static, generalized entity. However, real-time performance is in effect the consequence of fluctuating communicative demand for which test takers adjust to performance conditions by allocating their attentional capacities in appropriate ways. That is, test takers need to make use of different channels of their attentional capacities according to what is demanded by the given test-tasks. Skehan coins the term *processing competence*, which is test-takers' ability to handle the different demands imposed in tasks by flexibly making use of appropriate processing resources available. For instance, the linguistic modes that test takers make use of would be different from performing on a task requiring precise communication or emphasis on form to a task putting greater influence on effective, elaborated communication. In this sense, Skehan subscribed to the competitive nature of performance qualities such as *complexity*, *accuracy*, and *fluency* in that speakers put different processing goals within performance per tasks (see next section for Skehan's and Robinson's differing accounts on how the three dimensions of performance function within performance). Skehan further made a connection between the different processing goals test takers subscribe to and the scoring decisions on the emphasis of performance aspects prioritized by raters and rating scales.

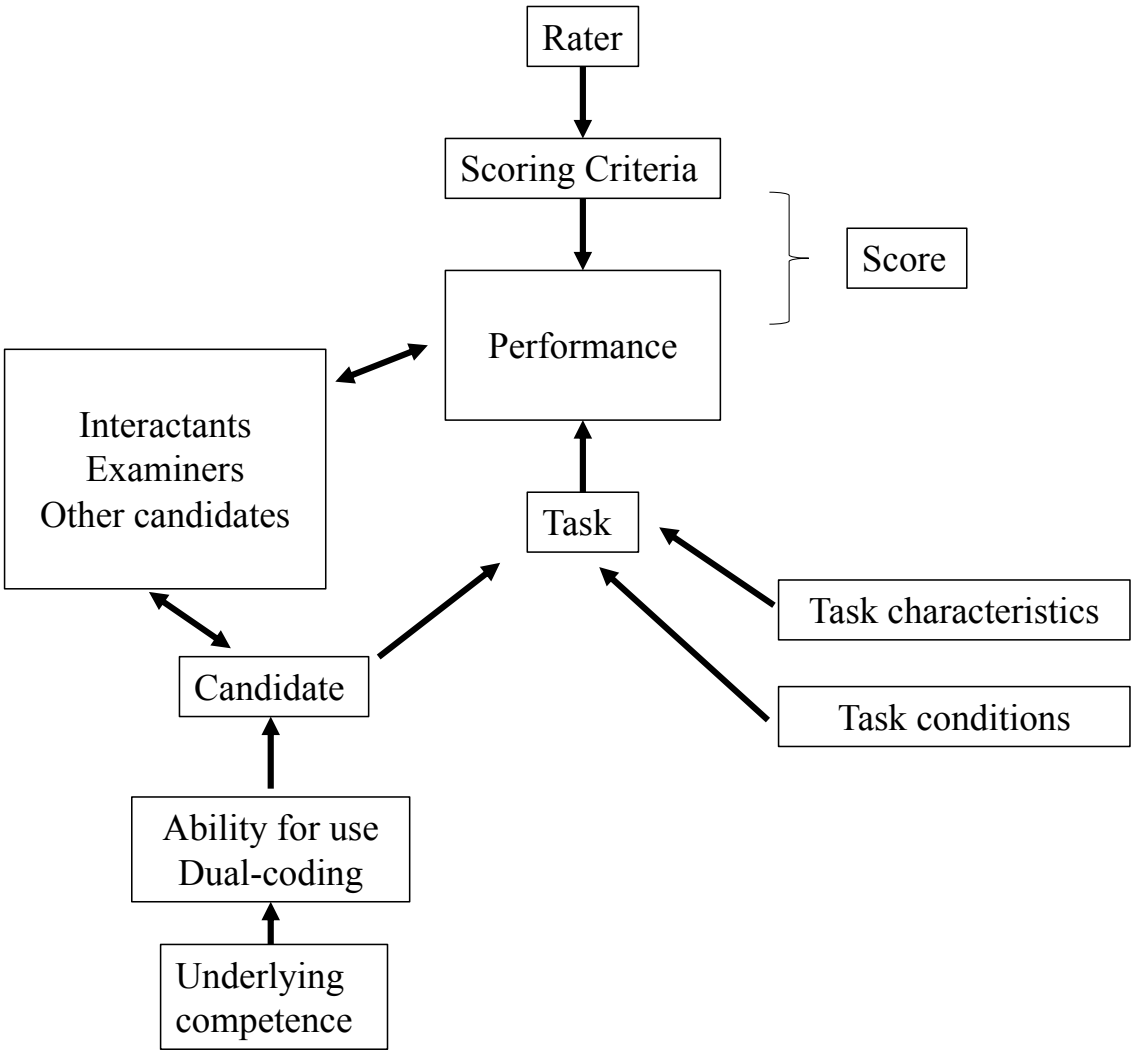


Figure 1 *Skehan's (1998, 2001) expanded model of oral test performance*

Therefore, it becomes vital to understand test-takers' oral test performance in conjunction to the conditions of task implementation (*task conditions*) and the inherent structures of tasks (*task characteristics*). *Task conditions* refer to task-external features that test takers are able to make use of; it is the manner in which the test-task is done (Skehan, 1998). For instance, in [+ planning] condition, test takers are able to shape their responses using the planning time given to them yet in [- planning] condition, they are required to make immediate responses. On the other

hand, *task characteristics* are the inherent entity or nature that makes up particular test-tasks; these elements are related to the content of test-tasks. For instance, integrated test-tasks require the dual- or triple demand in language abilities for task completion; test takers need to carry out reading and listening in addition to speaking to extract information necessary for task completion. Independent tasks, on the other hand, do not provide contextual information that test takers can refer to; hence, they do not need to make use of other language skills for task completion.

All in all, Skehan's model addressed two points: (1) the need to interpret test performance in terms of the design of test-tasks (elements that make up both *task condition* such as the provision of planning, and *task characteristics*, the intrinsic factors differentiating across test tasks) and how it influences test-takers' real-time processing orientations; and (2) the consideration of both test-tasks and test-takers' processing, which informs the sampling of test-tasks that is necessary for generalization.

1.2 Effects of task condition and task characteristics: Planning conditions and task features

Before I present research studies on the effects of planning conditions and task characteristics in detail, I introduce the constructs used in the task-based research on measuring language production. I provide this examination to further make clearer juxtaposition within and across planning-related variables commonly researched in current scholarship. I also do so to review appropriate measures and methodologies to be applied in the language testing research on planning conditions and task features.

1.2.1 Measuring the effects of planning and task characteristics

1.2.1.1 Conventional Task-based Research on planning and task characteristics: Complexity, Fluency, and Accuracy

Acknowledging the multidimentionality of language production (Tavakoli & Skehan, 2005; Housen & Kuiken, 2009), the effects of planning and task characteristics in task-based research have been comprehensively captured with regard to the three aspects of language production, namely, *complexity*, *accuracy*, and *fluency* (henceforth CAF; Ellis, 2005; 2009; Skehan, 1998; Skehan & Foster, 1999). While the three terms have been each elaborated by multiple perspectives in the field, relevant researchers have commonly subscribed to Skehan (1996) and Skehan and Foster's (1999) operational definitions as the starting point of conceptualization. In addition, by taking up such a three-way distinction, researchers have conceptualized each construct as consisting of a number of inter-related facets that establish a basis of units of data analysis for explaining the effects of planning and task characteristics.

Fluency, as defined by Skehan and Foster (1999), refers to "the capacity to use language in real time, to emphasize meanings, possibly drawing on more lexicalized systems" (p. 96). More precisely, it is a speaker's ability to produce real-time language "without undue pausing or hesitation" (Skehan, 1996, p. 22). In this sense, the notion of *fluency* takes up two standpoints from both a speaker and a listener; it is the extent to which a speaker can spontaneously come up with meaningful language units in real-time contexts as well as his or her ability to fluidly deliver the language, ultimately facilitating listener's comprehensibility (e.g., fewer signs of inappropriate and excessive pausing) (Trofimovich & Baker, 2006). Therefore, *fluency* is commonly broken down into three specific components, namely: (1) speed fluency (e.g., the number of syllables/words produced in a given time); (2) breakdown fluency (e.g., the number of

silent pauses or the number of fillers); and (3) repair fluency (e.g., speech phenomena pertaining to repetition, self-corrections, etc.).

Accuracy is defined as "the ability to avoid error in performance, possibly reflecting higher levels of control in the language as well as a conservative orientation, that is, avoidance of challenging structures that might provoke error" (Skehan & Foster, 1999, p. 96). While the concept of *accuracy* is relatively straightforward (i.e., error-free speech; Tavakoli & Skehan, 2005), there has been much debate in what constitutes *errors* and how one might measure them (Polio, 1997) in addition to the *comparative fallacy* (i.e., the native-speaker fallacy; Bley-Vroman, 1983) that the field of SLA has often been fallen into. As a result, researchers have preferred to use a mixture of both general and task-specific measures (Housen & Kuiken, 2009; Norris & Ortega, 2009) to account for the multi-faceted nature of *accuracy.* These include: (1) generalized, global measures such as percentage of error-free clauses in the given speech; and (2) specific grammatical (e.g., morpheme usages such as past tense *-ed*) or lexical features (i.e., appropriate use of words).

Lastly, *complexity* is broadly operationalized as "the capacity to use more advanced language" (Skehan & Foster, 1999, p. 97). More precisely, it involves "a greater willingness [of speakers] to take risks, and use fewer controlled language subsystems" (Skehan & Foster, 1999, p. 97) in that the resulting language would demonstrate "the size, elaborateness, richness, and diversity of the learner's linguistic L2 system" (Housen & Kuiken, 2009, p. 464). Because *complexity* can refer to various language subsystems such as vocabulary, morphology and syntax, it has spurred much controversy in task-based research as to its dimensionality and measurement methods (e.g., Bulté & Housen, 2012; Housen & Kuiken, 2009). Yet essentially, as in the case of *accuracy*, researchers have resorted to the analysis of (1) subordinization (e.g., the

number of clauses per syntactic unit); and (2) lexical complexity (e.g., type-token ratio or specific measures on speech cohesion).

1.2.1.2 Conventional Task-based Research on planning and task characteristics: Competition of Performance Constructs

As much as the variability in what holds conceptualizing CAF constructs, there has been much debates on how to interpret the mechanism behind the functioning of each dimension as per varied task conditions. Two prevailing accounts are of particular interest in the present study: The Limited Capacity Hypothesis (Skehan, 1998; Bygate, 1996; Skehan & Foster, 1997, 1999), and the Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2011).

The tenets of the Limited Capacity Hypothesis hold the view that the limited attention mechanism and processing capacity has collateral effects on language production. Speakers may not be able to adhere to the three CAF dimensions in a parallel manner, but rather selectively concern on one or two aspects owing to their constraints in attentional allocations. More specifically, researchers point out two levels of competition per planning and task characteristics: (1) *fluency* in competition for attentional resources with *accuracy*; and (2) *accuracy* and *complexity* facilitated at the expense of one another. The distinction between *fluency* and *accuracy* basically takes up the initial discussion of *focus on meaning* and *focus on form* (VanPatten, 1990); it is essentially a tension between getting the task completed (represented by the *fluency*/*meaning* dimension) and a commitment of attention to form; the latter further entails a concern for both *complexity* and *accuracy.* Skehan and Foster (1997) (and generally the task-based research) maintain that while *fluency* relatively remains stable in [+ planning] condition, *complexity* and *accuracy* does not go hand in hand. For instance, tasks inherently structured and concrete (e.g., picture-description tasks) would facilitate accurate but less complex speech since

the provided information sources depicted by the task frees up attentional space for idea elaboration in planning, and hence facilitating more focus on language forms. On the other hand, tasks that are less structured and abstract in nature (e.g., decision-making tasks) would have speakers to devote their attention to idea generation and elaboration during planning, while giving little room to attend to *accuracy*.

Robinson conversely proposed the Cognition Hypothesis, which essentially rejects the idea of competitions amongst performance constructs. He asserted that learners can simultaneously allocate attentional resources to both *accuracy* and *complexity* in speech. Essentially, he conceptualized tasks as resource-directing and resource-dispersing dimensions, and that higher complexity in these two types of task dimensions contribute to both accurate and complex speech. In the resource-directing tasks (e.g., requiring temporal reference, and/or spatial reasoning), speakers are able to simultaneously allocate their attention to the aspects of language code as well as elaboration. For instance, simple narrative tasks on an event happening now would impose lesser cognitive demand on speakers; yet tasks requiring speakers to elaborate on past events would encourage their control of morphology (e.g., past tense verb forms) as well as syntax (e.g., sentence connectors and adverbials of time and location) depicting what had happened in the past. Similarly, *accuracy* and *complexity* can be attended simultaneously in resource-dispersing tasks, for which complexity across relevant tasks is increased with the inclusion/exclusion of a number of task elements. For instance, as opposed from simple two-factor comparison tasks, tasks involving multiple comparison of factors may facilitate speakers' control of a wide range of deictic expressions, pronouns or relative clauses.

The theoretical underpinning of task-based research in understanding the effects of planning and task characteristics serve the basis of testing-linked task-based research (Tavakoli

& Skehan, 2005) (also see Ellis, 2009, pp. 478 – 490 for a synthesis of task-based and language testing research on planning effects). The linked research provides insight as to the inter-connectedness amongst speakers' (and presumably test-takers') processing capabilities and intrinsic task condition/characteristics on language production.

1.2.1.3 Language assessment literature on planning: Towards a triangulated approach

Task-based researchers have been embarking on illuminating the effects of planning and task characteristics in support of one of the two hypotheses (Skehan & Foster, 1999 for Limited Capacity Hypothesis; Gilabert, 2007; Kuiken & Vedder, 2008; Robinson, 1995, for Cognition Hypothesis). However, as researchers have pointed out (Ellis, 2009; Skehan, 2009), the evidence is still limited as the effects of task characteristics vary per study contexts (laboratory, classroom-based versus testing contexts) as well as the specific measures adopted for the three constructs. Particularly, how the three performance constructs operate have been mostly understood within the conventional classroom-based study contexts. Little is known about how the mechanisms apply to a unique research setting of language testing in which speakers are essentially imposed with added real-time pressures of performing, and whether the debate between the two hypotheses stand valid in those contexts as well (Ellis, 2009). For instance, thus far, task-based researchers have generally supported the positive role of [+ planning] on *fluency* while rendering mixed results for *accuracy* and *complexity* (e.g., Crookes, 1989; Foster & Skehan, 1996; Gilabert, 2007; Skehan & Foster, 2005; Tavokoli & Skehan, 2005; Wigglesworth, 1997; Yuan & Ellis, 2003). Yet the few relevant language testing research basically found the reverse: very weak effects of planning conditions and/or concrete task features as measured via three performance measures (e.g., Wigglesworth, 2001) (see next section for the review of these research studies). Implicit in this line of research is that there could be an underlying potential

discrepancy in study contexts on the effects of planning as well as how the three dimensions operate (Elder & Iwashita, 2005; Ellis, 2009). Particularly, the three constructs may be captured in an inherently different way in light of varied planning conditions between classroom-oriented contexts and testing environment.

Accordingly, few language testing researchers harmonized different data sources in addition to the three-dimension analysis such as (1) subjective ratings of speech (Elder & Iwashita, 2005; Elder et al., 2002; Wigglesworth & Elder, 2010) and test-taker perceptions (Elder & Iwashita, 2005), and (2) discourse-level qualitative differences in speech (Nitta & Nakatsuhara, 2014). Such a triangulated approach seems to be appropriate in finding evidence as to discerning the difference between the study contexts, but ultimately, it would exert more significance in undertaking the nature of validity argument inherent in language assessment research. Yet to date, only few studies have taken such an approach in the contexts of operational testing settings. Relevant researchers mostly simulated the study contexts of laboratory, task-based studies, but only to employ a shortened version of the conventional planning conditions (by reducing the length of planning/responding time) (Ellis, 2009). There is still more room for language testing research on planning and task characteristics that undertakes the conceptual framework posed from the conventional task-based research while taking a contextualized approach for studying the real-time effects of such task-related variables on test performance.

1.2.2 Research studies on the effects of planning time and planning conditions

As mentioned, previous researchers who investigated pre-task planning have acknowledged the discrepancy between the findings of classroom and language testing research for which a number of planning parameters have contributed (Elder & Iwashita, 2005; Ellis, 2005, 2009). In this section, I further present the potential mediating role accounting for such

discrepancy through reviewing combined factors such as the length of planning time (Wigglesworth & Elder, 2010) and the provision of detailed guidance for planning (Skehan & Foster, 1997).

1.2.2.1 Length of planning time

Of primary concern in what separates the language testing research on planning from the findings from conventional task-based research is the time variable (Ellis, 2009); namely, the amount of time allocated for planning. In his synthetized review of the effects of planning time, Ellis (2009) reported that the majority of classroom and laboratory-based studies that took the planning versus no-planning approach provided 5 to 10 minutes prior to eliciting oral output, with 10 minutes being the standard planning time among relevant literature (e.g.,Gilabert, 2007; Kawauchi, 2005; Mochizuki & Ortega, 2008; Skehan & Foster, 1997, 2005; Tajima, 2003; Yuan & Ellis, 2003).

On the other hand, the few testing studies on pre-planning allocated a much shorter time, conceivably prompted from reflecting on the demands of the timed condition of testing (Elder & Iwashita, 2005; Elder, Iwashita, & McNamara, 2002; Nitta & Nakatsuhara, 2014; Xi, 2005; Wigglesworth, 1997, 2000; Wigglesworth & Elder, 2010). In these studies, the effect of planning appeared inconsistent, being less likely to be positive in terms of test performance while demonstrating mixed extent of impact on CAF measures. For instance, Wigglesworth (1997) allowed one minute of planning time for adult ESL learners taking a tape-mediated oral test. Generally, only a minute of planning was reported to have effects on *fluency* (e.g., number of self-repairs), *accuracy* (e.g., article usage), and *complexity* (e.g., subordination) for the planning group; this was particularly true to those with higher language proficiency performing for a relatively more difficult task (e.g., summary of conversation). Yet Wigglesworth (1997) found

no substantial score differences between planners and no-planners based on an analytic rubric. In her subsequent study, Wigglesworth (2000) only focused on test scores with the provision of the same 1-minute planning; a consistent finding of insignificant score differences were found for unstructured tasks. More recently, Wigglesworth and Elder (2010) conducted a study comparing three conditions of no planning, 1-minute of planning, and 2-minutes of planning on ninety test candidates on an IELTS-type oral test format. Again, the researchers did not find any significant differences in oral performance according to the amount of planning time provided neither on analysis of the scores nor the employed CAF measures.

Elder et al. (2002) and Elder and Iwashita (2005) both provided 3 minutes of planning time (with additional 75 seconds to read instructions, which was also provided for the no-planning group) for adult test takers prior to a series of narrative tasks extracted from the Test of Spoken English (TSE). Similar to the research conducted by Wigglesworth and her colleagues, both studies failed to demonstrate significant effects of planning on test performance as depicted by the Multi-Faceted Rasch Modeling, as well as further statistical analyses. The researchers also noted hardly any difference between [+ planning] and [- planning] in terms of different data points related to CAF dimensions: namely, the subjective ratings resulting from the holistic rating scale of each CAF constructs that they had devised as well as the manual coding undertaken for individual CAF measures. The researchers also did not find a positive effect of [+ planning] on test-takers' affective orientation towards the planning task condition.

A recent study of Nitta and Nakatsuhara (2014) also provided 3 minutes of planning time in the context of paired oral testing. The researchers had 32 students perform two decision-making tasks in pairs, both under planned and unplanned conditions. The researchers conceptualized *test performance* from applying a range of CAF measures and conversation

analysis (CA). The researchers found a conflicting evidence as to the sub-components of *fluency*; while [+ planning] condition contributed to only a slight increase in *breakdown fluency* measures, it was detrimental to *speed fluency* measures. Yet the CA analysis indicated that in the [+ planning] condition, test takers attempted to produce longer turns while recalling what they had planned, which in turn reduced the intra-run speech. On the other hand, in [- planning] condition, test takers initiated more animated, spontaneous shorter turns that subsequently enabled them to talk faster despite the increased likelihood of cognitive demands of online planning under the unplanned conditions.

One rare study by Xi (2005), which was essentially not a planning-oriented study, found an effect of only a minute of planning time for a more structured test task (graph description) on the increase of the oral performance of college-level test takers of the Speaking Proficiency English Assessment Kit (SPEAK) exam. The researcher reported that the provision of planning time operated as to mediating test candidates' unfamiliarity with the given complex test task.

Such a consistent neutral finding renders a number of explanations (see Elder & Iwashita, 2005) yet importantly, the limited amount of time (minimum of 1 minute and maximum of 3 minutes) given may have been insufficient to contribute to significant score discrimination (Mehnert, 1998). This is a critical matter in that the actual time allotted for test tasks that are more structured and complex (e.g., academic-skills-oriented test tasks) are even lesser than a minute (Elder & Iwashita, 2005). In this sense, the aforementioned studies lack ecological validity, often not specifying justifications as to the provided amount of planning time. As Xi (2005) noted, the time allotted to planning in these studies is not contextualized within the existing testing contexts. This is only being recognized in recent scholarship: Wigglesworth and Elder (2010) mentioned that their use of 1 minute of planning was based on the current

instructions for the IELTS Part 2 that stated "one to two minutes to prepare" should be provided for test candidates. Still, little attention is directed to the need to contextualize the planning condition in the actual testing environment, which would be an essential research inquiry contributing to evaluate the authenticity and validity of the corresponding tests.

1.2.2.2 Planning activities

Another important parameter on planning that is still under-researched in the domain of language testing is the planning activities of which the test candidates could make use. As opposed to previous planning literature in task-based framework (Ellis, 2009), much attention has been given to the effectiveness of the provision of planning itself, thereby drawing comparisons between test performance derived from the planning, as opposed to the non-planning conditions (e.g., Elder & Iwashita, 2005; Elder et al., 2002; Wigglesworth, 1997, 2000; Wigglesworth & Elder, 2010). The under-representation of the effects of different planning types could be that informing test candidates about such information is yet to be a common practice in the actual oral testing. In effect, examinees are simply instructed that they are given a certain amount of time before responding to a test question. Moreover, test preparation kits of the current major test publishers are less likely to provide concrete information on how to make use of the offered planning time.

Yet one could argue for the urgency of initiating this discussion of planning activities in testing considering the extremely limited amount of time allowed for planning in existing tests. Elder and Iwashita (2005) contended that the ineffectiveness of planning on test performance could be due to test candidates being (a) not familiar with such a short amount of planning time (hence, not reflecting the normal classroom-based oral tasks conditions) and (b) not aware of ways to efficiently *use* such time in test situations. The latter explanation prompted from the

researchers' finding of the participants responding favorably of the planning time yet failing to produce quality responses. Accordingly, the researchers speculated that "whatever action they were taking to improve their performance was ineffective" (Elder & Iwashita, 2005, p. 233) within the 3-minute time frame, with such ineffectiveness intensified with the lack of concrete guidance (i.e., explicit instruction before presenting the tasks) or training on planning strategies. By quoting Rutherford (2001), Elder and Iwashita expressed the importance of investigating how the provision of planning guidelines or strategy training makes a difference in test environments. Given the above results, it could be assumed that test candidates may struggle with even less planning time as in the case of the current oral testing such as the TOEFL iBT. If this is to be true, reviewing and exploring the effectiveness of the provision of guidance of planning is a timely matter.

Previous task-based research has devised a distinction between *guided* versus *unguided planning* with regards to the type of planning (e.g., Foster & Skehan, 1996; Gilabert, 2007; Kawauchi, 2005; Mochizuki & Orgeta, 2008; Sangarun, 2005; Wendel, 1997; Yuan & Ellis, 2003). According to Ellis (2005, 2009), learners in the unguided condition are simply allowed a certain amount of time before the task and asked to freely make use of this time to prepare for his or her responses; this mostly resembles the current planning practices in oral testing. On the other hand, learners are instructed to focus on specific aspects of planning under the guided condition; that is, learners are asked to engage in varied planning activities (e.g., taking notes, verbal rehearsal) prior to responding.

Whilst it is still empirically unanswered about the precise circumstances under which guided or unguided planning facilitate speech production (Ellis, 2009), the former condition is considered to be theoretically supported from Swain's (1993) Output Hypothesis. The premise is

that such pre-planning pushes learners to "make use of their resources" to "reflect on their output and consider ways of modifying it" (p. 161) prior to constructing their responses.

Thus, various types of guided planning have been devised by researchers yet of great relevance to the current study is that of Foster and Skehan (1996) and Kawauchi (2005). Although not conducted in a test environment, both studies had significance in testing out the intertwined effects of the provision of guidance in planning with other task-specific features. Foster and Skehan (1996), for instance, explored a possible interaction amongst variables such as planning time, guidance in planning, and task types. Three study groups each performed three task types under distinctive planning conditions: Group 1 was not given any planning time, group 2 were given 10 minutes of planning time, and group 3 were given guidance as to planning in addition to the 10 minutes of planning time. The guidance included "suggestions as to how language might be planned, and also suggestions to develop ideas relevant to completing each test tasks" (p. Skehan, 1998, p. 70). The study tasks were a personal information exchange task, a narrative task, and a decision-making task. The result was that speakers in Group 2 ([+ planning], [- guidance]) put forth more accurate speech for under unguided planning condition for the narrative task, while Group 3 ([+ planning], [+guidance]) speakers produced more subordination of language in speech for decision-making tasks. The researchers proposed two main interpretations of the result. First, when guidance is provided that directs speakers' attention to task content (or the message to be expressed), the result is the suffer of *accuracy* and increase in *complexity* of speech; on the other hand, when simply given time to plan, speakers prioritize to plan for the language they will use, hence resulting in enhanced *accuracy* of speech. Second, the effects of planning conditions are to be mediated by task characteristics: under unstructured, input-depleted tasks such as the decision-making task, speakers may primary direct their

attentional resources to generate elaborated idea, yet under narrative tasks in which reference information is provided, speakers would prioritize the form of language.

Kawauchi (2005) further employed planning activities that are not only recommended in the current test preparation materials (e.g., note taking, rehearsing), but also based on learners' actual pre-task planning strategies (Ortega, 1999). Employing a within-subjects design, Japanese EFL learners performed the same picture-description narrative tasks first without planning and subsequently with planning (with an interval in between for answering to a background questionnaire). Kawauchi employed three types of planning; namely, writing (i.e., taking notes of what they want to say), rehearsing (i.e., saying out loud), and reading (i.e., reading a model passage of the picture story). Each participant performed three sets of narrative tasks within a three-weeks window under differing planning condition (see Kawauchi, 2005, p. 149, for a detailed outline of the study procedure). Whilst Kawauchi did not find a favorable effect for one type of planning quantitatively, a qualitative analysis on the language use suggested the differing effect of guided planning activities in terms of learners' proficiency levels. Low-level learners benefitted mostly from reading, making use of the L2 input (vocabulary) under planned condition; on the other hand, high-level learners did not show a preference of one activity, but stated that they benefitted from all activities by means of organizing their thoughts.

1.2.3 Task characteristics mediating the effects of planning

Generally, task-based researchers found mixed results for the planning conditions as well as the provision of guidance in planning. Several studies contended the positive effects of both guided and unguided planning on *complexity* and *accuracy* (e.g., Foster & Skehan, 1996; Wendel, 1997) while others reporting favorable findings for one condition (e.g., Gilabert, 2007; Yuan & Ellis, 2003) or even neutral results (e.g., Kawauchi, 2005; Sangarun, 2005). Language

testing research, on the other hand, demonstrated rather consistently the limited effects of planning across different planning conditions. Alternatively, researchers have pointed out the mediating effects of certain task-specific features (e.g., narrative tasks versus decision-making tasks) to explain for such mixed findings (Elder et al., 2002; Foster & Skehan, 1996; Kawauchi, 2005; Tavakoli & Skehan, 2005). Foster and Skehan (1996) mentioned above is a precise example. This line of thought has been established by researchers drawing onto the earlier task-characteristics framework (Robinson, 2001; Skehan, 1998).

For instance, both Robinson (2001) and Skehan (1998) synthesized a number of task characteristics researched by the task-based researchers. These include a broader categorization of task features, but I list four categories that are of interest of the current study that concern on the manipulation of task information:

(1) Type of given information: Concrete/immediate versus abstract/remote information (Brown et al., 1984; Foster & Skehan, 1996; Robinson, 1995; Skehan & Foster, 1997)

(2) Organization of given information: Structured tasks versus unstructured tasks (Foster & Skehan, 1996; Skehan & Foster, 1997)

(3) Familiarity of information: Information with background knowledge versus unfamiliar information (Foster & Skehan, 1996; Robinson, 2001)

(4) Operations of information: Retrieval of given information versus transformation (manipulation) of given information (Foster & Skehan, 1996; Prabhu, 1987; Skehan & Foster, 1997)

The idea is that each parameter of task characteristics has differential impact on subsequent oral performance (more precisely, the CAF dimensions). For instance, speakers

benefit from concrete information given for task completion as it frees up information-processing load in planning, thereby releasing attention for *accuracy* and *fluency*; clear storyline presented in the provided pictures are one example. On the other hand, information abstract in nature impose on more processing load, as there are fewer external evidence for speakers to refer to; hence, *accuracy* of planned speech may be reduced while *complexity* gets enhanced. How the information is organized is also known to impact on oral performance. Tasks that are highly structured provide speakers with clearer macrostructure of the information, hence speeding up the conceptualization of task contents during planning. These may include information progressing in a sequential order (as in narrative texts) or in a salient pattern (as in argumentative texts that flows from introduction, main body, to conclusion). Speakers also benefit from task contents that they are familiar with during planning as they facilitate expedited processing and understanding of the given information while putting more conscious efforts in *accuracy* and *fluency*. Lastly, tasks requiring the manipulation of given material push speakers to try out more and use complex language (increased *complexity*). On the other hand, tasks asking for simple delivery or repetition of task contents may grant easy access to idea generation, hence reducing information-processing load during planning.

The presented task characteristics specific to the provision of test-task contents are of particular interest as they relate to the TOEFL iBT speaking test-tasks employed in the current study, and thus can shed light on ascertaining the influence of planning conditions in light of task-specific characteristics. This is because the primary difference between the independent and integrated test-tasks is the role of information processing of test takers; more precisely, the option of the provision of additional information (written and aural source material) in test-tasks that draws on test-takers' ability in not only speaking, but also reading and listening for

completing the task. Furthermore, given the focus of testing, studying the effects of task properties would be useful to be able to design tasks of predictable levels of difficulty which can be manipulated to elicit appropriate performances across candidates (Bachman, 2002; Wigglesworth & Foster, 2017).

To date, such an attempt to illuminate the impact of task characteristics has not been extensively demonstrated for oral task performance (Wigglesworth & Foster, 2017), particularly within the TOEFL iBT speaking contexts. Researchers thus far have explored the difference in task properties in independent and integrated tasks of the TOEFL iBT in terms of (1) test takers' strategic behaviors (e.g., Barkaoui, Brooks, Swain, & Lapkin, 2013; Huang & Hung, 2010; Hong, Huang, & Hung, 2016); (2) rater orientations to the test-tasks (Brown, Iwashita, & McNamara, 2005; Lee, 2006); and (3) the way in which test takers incorporate source materials into spoken performances (Brown et al., 2005; Crossley, Clevinger, & Kim, 2014; Frost, Elder, & Wigglesworth, 2012; Kyle, Crossley, & McNamara, 2015). However, none of the researchers of the studies listed here have focused on illuminating the role of planning in conjunctions to the TOEFL iBT test-task types for explaining a possible qualitative and/or quantitative difference in test performance; or more generally, the researchers indirectly explored the precise aspects of the test-tasks that inform about their inherent design. Given the varied time allotment of planning within and across test-task types in the TOEFL iBT, one can only speculate as to the rationale behind such task implementation condition, and in effect, whether they impact on subsequent test performance.

1.3 The study

In this present study, I take into account both task-based and language-testing research on

planning and task characteristics in the testing context of TOEFL iBT speaking. This research is needed because understanding what test takers are required to do, and what they are imposed to do through test-tasks fundamentally influence the nature of the performance obtained in performance-based oral testing (Skehan, 2016). Fundamentally, exploring planning and task-specific features of the TOEFL iBT could further address the construct validity of speaking tasks included in the test (Kyle et al., 2015).

I revisit the construct of planning in the context of the TOEFL iBT speaking test, with an additional interest in the interplay of planning and test-task characteristics. Specifically, I focus on the effects of the differing amounts of planning time provided across tasks as well as the provision of specific instructions (i.e., guidance) of planning. In so doing, I investigate the quantitative (test scores) and qualitative (discourse qualities) differences in test taker's speaking performance under specific task and planning conditions. I also explore test takers' use of and perceptions on the planning times employed across the two task types, and whether differences are contingent upon test tasks.

CHAPTER 2. THE CURRENT STUDY

2.1 Research questions and study variables

In this study, the unguided planning condition is equivalent to the kind of planning practiced in current academic oral testing contexts; its effects are compared to a condition where the test takers are given specific guidance on planning. This research inquiry is essentially an attempt to (a) document the actual planning actions and perceptions of test takers in this specific testing condition and (b) explore the optimal planning condition for facilitating academic speech performance within the current testing environment. Thus, to paint a comprehensive picture of the current planning practices, I investigate test takers' planning actions as well as perceptions. I also investigate the potential mediating effect of task characteristics of the planning conditions (Kawauchi, 2005; Wigglesworth & Elder, 2010). As the existent test procedure and its effectiveness on test performance is the subject of the current study, the results could be taken to understand the authenticity and validity of the corresponding oral tests. Most importantly, the findings could augment the ecological validity of the current planning conditions in language testing.

The following questions guided the current study:

1. Does the type of planning (guided versus unguided) affect the test scores of test candidates? If yes, what are the influences of the different task types?
2. Does the type of planning (guided versus unguided) affect the discourse quality of test candidates? If yes, what are the influences of the different task types?
3. How do test candidates use their planning times?
4. How do test candidates perceive the given planning times?

The variables of primary interest, therefore, are (a) the type of planning (guided versus

unguided) and (b) the type of guided planning (writing versus rehearsal). In terms of the type of planning, guided planning in the present study differs from the notion coined by Mochizuki and Ortega (2008) whose definitions refer to directing learners to attend to specific language aspects (e.g., "the syntax, lexis, content, and organization of what they would say", p. 307); instead, guided planning comprises interventions that suggest to the test takers in which activities to engage (Kawauchi, 2005). I chose two planning activities for the guided planning condition; namely, a regular time in which to plan, and a suggested plan to write during the same amount of time of planning. The latter could be considered *note taking* (Ortega, 1999; Wendel, 1997), requiring the act of writing or jotting down whatever comes into mind. The former is an act of thinking about one's ideas or sentence phrases that could be said in the actual response. Previous accounts have suggested that learners frequently engage in these activities when they are left to their own devices to use time for preparing for performing academically oriented oral tasks (e.g., presentation, role-plays) (Kawauchi, 2005; Ellis, 1987; Ortega, 1999). The difference in this study is that in the second condition with writing, test takers are encouraged and suggested explicitly to write, whereas in the first condition, they are not (though they may write if they would like, as is normal in current tests where there is time to plan).

An additional variable employed in the present study is a covariate: the test-task type. In particular, the task types are similar to those assessed on the TOEFL iBT speaking test; namely, the independent and the integrated tasks. Researchers and test developers have suggested that planning time is a major component in determining task difficulty, in that the varying planning conditions represent the items' different cognitive-demand levels (Norris, 2009; Robinson, 2009). In the TOEFL iBT speaking test, the lengths of planning time differ across task types. Researchers need to document the extent to which such varied time allotments on planning per

specific task type affects test takers' performance or perceptions on the planning conditions. In addition, such an analysis is essentially (although indirectly through the present study to some extent) a response to the call for more research on the test-task differences in the TOEFL iBT speaking test (Kyle et al., 2015).

Finally, I took into account participants' proficiency levels within each experimental group. In other words, each group consisted of a balanced number of high- and low-level learners. This grouping is primarily to illuminate a potential interaction effect of proficiency level and planning condition (Kawauchi, 2005).

2.2. Methods

2.2.1. Participants

Initially, a total of 120 college students participated in the beginning phase of the study. Yet due to various reasons (e.g., disqualification of participation, n=14; discontinuance in participation during the course of data collection, n=3; missing audio files, n=3), a total of ninety-nine participants remained for the final analysis.

In the summer of 2016, I recruited the participants from four large universities (University A, B, C, and D) located in Seoul, South Korea. I advertised the recruitment by posting an electronic version of the flyer on each university's online forums. Upon receiving contacts from the interested individuals, I only contacted those who met the qualifying criteria, which are as follows: (a) test takers with unexpired (effective) test scores from select standardized English language proficiency tests (e.g., IELTS, TOEIC, TOEIC Speaking, and TOEFL iBT); (b) speakers at intermediate-high to advanced in English proficiency as evidenced through their reported test scores (threshold score ranges: IELTS 6.5 to 7.5; TOEIC 880 and

above; TOEIC Speaking 130 to 180; TOEFL iBT: 87 to 110); (c) test takers with minimum test-taking experiences (took the test once or twice, at most) on the TOEFL iBT test; and (d) full-time university students. The reasons for establishing each criterion were the following: (a) to capture the participants' most up-to-date English language proficiency; (b) to ensure participants would be able to produce speech in English to the extent of yielding a meaningful amount of data for CAF analysis; (c) to ensure practice effects are minimal, due to the fact that the main instruments of the current study came from the TOEFL iBT practice test[1]; and (d) to control for the occupational status of the participants.

Among the four universities, University B was the only all-women's college while the others were all coed colleges. All of the participants were regularly enrolled college students in one of the universities: 36 students (18 males; 18 females) from University A, 33 students (all females) from University B, 20 students from University C (4 males; 16 females), and 10 students from University D (4 males; 6 females). The participants were primarily in their early- to mid-twenties (M = 24.39; SD = 2.26; *Max.* = 31; *Min* = 19), with more than half of the participants being seniors (N = 43) or beyond seniors (N = 17). A relatively small number of participants were in their early years in the college (Freshman: N = 5; Sophomore: N = 8; Junior: N = 18) or pursuing master's degree (N = 8). The participants came from a variety of academic backgrounds, yet approximately 58% of them were pursuing social sciences (N = 52); the

[1] Those with extensive test-taking experiences of the TOEFL iBT would have already internalized a certain degree of test-wiseness as to test tasks and testing conditions. This is particularly true in the Korean context where extensive test-taking experiences are inevitably related to the exposure to intensive test preparation practices. It is not an overstatement to say that those with the intent to prepare for this test begin their journey of test taking by taking test-preparation classes at private institutions, which are well known for its intensive training (coaching) of test-taking strategies (Choi, 2008). Planning strategies are one of which the programs are likely to instruct the students.

remaining disciplines included humanities and languages (N = 26), science (N = 15), and education (N = 6).

With the exception of 11 participants who indicated living in the English-speaking countries in their early years, most of the participants had learned English as a foreign language. They were relatively early learners of English (M = 7.92; SD = 2.34; $Max.$ = 15; $Min.$ = 6), having been exposed to English through both regular school instruction and private tutoring. When asked to self-assess their English language skills on a 6-point Likert scale (1 being very poor to 6 being advanced), the participants indicated higher confidence in their receptive skills (Reading: M = 4.96, SD = 0.72; Listening: M = 4.64, SD = 1.10) than in their productive skills (Speaking: M = 3.89, SD = 1.14; Writing: M = 3.98, SD = 1.14). Notably, the participants tended to give the lowest ratings to their speaking skills. Of the participants, 62 reported that they had studied a foreign language besides English. In terms of test-taking experiences, 22 participants once took the TOEFL iBT test, while vast majority of the participants took the TOEIC (N = 87) due to its significant use for employment and admission purposes in South Korea. The participants, in the end, were assumed to be quite homogenous in educational background, occupational status, age, and English-learning experiences.

I randomly assigned 33 participants to comprise each group, resulting in three experimental groups for the present study. Through the administration of a pre-test (which is described in detail in the following section), participants were further divided into two sub groups, each representing high and low proficiency of English (relative to each other). Participants performed three test sets under one of the three following conditions with corresponding instructions (adapted from the TOEFL official guide, ETS, 2012, and Ortega, 1999).

Unguided Planning (UG): In this condition, participants were left at their will to plan for their responses. They received the following instruction before responding: "Begin preparing your response after the beep."

Guided Planning-Writing (GW): In this condition, participants were provided with a blank sheet of paper, which they were required to use during planning time. Before planning, they received the following instruction: "On the provided piece of paper, please write out what you wish to say. It doesn't have to include everything in detail or in full sentences."

Guided Planning-Silently thinking (GT): In this condition, participants were asked to silently think what they would say for the actual response. They were not allowed to write or make notes during planning time. Before planning, they received the following instruction: "Please think silently inside your head about whatever you wish to say. You will not be allowed to take notes while you plan your responses."

2.2.2. Materials

2.2.2.1. Background questionnaire

I administered a pre-experiment questionnaire (written in Korean) to elicit participants' language learning background and test-taking experiences prior to their participation in the main study (see Appendix A for a complete list of questions). I administered the survey online by using Qualtrics (https://www.qualitrics.com). This was to additionally screen participants with extensive test-taking experiences on the TOEFL iBT test.

2.2.2.2. Elicited imitation task

I used an elicited imitation (EI) task (Ortega, Iwashita, Norris, & Rabie, 2002) (see Appendix B for the practice and main test sentences) as a pre-measure for three purposes: (a) for ensuring homogeneousness in grouping participants into the current study's experimental

conditions; (b) for dividing the participants into high and low proficiency levels in English; and (c) for minimizing participants with the exposure to any type of exposure to typical oral-testing conditions, EI was used instead of implementing a different type of speaking test. The decision to use EI was that not only it is an authentic task that is reflective of everyday conversation, but also it has been attested as a global measurement of oral proficiency in the L2 (Tracy-Ventura, McManus, Norris, & Ortega, 2013). More specifically, researchers have favored EI for directing learners to attend to the form and meaning of the target structure while requiring both comprehension and production (Bowles, 2011; Cox, Bown, & Burdis, 2015; Van Moere, 2012). According to the EI test results (discussed further in the next section), I balanced each testing condition in terms of the same average score and standard deviation.

In the current EI task, the participants were asked to first listen to English sentences (that is read once) and then verbally repeat them. The full test consisted of 30 test sentences, preceded by six practice sentences for the purpose of familiarizing participants with the test procedures. These practice sentences were given in the participants' L1, Korean. Participants' responses were recorded by Audacity (http://web.audacityteam.org).

2.2.2.3. Test tasks

The main oral test tasks in the present study come from the three practice test sets (Set A, B, and C) from the *TOEFL iBT*® official guide, which are electronically accessible (ETS, 2012). From each test set, I further selected three test tasks, which are as follows: (a) independent task (providing one's opinions about a given topic) (henceforth IP task), (b) integrated task with listening and reading (e.g., providing responses by combining relevant information from reading and listening to two sources) (henceforth IT-RL task), and (c) integrated task with listening only (henceforth IT-L task) (see Appendix C for complete sets of questions). I chose to use sample

TOEFL iBT test items and the specific test task types that I did for their academic orientation of speaking skills as well as the planning conditions of interest. Planning times differed across tasks: (a) 15 seconds for IP task, (b) 30 seconds for IT-RL task, and (c) 20 seconds for IT-L task. The response time for IP task was 30 seconds, while for IT-RL task and IT-L tasks, 60 seconds were given.

The electronic version of the practice sets simulates the actual testing screen of the TOEFL iBT speaking section. Yet to enable a natural flow of the experimental procedures of the current study (which is described in the next section below), I used Camtasia (www.camtasia.com) to make screen-captured video clips of each test task. I then uploaded the nine video clips (3 test tasks times 3 test sets) online via Youtube (www.youtube.com) due to its compatibility with Qualtrics (www.qualtrics.com), an online platform used to construct surveys and assessment tools. In Qualtrics, I constructed a web-based assessment on which participants were able to read the guided planning instructions, respond to the test tasks, and take the post-survey questions at their own pace. To respond to the test tasks, the participants were instructed to press the play button on the test-task videos. In each video, a speaker guided the participants throughout the test procedures. The speaker gave general instructions of the corresponding test task, which were simultaneously displayed in a written form on the test screen. The speaker then prompted participants to read the test question, which was followed by an indication about the amount of planning and responding time for the test task at hand. Subsequently, a time-running bar on the bottom side of the test screen was immediately activated, while a countdown timer appeared to display the remaining time. After the planning time was completed, the response time was given. The guiding speaker instructed the test takers when to begin planning and responding.

2.2.2.4. Post questionnaire

I gave a post questionnaire (written in Korean) using Qualtrics (that appeared immediately after the last test task in each test set) to gauge the participants' perceptions of the current planning practice (see Appendix D). The questions (adapted from Rutherford, 2001; Wigglesworth & Elder, 2010) particularly elicited (a) participants' self-assessment on how they performed on each test task; (b) participants' perceptions on the effectiveness of the length of planning time provided (15 seconds, 20 seconds, and 30 seconds) for each test task; (c) participants' engagement of other types of planning activities (a question asked to those who reported that they did not make extensive use of the examined planning activity in each planning condition) and (c) participants' evaluation of the effectiveness of the type of planning (guided versus unguided).

2.2.2.5. Interview

To further probe into the participants' perspectives on the effects of planning, I conducted a brief interview with participants after testing. The participants were able to respond in Korean or English. I adapted the interview questions from Wigglesworth and Elder (2010) (see Appendix E). Questions concerned (a) the appropriateness of the planning conditions (in terms of the length of time and type of planning), (b) the use of the planning time in the differing planning conditions as well as differing test tasks, (c) the strengths and weaknesses of the planning activities, and (d) further suggestions, if any, of improving the current planning practice.

2.2.3. Study Design

As seen in Table 1, the order of the three planning conditions (UG, GW, GT), test sets (Set A, B, and C), and test tasks (IP task, IT-RL task, and IT-L task) were counterbalanced with a Latin Square design. For instance, a group of participants employed a certain planning activity

for a test set while the other two groups of participants took the remaining test sets, each under different planning condition. In the end, all test tasks were performed under all three planning conditions. Such a study design controls the possible intervening effects of the sequences of the planning conditions, test sets, and test tasks (Wigglesworth & Elder, 2010).

Table 1

Research Design

Groups	Session 1	Session 2	Session 3
Group 1 (N = 30)	UN-Set A	GW-Set B	GS-Set C
	IP task	IP task	IP task
	IT-RL task	IT-RL task	IT-RL task
	IT-L task	IT-L task	IT-L task
Group 2 (N = 30)	GS-Set B	UN-Set C	GW-Set A
	IT-L task	IT-L task	IT-L task
	IP task	IP task	IP task
	IT-RL task	IT-RL task	IT-RL task
Group 3 (N = 30)	GW-Set C	GS-Set A	UN-Set B
	IT-RL task	IT-RL task	IT-RL task
	IT-L task	IT-L task	IT-L task
	IP task	IP task	IP task

Note. UN, GW, and GS each indicate Unguided Planning, Guided Planning-Writing, and Guided Planning-Silently thinking, respectively. IP, IT-RL, and IT-L each indicate Independent, Integrated-Reading and Listening, and Integrated-Listening only tasks, respectively.

2.2.4. Procedure

Prior to proceeding to data collection, I conducted a pilot study with 25 college students (with varying English proficiency levels) in South Korea. I tested out the testing procedures as well as the post-survey items. In so doing, I replaced Kawauchi's (2005) *Speaking out loud* condition to the regular *timed planning* condition. This was due to the fact that the former condition was practically difficult for the majority of learners to carry out in the actual testing context; in other words, the act of verbally rehearsing within a brief amount of planning time appeared to be disruptive to them in organizing their thoughts and preparing their responses. Silent planning time was a condition that most of the students reported as easy in which to engage during cognitively demanding situations such as test taking.

For the main experiment, I sent out a link to the online pre-experiment questionnaire to the students who were eligible to participate. I then randomly assigned each participant one of the three study groups. Upon assignment, I asked each participant to schedule themselves for three separate testing days (with at least a one-day interval in between; consecutive test dates were avoided to prevent immediate practice effects).

I asked the participants to attend three testing days in a private study room I reserved at each of the four universities. In all sessions, I had them sit in front of a laptop computer, and I provided them a headset for listening and speaking. As soon as I gave them brief instructions about the procedures in general, they were left alone in the room. On the first testing day, participants first signed the consent form. They then took the EI test for 20 minutes and moved to the first test set corresponding to their assigned groups. In the second testing day, the participants came in to take the second test set, followed by the post-questionnaire. On the third testing day, the participants took a third test set and partook in the post-hoc interview session. As in

operational testing, participants were able to make notes during the test-taking of integrated tasks. These notes were differentiated with the written planning participants produced in the guided-writing planning condition. In the unguided planning condition, participants were instructed to use the time at their will. All participants were monetarily compensated ($30 worth of a gift card) for their time.

2.3. Measures

Following the previous literature on planning time, I analyzed each participant's responses in regard to the discourse quality measures of *complexity*, *accuracy*, and *fluency* (henceforth CAF measures) (e.g., Elder & Iwashita, 2005; Kawauchi, 2005; Wigglesworth, 1997; Wigglesworth & Elder, 2010).

The three quality measures were further broken down to include specific features. In the following section, I provide the definitions and subcomponents related to each quality measure.

2.3.1. Fluency measures

Following previous researchers (Kawauchi, 2005; Housen & Kuiken, 2009; Skehan, 2009; Wigglesworth & Elder, 2010), I conformed to Segalowitz (2010) in conceptualizing *fluency* as subscribing to objective acoustic measures of an utterance (see Segalowitz, 2010, on a broader definition of L2 fluency that additionally includes *cognitive* and *perceived fluency*). More specifically, I defined *fluency* as three subdimensions; namely, *speed fluency*, *breakdown fluency*, and *repair fluency*.

First, I indexed *speed fluency* through (1) identifying the total number of syllables produced in each response per minute by using the online software, Syllable Counter (SyllableCount.com, n.d.); (2) calculating the *speech rate* of all responses (Freed, 2000) by

dividing the total number of produced syllables in a given speech by the total amount of utterance time (Tavakoli & Skehan, 2005); and (3) identifying the silent duration, or the *time spent before articulation* each participant took before actually articulating his or her speech. I initially derived *time spent before articulation* to better account for the actual phonation time to calculate *speech rate*, which makes the measure to only encompass the production time. Yet I decided to include it as a separate *speed fluency* measure after inspecting participants' interview responses. As will be reported in the next chapter, a majority of participants consistently indicated in the interview a possible relationship between fluidly delivering a speech and the silent buffering time taken before articulation. It was seemingly plausible that such a buffering time could have influenced how the subsequent speech was delivered. I defined the measure as the amount of time of a silent duration between the initiating beep sound (that prompted participants to speak) and the first articulation point.

After data collection, I applied the following specific criteria to qualify the data: (1) *silent pauses* pertained to the inaudible sounds that were equivalent to or longer than 250 milliseconds (see De Jong, Groenhout, Schoonen, & Hulstijn, 2013 for the discussion of a threshold value for identifying silent pauses); (2) *the total number of syllables* only included intelligible language; that is, I excluded filled pauses (*uh*s, and *um*s) as well as disfluency features (e.g., repetitions, hesitations, false starts marked with hyphens in the transcriptions) when counting the number of syllables in the entire speech; (3) *speech rate* pertained to actual phonation time, excluding the silent *time spent before articulation*; and (4) *time spent before articulation* excluded any verbal articulation of produced by participants; for instance, participants' production of lengthened filled pauses (e.g., *uhhhh*, *ummm*) was not included in the measure. Table 2 presents the measures used for indexing *speed fluency.*

Table 2

Speed fluency measures

Measures	Description	Values
Tot. Num. of syllables per minute	Total number of syllables in each response	Quantified value
Speech rate	Number of syllables / Total duration of utterance (excluding *time spent before articulation*)	Quantified value
Time spent before articulation	Total duration of silence before being prompted to speak	Quantified value (in seconds)

Second, I identified *breakdown fluency* by: (1) calculating the number of filled and unfilled pauses per minute (Elder & Iwashita, 2005; Kormos, 2006) and (2) mean length of run (De Jong et al., 2012). Table 3 summarizes the *breakdown fluency* measures employed in the present study.

Table 3

Breakdown fluency measures

Measures	Description	Values
Num. of filled pauses per minute	(Total number of filled pauses / Duration of utterance) * 60	Quantified value

Table 3 (cont'd)

Num. of unfilled pauses per minute	(Total number of unfilled pauses / Duration of utterance) * 60	Quantified value
Mean length of run	Total number of syllables / Total number of unfilled pauses +1	Quantified value

Finally, for *repair fluency*, I identified the number of disfluency features, particularly those representing *repetitions*, *replacements*, *reformulations*, *hesitations*, and *false starts* (Bygate, 1996; Foster & Skehan, 1996; Kormos, 2006; Riggenbach, 1991). I classified these qualitative features according to Bygate's (1996) sub-dimensions: *verbatim repetition* (i.e., "occurs when hesitating, creating time to find an appropriate word", p. 141) and *substitutive repetition* ("employed when correcting a word or grammatical feature", p. 141). According to these conceptualizations, I included *repetitions* and *replacements* in the former, and *reformulations*, *hesitations*, and *false starts* in the latter component. Table 4 depicts the *repair fluency* measures with precise definitions and examples from the data.

Table 4

Repair fluency measures

Measures	Sub-category	Description	Example from the data	Values
Verbatim repetition	Repetitions	Words, phrases, or clauses that are repeated with no modification whatsoever to syntax, morphology, or word order	*Uh, for this* ***problem-problem***	Quantified value
	Replacements	Lexical items that are immediately substituted for another	*the* ***awa- ceremony*** (the award -> the ceremony)	
Substitutive repetition	Reformulations	Phrases or clauses that are repeated with some modification to syntax, morphology, or word order	*I believe that working together* ***it- this*** *makes synergy*	
	Hesitations	Initial phoneme or syllable(s) uttered one or more times before the complete word is spoken	*to write a paper for* ***inste- instead***	
	False starts	False starts are utterances that are abandoned before completion and that may or may not be followed by a reformulation	*So, if she, if student considered, so,* ***she think that um*** (*if clause* abandoned and not repeated before being modified to SV structure)	

As in Kawauchi (2005), I explored the proportion of *repair fluency* measure per performance for a task to account for the varied amount of speech from individual participants. This was accomplished by dividing the total number of each *repair fluency* measure by the total duration of speech time.

2.3.2. Complexity measures

In this study, I followed Kawauchi (2005) in analyzing *syntactic complexity*, and I also followed Kawauchi by using *lexical diversity* for measuring *complexity*.

I analyzed *syntactic complexity* by exploring the following measures: Analysis of Speech Unit (AS-unit) and subordinate clauses (Ferrari, 2012; Frost, Elder, & Wigglesworth, 2011). An 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, p. 365). While an AS-unit is a syntactic unit encompassing the definition of a typical T-unit (see Hunt, 1965), it differs from the T-units by taking into account the specific features of spoken data. According to Foster et al. (2000), AS-units reflect such a nature of oral speech by (a) including the independent sub-clausal units that are very common in speech, and are defined as minor utterances (i.e., irregular sentences or non-sentences; *yes*, *thank you very much*, *oh how wonderful*, p. 366); (b) including *breakdown fluency* measures such as *false starts, repetitions,* and *reformulations* as the same unit; and (c) using intonation and pausing phenomena to distinguish syntactic boundaries. In terms of the second point, such principled use of temporal features applies to specific cases of discerning whether coordinated phrases are included in the previously occurring AS-unit. In the case of T-units, coordinated phrases are generally considered as being included as the same T-unit (Hunt, 1965). However, coordinated units can be independent AS-units when a pause longer than 0.5 seconds is preceding and the

conjunctive markers (e.g., *and*, *but*) themselves are articulated either in falling or rising intonation (Foster et al., 2000). Foster et al. (2000) asserted that AS-units were also useful for analyzing monologic discourse (i.e., non-interactive speech) as t-units are, thus I used AS-units for analyzing discourse in the current study.

In addition to the broader syntactic boundary, I explored the number of *subordinate clauses*. Following Foster et al. (2000), I defined an independent clause as a clause including a finite verb, whereas a subordinate clause is "a clause consisting of a finite or nonfinite verbal element with at least one other clausal element such as a subject, object, complement, or adverbial" (p. 365).

Taken altogether, the two syntactic units were meant to quantify the participants' abilities in producing both macro- and micro-units for establishing a more complex message. In the end, with these indices, I used two global quantitative measures of *syntactic complexity* as follows: (1) AS-unit length to account for the density of the syntactic unit (Ferrari, 2012) by dividing the total number of produced syllables by the identified number of AS-units (Bulté & Housen, 2012); and (2) the number of subordinate clause to AS-unit for indication of subordination (Foster et al., 2000). See Table 5 for the summary of the independent measures for *syntactic complexity*.

The following excerpt displays the identification of AS-units and subordinate clauses. Double forward slashes (//) were placed at AS-unit boundaries while double colons (::) indicate the locations of subordinate clauses. Numbers in parenthesis are the duration of an unfilled (silent) pause. Note that the coordinated phrase starting with "so" in the last sentence of the excerpt (bolded segment) was treated as a separate AS-unit as it was preceded by a long silent pause (of 1.14 seconds) and was articulated with a rising tone.

Um, I often go to the baseball stadium with my friends :: because I like to watch Korean baseball game. // It's my like, kind of like ori- original life. // So, I always watch the movie :: even though I can't go to the stadium but I try to go to the stadium :: when I have time. ***(1.14)*** ***// So*** *I, I meet my friends and we, enthusiastically, um, enjoy the game. //*

For *lexical diversity* (see Table 5), I used the Tool for the Automatic Analysis of Cohesion (TAACO; Crossley, Kyle, & McNamara, 2016) for analyzing a number of indices denoting to lexical sophistication and complexity. TAACO (available here: http://www.kristopherkyle.com/taaco.html) is an automated text analysis, which specifically concerns on measures of lexical cohesion at word, sentence, and paragraph level. From TAACO, I identified both global and refined measures of lexical diversity. First, I explored *type-token ratio* (TTR), which is the ratio in percent between the different lexemes in the text (e.g., nouns, pronouns, verbs, adjectives, articles, adverbs, prepositions, and conjunctions) and the total number of lexemes (Laufer, 1991; Laufer & Nation, 1995; Ortega, 1999; Robinson, 2001). Yet supplemental information to TTR is needed as the value itself is strongly affected by the text length (Skehan, 2009); in other words, TTR could yield unstable results for a shorter amount of discourse (as is expected to be elicited from the task types employed in the current study) as opposed to longer stretch of utterances (Laufer & Nation, 1995; Malvern & Richards, 2002). Therefore, I looked at the number of unique words (i.e., types) as well as the cohesive devices used in the speech samples such as *conjunctions* (*and*, *but*) and *sentence linking words* (e.g., *nonetheless, therefore, although*).

Table 5

Complexity measures

Measures	Sub-category	Description	Values
Syntactic complexity	Number of AS-units	The total number of AS-units in each utterance	Frequency counts
	Number of subordinate clauses	The total number of subordinate clauses in each utterance	
	AS-unit length	Total number of AS-units / Total number of syllables	Quantified value
	Subordinate clause ratio	Total number of subordinate clauses / Total number of clauses	
Lexical diversity	Number of word types	Unique lexical words in each utterance	Frequency counts
	Type-token ratio	Total number of unique words (types) divided by the number of total words (tokens)	Quantified value
	Cohesive devices	Total number of conjunctions and sentence linking words / Total number of words	

2.3.3 Accuracy measures

I measured overall accuracy in terms of employing a broader accuracy measure of *error-free clauses* and a refined index of *lexical errors* (Kuiken & Vedder, 2012). Adopting the latter element complemented the incomplete nature of errors represented by the global accuracy measures that are associated with syntactic units (e.g., the number of error-free units, the number of errors per units) (Kuiken & Vedder, 2012). Following Tavakoli and Skehan (2005), I defined error-free clauses as those with no grammatical errors in terms of syntax, morphology, and word order. To be more specific, for a clause to be error free, clauses should contain correctly used verb forms (e.g., tense, aspect, voice, modality, and subject-verb agreement; Wendel, 1997), articles, and plural –s. I did not include errors in stress, intonation patterns or pronunciation of words and utterances. *Lexical errors*, on the other hand, subscribed to deviations in pronunciation, meaning, grammatical form, word order, collocation, idioms, and awkward phrasing that may interfere with the overall *comprehensibility* of the speech (e.g., It could *happen* a lot of troubles; words such as *trigger* or *cause* are more appropriate) (Mehnert, 1998).

With these two indices, I calculated two quantitative measures of *accuracy* as follows: (1) the number of error-free clauses to total number of clauses (Tavakoli & Skehan, 2005); and (2) lexical errors per 100 words (Mehnert, 1998). I used the latter measure for normalizing the raw error counts. This was to account for the impact of speech lengths on skewing the occurrences of errors (a higher word count creates more opportunities for errors; Plakan & Gebril, 2016). Table 6 reports further examples from the data coded for the two accuracy measures.

Table 6

Accuracy measures

Measures	Description	Values
Error-free clauses ratio	Total number of error-free clauses / Total number of clauses	Quantified value
Lexical errors per 100 words	(Total number of lexical errors / Total number of syllables) *100	

2.4 Analysis

2.4.1 Research question 1: Does the type of planning (guided versus unguided) affect the test scores of test candidates? If yes, what are the influences of the different task types?

The first research question pertains to whether there are effects of the type of planning (guided versus unguided) in conjunction to the test-task type on test taker's test performance. In the following section, I describe the procedures and statistical analysis taken to address this inquiry.

2.4.1.1 Subjective ratings of the Elicited Imitation task responses

Prior to answering research question 1, I examined participants' EI test performance for the purpose of having two sub-groups within each study group that represent high and low oral proficiency. I hired a female rater (an M.A. student in TESOL at a large Mid-western university) to evaluate participants' EI test performance. I participated as the second rater, to establish inter-rater reliability with the first rater. The first rater and I met for multiple sessions to train

ourselves on the EI rubric (Park, 2015; Ortega et al., 2002; see Appendix B). After we reached consensus in our ratings on the sample items, we each scored all 99 participants' responses. Based on the rubric, we scored each response on a 4-point scale. We met multiple times during the course of our ratings to discuss any discrepancies and concerns we had when rating particular responses. In the end, our inter-rater agreement on all 30 EI items for all participants was estimated at .88 (Cronbach's alpha), which was interpreted as moderately high. In addition, the EI test appeared to have high overall reliability, which was estimated at .93 (Cronbach's alpha).

It should be noted that the scoring method of the EI responses employed in the current study (based on Ortega et al., 2002) differed from the conventional approaches that primarily concern the completion and/or grammatical accuracy of repetitions (e.g., Erlam, 2006; West, 2012). Instead, the responses were evaluated for its retention of key idea units in addition to the language forms (see Appendix B). Ortega and her colleagues' rationale of such evaluation comes from the evidence that after listening to a sentence, the utterance's meaning is stored for a significantly longer time than its linguistic form and specific wording (Sachs, 1967); hence, assessment of the responses becomes important to take into account the degree to which speakers are able to maintain both form *and* meaning of an input sentence (Wu & Ortega, 2013). Aside from this, integrity of the meaning in the repetition was deemed important as to its potential link to the construct measured by the TOEBL iBT speaking.

Table 7 reports on the average scores between the first rater (rater 1) and I (henceforth rater 2). Although the mean score for Group 1 participants was the lowest (from both raters), a one-way ANOVA on the average mean scores of rater 1 and rater 2 (see the fifth column in Table 7) indicated statistically insignificant difference among the three groups ($F = 0.147$, $p = .086$, $\eta p^2 = 0.21$). The results suggest that participants in each group were comparable in terms

of the oral proficiency measured through the EI test.

Table 7

Descriptive statistics for the EI TEST ratings given by rater 1 and rater 2

Group	N	Rater 1	Rater 2	Average Mean
		M (SD)	M (SD)	(Rater 1 + Rater 2)
Group 1	33	81.30 (15.61)	81.04 (14.20)	80.67 (14.91)
Group 2	33	82.30 (18.33)	82.04 (17.66)	82.17 (18.00)
Group 3	33	82.70 (11.94)	82.51 (12.32)	82.60 (12.13)

Note. Maximum possible total score a participant can receive is 120 (30 items * highest rating of 4).

Following Ortega (2000), I averaged the three combined mean scores from rater 1 and rater 2 (i.e., scores reported in the fifth column of Table 7). I then derived an arbitrary cut-off point of 81.8 as a benchmark for dividing participants into two subgroups (low and high proficiency). Those who scored above 81.8 on the EI test were considered as speakers with relatively higher oral proficiency than the lower group.

As in Table 8, the two sub-groups within each of the three experiment groups were all balanced in number. In addition, the mean EI test scores of all High-Proficiency (henceforth HP) and Low-Proficiency (henceforth LP) groups were comparable across the three experiment groups.

Table 8

Descriptive statistics of the EI test according to the proficiency sub-groups

Group	Sub-group	N	Averaged descriptive		
			M (SD)	Max.	Min.
Group 1	High	16	91.24 (7.91)	109	84
	Low	17	65.69 (7.87)	77	52
Group 2	High	17	93.65 (7.55)	105	82
	Low	16	66.69 (10.76)	78	46
Group 3	High	16	92.94 (8.73)	110	83
	Low	17	69.82 (9.86)	79	40

Two separate one-way ANOVAs conducted once with the three HP groups and once with the three LP groups confirmed that participants in each proficiency band performed similarly (HP: $F(2, 47) = 0.400$, $p = .673$, $\eta p^2 = 0.36$; LP: $F(2, 46) = 0.841$, $p = .438$, $\eta p^2 = 0.58$). From these results, it can be inferred that the distribution of participants (at least with respect to their EI test scores) were relatively even across the three experiment groups. In addition, a series of three independent *t* test revealed that the magnitude of difference between the HP and LP participants within each experiment group was significant (Group 1: $t = 9.261$, $p < .001$; Group 2: $t = 8.374$, $p < .001$; Group 3: $t = 7.114$, $p < .001$).

2.4.1.2 Subjective ratings on spoken responses

To address research question 1, I hired three raters to score the speech samples according to the TOEFL iBT speaking rubric (see Appendix F). These individuals (two males and one female) were all native speakers of English, who, at the time of scoring, were pursuing their

master's degrees in Linguistics and TESOL at a large Mid-western university. While all three raters had varying levels of experience in teaching English, they were novice to rating speech samples produced in the language assessment context. As such, prior to moving onto the main phase of rating, they were asked to receive an intensive training session on rating from a rater-training expert affiliated to the same university. The rater-trainer had extensive experiences on rating as well as training raters for high-stakes language proficiency tests. Prior to the training session, the rater-trainer took 30 speech samples from the current dataset to derive benchmark speech samples (with respect to the TOEFL iBT Speaking rubric) and to categorize characteristics of the oral responses pertaining to a specific score band. The actual training session was held for about two hours, during which the raters accustomed themselves with the rubric as well as the benchmark responses. There was an additional intervention session during the individual rating period in which the rater trainer and all three raters came together to discuss possible concerns and inquiries in rating.

In terms of the rating design, I adopted an incomplete, connected block design of rating (Eckes, 2009; Fleiss, 1981). From this approach, raters do not score each and every speech response, but as pairs, they jointly rate a fixed number of responses (Fleiss, 1981). As opposed to the fully crossed, complete block designs (i.e., every rater rates every test response), incomplete block designs have been commonly employed in research studies and large-scale rating projects due to their cost- and time-efficient nature (Myford & Wolfe, 2004). I employed the incomplete design precisely for such a reason, while I had to acknowledge the existence of a sparse data set with missing observation (Myford & Wolfe, 2004). To minimize such a limitation of the rating design, researchers have suggested randomizing the assignment of participants (or observations) to the raters (Bechger, Maris, & Hsiao, 2010; Eckes, 2009). Following this suggestion, I devised

a partially randomized rating assignment as displayed in Table 9. More specifically, I divided the speech samples into three large sets (each consisted of 33 participants' speech samples). I randomized the order of participants as well as the test-task orders (IP task, 2, and 3) in each data set, and took care to balance out the number of participants from each experiment group within each set. The order of raters (rater A, B, and C) were balanced across each set of speech samples. Furthermore, I ensured the rating design adopts *connectedness* amongst the individual raters as well as test takers. A connected design is critical in performing Rasch modeling (which will be dealt further in the next section) as it fulfills the unidimensionality of Rasch modeling (Linacre & Wright, 2002); in other words, it makes it possible to calibrate all measurement factors (e.g., planning conditions, test-task types) onto the same scale in terms of score variation (Rasch modeling will be dealt in the next section in more detail) (Eckes, 2009). As shown in Table 9, individual raters are linked to one another (e.g., Raters A and B) through common ratings of the same examinees (e.g., Participant D), while each examinee (e.g., Participants D and E) is linked to one another through common ratings by at least a pair of raters (e.g., Raters A and B).

As stated above, the raters scored using the TOEFL iBT Speaking rubric (https://www.ets.org/s/toefl/pdf/toefl_speaking_rubrics.pdf). The rubric is based on a holistic scale, which specifically focused on three descriptors: *delivery* (i.e., fluidity of the response), *language use* (i.e., complexity of the language), and *topic development* (i.e., coherence and relevance of the ideas to the topic). The TOEFL iBT Speaking rubric has all responses rated on a 4-point scale (i.e., possible maximum score: 4).

Table 9

Partially-balanced, incomplete rating block design

Set 1	Task	Rater A	Rater B	Rater C
Participant D	IP task	X	X	
	IT-RL task		X	X
	IT-L task	X		X
Participant E	IT-RL task	X	X	
	IT-L task		X	X
	IP task	X		X
Participant F	IT-L task	X	X	
	IP task		X	X
	IT-RL task	X		X
…	…		…	
Set 2	Task	Rater B	Rater C	Rater A
Participant G		X	X	
Participant H		X		X
Participant I			X	X
…	…		…	
Set 3	Task	Rater C	Rater A	Rater B
Participant J		X		X
Participant K			X	X
Participant L		X	X	
…	…			

2.4.1.3 Subjective ratings on language use of spoken responses

Given the differential task conditions among the three test tasks, I collected additional

information to enlighten a possible task effect and its interaction with the planning conditions under which a specific task was performed. More specifically, this analysis concerned how the raters perceived the extent to which participants benefitted from a particular task type or not.

The same three raters were asked to rate on a 4-point scale on how much of the specific language features (e.g., phrasal verbs, collocations, and vocabulary) in the input sources (e.g., reading texts in IT-RL task, and listening dialogue in IT-L task) were directly utilized in each response. With regard to the 4-point scale, 1 indicated a complete originality of the response, while 4 represented that a good amount of the language was extracted (or repeated) from the sources.

2.4.1.4 Statistical analysis

I carried out Multi-Faceted Rasch Measurement (MFRM) (Linacre, 1989) to discern how raters' subjective scorings differed according to the testing variables at interest. I used FACETS 3.80.3 (Linacre, 2017) for the MFRM analyses. While testing programs and researchers have commonly utilized the approach for assessing the severity of the raters and to establish test quality control (Bachman, 2004; Eckes, 2009; Weigle, 1998), the emphasis in the present study was placed on observing the impact of diverse measurement factors (e.g., raters, examinees, test tasks), or also known as *facets*, on quantitative test outcomes (Bonk & Ockey, 2003; Papageorgiou, Stevens, & Goodwin, 2012). With the MFRM model, the parameters of each facet can be estimated independently of the rest of the facets and are calibrated onto a single linear scale (i.e., the logit scale) (Myford & Wolfe, 2004). This makes it possible to carry out comparative interpretations among specified facets and the additive contribution of each indicator to score variation. Among a number of facets, I put a primary emphasis on the planning conditions and test-task types employed in the present study. Thus, I utilized MFRM modeling in

response to the question, "To what extent do the core facets of interest – the three planning conditions and the test-task types – modeled in the current testing condition contribute to score variance?"

MFRM modeling using FACETS was also critical for the current analyses for its robustness to model missing data (Bond & Fox, 2007; Eckes, 2009). FACETS accounts for these missing observations through the Joint Maximum Likelihood Estimation (JMLE) (Fisher, 1922), which is an iterative procedure taken in FACETS to calculate the estimates of each facet and its element defined in the analysis. JMLE operates under almost all conditions and overrides an incomplete data set for making estimations. This is because estimates are only based on the active data, or the data that has been observed (Eckes, 2009; Linacre, 2017). Any extent of randomness exerted in the data by missing observations is treated as "well-behaved" (Linacre, 2017, p. 15). In the end, as long as the data set sufficiently fits the specified Rasch model, there is no bias in the produced measure estimates and fit statistics caused from the presence of missing data. This feature of FACETS and MFRM analysis carried significance in the current analyses; due to the incomplete block rating design incorporated in the present study, test takers did not receive full ratings from all three raters (see Table 9). Therefore, in the current study's data file prepared for the MFRM analyses, I coded the missing observations as "m" and identified the code in FACET's model specification. Entering the code for missing data helped the corresponding data points to be ignored and bypassed in the estimation.

Following Bonk and Ockey (2003), I conducted three separate MRFM analyses in FACETS with regard to three administrations of test sets (Test Set A, B, and C). For all three analyses, I used a Rating Scale Model (RSM) as opposed to a Partial Credit Model (PCM) owing to the fact that the same holistic rating scale (e.g., the TOEFL iBT Speaking rubric) was used

across the board for rating on each task (Winke, Gass, & Myford, 2012). In the present study, I specified six facets and the corresponding elements to define each facet (see Table 10).

Table 10

Six facets specified for the MFRM analyses

Facets	Elements	Description
Raters	3	Rater A, B, and C
Examinees	98 for Set A 99 for Set B 97 for set C	Varied by the number of participants with complete responses for all three test tasks in the test sets
Participant groups	3	Group 1, 2, and 3
Participants' oral proficiency level	2	High and low oral proficiency as determined by the EI scores
Planning conditions	3	Unguided (UG) Guided-Writing (GW) Guided-Thinking (GS)
Test-task types	3	IP (Independent) IT-RL (Integrated-Reading & Listening) IT-L (Integrated-Listening)

In all three analyses, I specified the following MFRM model:

?, ?, ?, ?, ?, ?, R4

where *?* each indicates the six facets of interest (raters, examinees, participant groups,

participants' EI-based oral proficiency levels, planning conditions, and test-task types), and *R4* indicates the highest possible score awarded to a test taker is 4 on the current rating scale. Each *?* controls the selection of elements in each facet (first, second and so on). Specifying *R4* controls for data-entry errors in that any data points greater than 4 would not be treated as valid data for estimation. Overall, the model "?, ?, ?, ?, ?, ?, R4" means: any element of *facet 1* adds to any element of *facet 2* adds to any element of *facet 3* adds to any element of *facet 4* adds to any element of *facet 5* adds to any element of *facet 6* producing an observation on a rating scale whose highest category is 4 or less.

In the end, I made use of a number of outputs from FACETS in the current analyses. These included: (1) estimates from the fit statistics (e.g., infit and outfit mean square values) for assessing the overall fit of the data to the Rasch model; (2) reliability of separation index for assessing the extent to which the elements specified in each facet are separated (e.g., how distant are test scores awarded in a UG condition from other planning conditions); and (3) the Wright map for visual inspection of the data.

As a follow-up on the MFRM analyses, I conducted a series of descriptive as well as inferential statistics for the test scores across the group. The primary dependent variable was the total test scores of each participant across the nine test tasks from the three test sets, which ranged from 0 to 33 (maximum of 12 points were assigned to a test set, with maximum of 4 points were possible for all three test tasks). The independent variables of interest were (a) the three planning conditions, (b) proficiency levels of participants, and (c) task types (independent, integrated tasks). For illustrating the effect of the planning condition, I performed repeated measures ANOVA (using IBM SPSS) with the total test score as the dependent variable and the task types and test sets as the within-subjects variables; additionally, proficiency levels and

planning conditions were entered as the between-group variables. I investigated a possible main and interaction effect of the type of planning condition and proficiency levels of participants (as well as the effect of the task types).

As a final follow-up analysis, I conducted Generalized Equating Estimations (GEE) in SPSS for further exploring the extent to which planning conditions and task types each contributes to better test performance. GEE extends regular regression modeling as it accommodates repeated observations in the data, thereby reducing the overestimations of significance statistics (Type 1 error) (Ghisletta & Spini, 2004). For the analysis, I created a binary dependent variable by recoding each participant either as a "0" (low scorers) or "1" (high scorers) to be entered in the model. I added participants' average scores (i.e., average scores between the two raters) from all nine tasks to derive a single sum score for all individuals. Based on the median of the sum score, I was able to divide the participants largely into two sub-groups of *low*- (n = 49) and *high-scorers* (n = 50) on the speaking test. Independent variables were *planning condition* (with Unguided Planning condition being the reference or the "dummy" variable) and *test-task type* (with IP task being the reference variable).

2.4.2 Research question 2: Does the type of planning (guided versus unguided) affect the discourse quality of test candidates? If yes, what are the influences of the different task types?

The second research question was raised to investigate whether the type of planning (guided versus unguided) had effects on the quality of the spoken responses. In the following section, I provide the procedures and statistical analysis taken to address this inquiry.

2.4.2.1 Transcription of spoken responses

Prior to conducting analyses, I transcribed all verbal responses verbatim. This yielded

889 transcribed texts from all 99 participants (with the exception of two missing data, 99 participants each responded to 9 test tasks). The transcriptions only encompassed responses produced within the given responding time; that is, I did not transcribe language produced beyond the given responding time (see Weigle, 2004, for the discussion of scoring incomplete responses in accordance to different scoring contexts).

As it was beyond the scope of the current study, I did not refer to a more rigorous transcription convention (*cf.* see Lazaraton, 2002, for the application of Conversation Analysis conventions on spoken data from the language assessment perspective). However, during the course of transcribing, I internally developed three specific codes to mark for certain utterance phenomena to support further qualitative analysis. As shown in Table 11, these codes helped in (a) identifying the occurrences of the filled pauses in the data; (b) indicating the existences of utterance phenomena pertaining to a specific *fluency* measure in the transcriptions; and (c) specifying unintelligible language; it should be noted that 9 out of 889 transcriptions (approximately 0.01% of the whole dataset) contained these incidences.

Upon completion of transcribing, I had two native speakers of English (who were both undergraduate students each pursuing his and hers bachelor's degrees in Linguistics at a large Mid-western university) cross-examine the accuracy of the transcriptions. I assigned approximately 50% of the data each to these individuals. The students read through the transcriptions while listening to the assigned audio files. In so doing, they each noted transcription errors, which primarily pertained to minor mechanic errors (e.g., spelling errors, word-level errors, punctuation errors, etc.). I then went through the transcribed texts again and revised the identified transcription errors accordingly.

Table 11

Transcription codes used in the present study

Codes/Marks	Descriptions	Examples from the data
uh, *um*	Filled pauses	*...I will **uh**, have my own stuff to do and **um**, therefore, I can build up my responsibility.*
Dash (-)	A cut-off, usually a glottal stop. This is to indicate: 1. *Breakdown fluency* measures such as *repetitions*, *replacements*, *reformulations*, *hesitations*, and *false starts* 2. Incomplete response	*...the woman is **kin- uh, sort of** worried about* (Indication of *replacements*) *...um, well was the, was about the um, saving money* – (end of responding time)
Square brackets with 'xxx'	Transcription doubt, uncertainty; words within squared brackets are uncertain or unintelligible.	*...the decision of university is ridiculous and she is um, **[xxx]**.*

2.4.2.2 Coding spoken responses according to CAF measures

I hired two external coders for coding and rating the verbal responses in terms of the CAF measures for the main analysis regarding spoken quality. Both of the coders were graduate students at the same Mid-western University and had extensive teaching experiences of English

in both the U.S. and in foreign countries. Coder A was a native speaker of English, pursuing his Master's degree in TESOL. Coder B was a non-native speaker of English and a doctoral student in Applied Linguistics, with extensive experience in interacting with L2 learners and assessing their productive language in the classroom contexts.

The coding and rating processes were rigorous and detail-oriented and lasted about four months until completion. The coding procedures can be broadly summarized as follows: (a) training phase on coding/rating by using the devised coding scheme and Iwashita and Elder's (2005) CAF rubric (which is described in detail in the next section); (b) "playing around" with the dataset, with the two coders freely spending a specific amount of time on getting familiarized with the dataset and coding scheme; (c) a number of interim meetings (between the coders, and among the coders and I) in between on resolving concerns and questions prior to the main coding/rating phase; (d) joint coding/rating phase on a subset of data; (e) interim sessions (again, between the coders, and among the coders and I) on revisiting the coding scheme, and discussing emerging discrepancies in the dataset; (f) recoding of specific parts in the dataset to meet a certain level of agreement; and (f) independent coding/rating of the rest of the dataset.

The purpose of the joint coding/rating was to establish reliable coding/rating; that is, to obtain inter-coder reliability, which in turn can ensure reliability for the qualitative results. For this, I assigned approximately 30% of the whole dataset commonly to the two coders, which consisted of 30 participants' spoken responses. More specifically, I extracted 10 participants each from the three experiment groups. In the end, the coders analyzed a total of 270 same set of speech samples (30 participants each responding to 9 test tasks) in this joint coding phase.

Another project assistant (an undergraduate student at the same university) helped on various phases of coding, particularly on the mechanic aspects of coding (e.g., deriving the total

number of words and syllables in each speech sample).

2.4.2.3 Subjective ratings on Iwashita and Elder's (2005) CAF rubric

The same two coders holistically rated the participants' spoken performance using Elder and Iwashita's (2005) CAF rubric (See Appendix G). Such a holistic rating was essential for complementing the quantified CAF measures in the manual coding phase (Elder & Iwashita, 2005).

All three CAF constructs were each examined on a 5-point scale. The descriptors for each construct represented the sub-measures employed in the present study. The *complexity* descriptors concerned the trade-off between *complexity* and *accuracy*; higher ratings were given to responses in which complex meanings (expressed through the use of a variety of verb forms and syntactic units) were conveyed at the expense of grammaticality. *Accuracy* descriptors referred to participants' linguistic control over correct language forms. For *fluency*, speech rate as well as measures representing *breakdown fluency* was considered.

2.4.2.4 Statistical analysis

Prior to conducting the main statistical analysis, I went through the coding and rating results and obtained inter-coder reliability. I conducted Intra-class correlation coefficient (ICCs) using SPSS. ICC is an extension of the conventional Pearson correlation coefficient in that it reflects both degree of correlation and agreement between assessors (*cf.* Pearson correlation coefficient is only a measure of correlation; McGraw & Wong, 1996). In addition, following Tavakoli and Skehan (2005), I also conducted a series of exploratory factor analyses to inspect whether the qualitative measures each represent the larger CAF dimensions to which they were assigned.

For the main analysis, I treated each measure from the three discourse qualities as a

dependent variable with the independent variables being the planning condition. I performed seven repeated measures MANOVA (RM MANOVA) with regards to seven sub-categories of CAF dimensions. Subsequently, I carried out a series of post-hoc pairwise comparisons and one-way ANOVA for addressing the interaction effects found in the RM MANOVA analysis.

Finally, I conducted GEE analysis for exploring how speech quality measured for each planning condition contributed to overall test performance. I collapsed speech quality data pertaining to each planning condition and treated them as independent variables. I entered the same binary dependent variable of test performance (*low*- and *high-scorers*) in the model.

2.4.3 Research question 3: How do test candidates use their planning times?

For addressing research question 3, I downloaded the post-survey responses from Qualtrics, and formatted them into spreadsheets in Microsoft Excel. In Excel, I re-coded the responses into numerical values (e.g., responses on Likert-scale items) to further derive raw frequencies of each category of survey questions. Following previous researchers (Elder & Iwashita, 2005; Tavakoli & Skehan, 2005), I used SPSS to perform RM MANOVA in comparing the survey responses by planning condition and test-task types. General Linear Models (such as RM MANOVA) are fairly robust for ordinal data with small sample size (e.g., 40 – 60) in the dataset (See Stiger, Kosinski, Barnhart, & Kleinbaum, 1998 for detailed discussion).

2.4.4 Research question 4. How do test candidates perceive the given planning times?

To answer research question 4, I had the project assistant extract the precise segments of the interview responses from the entire audio file yielded from each participant (the entire testing session were audio-recorded for every participant). With the exclusion of 9 missing interview segments from 9 participants (due to poor audio-recording quality and practicality of data

collection procedures), I imported a total of 90 video segments into NVivo (version 10). In NVivo, I created separate individualized nodes (i.e., coding category) for each participant and designated attribute codes (e.g., his or her background variables) that represented their assigned groups and proficiency level. I then took an emergent and grounded-theory methods (Strauss & Corbin, 1998) while adopting the following coding procedures: (a) initial, open coding phase (Friedman, 2012) for developing a preliminary set of coding schemes and (b) axial and selective phases (Strauss & Corbin, 1998) for refining the coding schemes and gauging specific patterns in responses.

CHAPTER 3: RESULTS

In this chapter, I present results in the order of the four research questions that guided the current study. Therefore, four subsections largely constitute this chapter. In the first section, I report on participants' speaking test performance through the ratings given by the three raters. With the rating data, I first illuminate whether participants' test scores vary across testing conditions, which differed in terms of planning as well as task types. Next, I report on findings pertaining to the predictive relationship that planning conditions and task types each has with how participants performed on the speaking test. In the second section, I report on the extent to which participants' speech quality differs across testing conditions as indexed through CAF dimensions. In the third section, I turn to participants' survey responses and illustrate how the participants used and perceived the given planning times and conditions. In the final section, I provide more detailed responses from participants through their retrospective data.

3.1 Research question 1: Speaking test scores

3.1.1 Descriptive statistics for the speaking test scores

To answer the first research question, I first obtained descriptive statistics for participants' test performance across the three test sets. I broke down the test scores for each test set, and further by the three planning conditions as well as the three test tasks. Table 12 summarizes the mean test scores and the standard deviations (SD) pertaining to each testing condition. It should be noted that the mean test scores for each testing condition showcased in Table 12 are essentially the means of the averaged scores each individual received from the two raters.

As described in Table 12, it was apparent that participants' mean test scores did not differ to a substantial extent across the three test sets. While the mean scores did seem to slightly increase in Test Set C, the score ranges from Test Set A to Test Set C were not heavily dispersed. In particular, the mean test scores especially had marginal differences amongst the three planning conditions within and across the three test sets. For instance, on average, participants maintained in the 2.50 to 2.56 score range for the IP task regardless of the differing planning conditions within and across test sets. For the IT-RL and IT-L tasks, the mean test scores had minimal differences across the three planning conditions in every test sets as well (while as noted above, there were increases in the mean scores for these two tasks in Test Set C). This seem to suggest that participants performed similarly regardless of the differing test sets and the types of planning activity they engaged in.

Interestingly, a subtle yet noticeable pattern could be seen in the mean test scores for the three test tasks. Within and across test sets, and particularly regardless of the planning conditions, participants generally scored the least for the IP task, while scoring higher for the two integrated tasks. Between the IT-RL and IT-L tasks, it was the former that participants generally scored higher; however, the score difference between the two integrated tasks seem to be less clear when making the comparison between the two integrated tasks. In all test sets, participants' score ranges for the IT-RL tasks maintained within the 2.80 to 2.90 range, while for the IP tasks, the range was within 2.50 to 2.56. The mean test scores for the IT-L task, on the other hand, somewhat maintained in between the IP and the IT-RL tasks.

Table 12

Descriptive statistics for participants' speaking test scores

	Test Set A			Test Set B			Test Set C		
	UG (N = 32)	GW (N = 33)	GT (N = 33)	UG (N = 33)	GW (N = 33)	GT (N = 33)	UG (N = 32)	GW (N = 33)	GT (N = 32)
IP	2.45 (0.62)	2.50 (0.67)	2.56 (0.61)	2.50 (0.66)	2.49 (0.58)	2.52 (0.63)	2.53 (0.69)	2.56 (0.74)	2.54 (0.51)
IT-RL	2.80 (0.66)	2.82 (0.66)	2.79 (0.69)	2.78 (0.42)	2.80 (0.49)	2.83 (0.67)	2.88 (0.57)	2.89 (0.66)	2.90 (0.49)
IT-L	2.71 (0.60)	2.76 (0.72)	2.76 (0.60)	2.71 (0.45)	2.73 (0.52)	2.72 (0.65)	2.82 (0.66)	2.80 (0.61)	2.77 (0.41)

Note. UG, GW, and GT each indicate Unguided-Planning, Guided-Writing Planning, and Guided-Thinking Silently Planning. IP, IT-RL, and IT-L each indicate Independent task, Integrated-Reading and Listening task, and Integrated-Listening only task. Standard deviations are in parenthesis.

Note. In Test Set A, one participant's responses for all test tasks were missing. In Test Set C, there were two participants' responses for all test tasks missing.

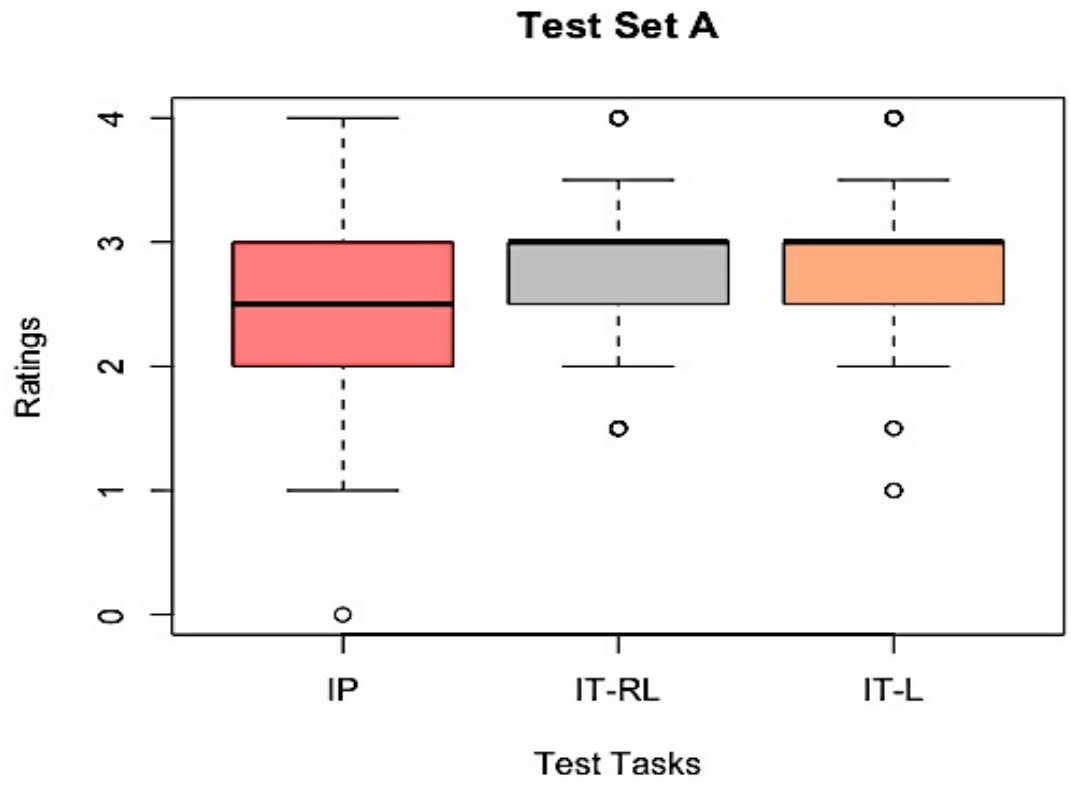


Figure 2 Boxplots representing the distribution of the test scores from Test Set A by test tasks

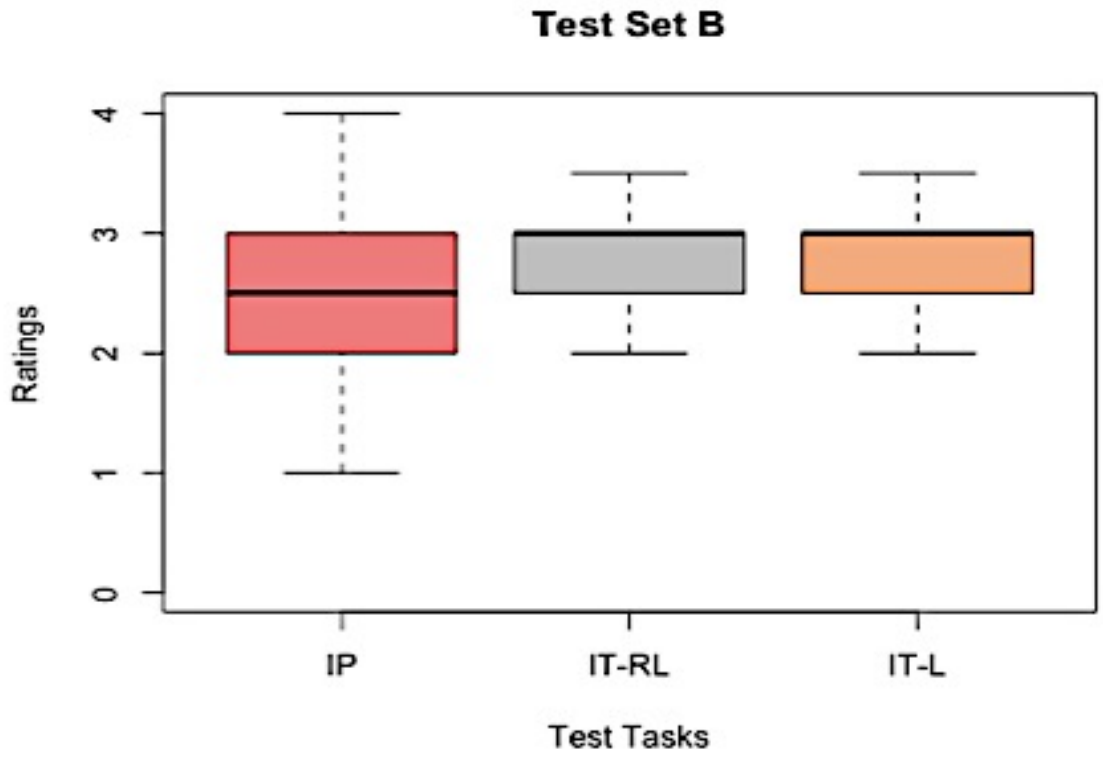


Figure 3 Boxplots representing the distribution of the test scores from Test Set B by test tasks

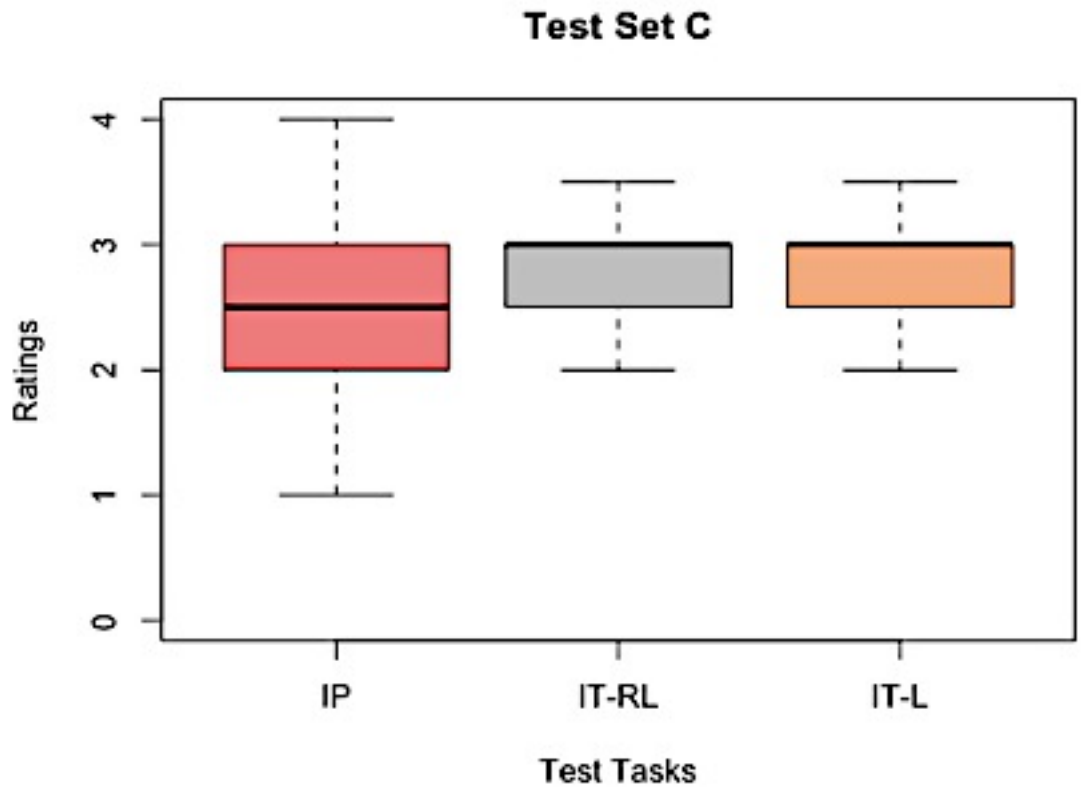


Figure 4 Boxplots representing the distribution of the test scores from Test Set C by test tasks

Taken altogether, the trend emerging in Table 12 was that participants' test scores did not vary to a noticeable extent in accordance to the type of planning. The differences in the mean scores were amongst the type of test tasks; that is, the scores were generally higher for the two integrated tasks and lower for the IP task. Such a trend is additionally illustrated in Figures 2, 3, and 4 through the boxplots generated for the overall mean scores by the test tasks within each test set from all participants (i.e., boxplots are not broken down by the planning conditions). It seems that for the IP task in all three test sets, participants consistently have lower medians (which generally is close to the average score) than the two integrated tasks. In addition, in all three test sets, participants' scores are relatively more spread out for the IP task yet for the two integrated tasks, participants' scores tend to vary from one another to a lesser extent.

3.1.2 FACETS analysis

In this section, I report on how the raw data of the test scores reported in the previous section are interpreted through Rasch modeling. I specifically focus on comparatively reporting how each factor or facet specified in the analyses explains the variance in the test scores. Because I analyzed each test administration (i.e., test sets) independently from one another, I present three MFRM analyses in accordance to the three test sets administered.

Prior to exploring the outputs for individual facets in detail, I first inspected the fit statistics generated by FACETS to confirm whether the data fit the specified Rasch models. When the data set shows sufficient fit to a particular Rasch model, invariance among specified measurement factors are verified (Eckes, 2009). Measurement invariance (Bond & Fox, 2007) in Rasch modeling is particularly important as it entails (a) test scores (i.e., observation points entered in the model) contain sufficient information required for estimation; and (b) the

unidimensionality of the test (i.e., the test items are measuring the same latent construct; in the current analyses, the latent construct is the speaking test performance) (Eckes, 2009).

Table 13 summarizes the fit indices generated for the three Rasch models. It should be noted that the reported values are composite values from averaging the fit indices generated for the individual elements within each facet; for instance, the fit values for *raters* are the means of the corresponding fit values from the three individual raters. *Model Standard Error of Measurement* (henceforth *Model S.E.*) and the *Infit* and the *Outfit Mean Square* statistics each represent *measurement precision* (i.e., consistency of estimation) and *measurement accuracy* (i.e., correctness of estimation) in the models (Linacre, 2017). As in Table 13 (see columns 2, 5, and 8) in all three test sets, the *Model S.E.*s for the six facets were all small, mostly clustering around 0; this indicates that the corresponding facet in the models were measured with relatively high precision (Harvill, 1991). In terms of the *Infit* and *Outfit* statistics (see columns 3, 4, 6, 7, 9, and 10 in Table 13), the indices all fell within the range of the lowest of 0.89 to the highest of 1.05, which correspond to the conventional range of acceptable fit (i.e., values located between the range of 0.7 or 0.8 to 1.2 or 1.3, Linacre, 1999) (See Linacre, 2000 for the discussion of a broader range of model fit and how different ranges apply differently depending on the assessment purpose and data size). Standardized Z (*Zstd*) values provide additional information on good model-fit in the data. Values closer to 0 demonstrate that the data did fit the model sufficiently. Table 13 suggests that all *Zstd* values were close to 0.

Taken altogether, the fit values reported in Table 13 did not flag extreme tendencies of either misfit or overfit of the data to the models.

Table 13

Model-fit statistics Summary for Test Sets A, B, and C

	Test Set A			Test Set B			Test Set C		
Facets	Model *SE*	Infit mean square (Zstd)	Outfit mean square (Zstd)	Model *SE*	Infit mean square (Zstd)	Outfit mean square (Zstd)	Model *SE*	Infit mean square (Zstd)	Outfit mean square (Zstd)
Raters	0.16	0.97 (-0.30)	1.04 (0.30)	0.17	1.00 (0.00)	0.94 (-0.40)	0.18	0.98 (-0.30)	0.89 (-0.60)
Examinee	0.95	0.99 (0.00)	1.05 (0.00)	0.98	0.96 (-0.10)	0.94 (-0.10)	0.97	0.89 (0.00)	0.89 (0.00)
Group	0.13	0.98 (-0.20)	1.04 (0.30)	0.17	1.00 (0.00)	0.94 (-0.40)	0.18	0.98 (-0.20)	0.89 (-0.50)
Proficiency level	0.14	1.00 (-0.30)	1.05 (-0.10)	0.14	0.99 (-0.10)	0.94 (-0.50)	0.15	0.98 (-0.20)	0.89 (-0.70)
Planning condition	0.13	0.98 (-0.20)	1.04 (0.30)	0.17	1.00 (0.00)	0.94 (-0.40)	0.18	0.98 (-0.20)	0.89 (-0.50)
Task Types	0.16	0.97 (-0.20)	1.04 (0.40)	0.17	0.99 (-0.10)	0.94 (-0.40)	0.18	0.98 (-0.20)	0.89 (-0.50)

Note. Zstd indicates Standardized fit statistics.

3.1.2.1 Test Set A

In this section, I present the variable map and the corresponding outputs from FACETS that further explain the visual information. As can be seen in Figure 5, the variable map (also known as the "Wright Map") displays comprehensive information on how all the facets entered in the model are represented in a single frame of reference; for MRFM modeling, this is the logit scale. The "+" and "-" in front of the facet headings indicate whether the corresponding facet measures were positively or negatively oriented. Positively oriented facets in the ruler mean that the data points in the higher positions have higher measures. In reverse, negatively oriented facets represent that highly positioned data points in the column have lower measures.

In the first column in the map, *measr* (measure) depicts the logit scale that positions all measures of facets; the scale range for Test Set A spanned from 6 logits to -6 logits.

In the second column, *judge* provides information on the level of severity or leniency of the three raters (A, B, and C) when they assessed participants' speaking test scores for Test Set A. As the facet is positively oriented, more sever raters are to appear lower in the column, while more lenient raters are to be positioned higher. With the relatively tight clustering of rater A, B, and C in this column, it seems that the variability across raters in terms of their level of severity was not substantial. Rater A (appearing highest in the column) had a severity measure of 0.90 logits while rater B and C each had severity measures of 0.63 and 0.55 logits, respectively. This essentially corresponded to less than 1-logit spread amongst the raters. The raters, even with a relatively short period of time to be accustomed to the rubric, did not have extreme differences in rating behavior. The *Model S.E.*s and the *Infit* statistics for the three raters were all within the acceptable range (*Model S.E.*: 0.16 to 0.17; *Infit*: 0.84 to 0.94), demonstrating that raters were relatively consistent in giving scores.

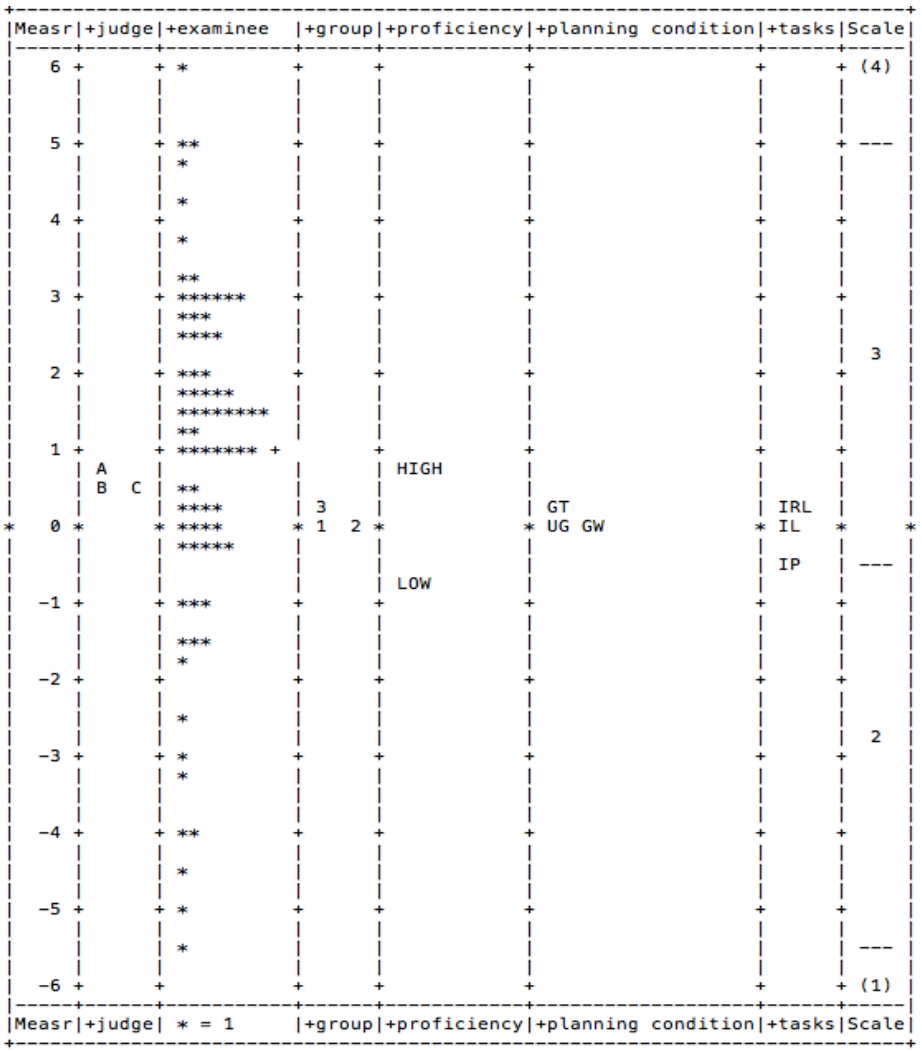


Figure 5 *Variable map for Test Set A*

The third column, *examinees*, depicts the estimates of participants' proficiency on the speaking test (and particularly on Test Set A). Here, each star refers to one examinee. As the facet is positively oriented, participants who scored higher appear in the higher end of the column. Estimates for participants ranged from -6.61 logits to 5.63 logits. While the spread seems to be relatively wide, the *examinee* separation value (i.e., the spread of test taker proficiency that displays the separation among test takers as defined by their performance; Stone & Wright, 1988) of 2.24 denoted that participants were measurably separable into (at least) 2

levels of ability. The reliability of the separation index of .83 suggested that the separation levels were fairly acceptable. This may suggest that participants were likely to cluster around certain rating categories, with few participants scoring in the extreme ends of the scoring scale (i.e., 0 or 4).

The fourth column, *group*, compared the three experimental groups in terms of their overall performance on the speaking test. Estimates for groups spanned from -0.13 logits to 0.07 logit; the spread was narrow. Indeed, the separation value was 0 with a reliability of .90, indicating that the groups performed similarly. The fixed chi-square value provides additional information on whether the identified subgroups within a facet differ from one another to a statistically significant extent. There was a statistically non-significant chi-square value of 0.9 ($df = 2$, $p = .62$), which suggested that the three subgroups performed similarly.

The fifth column, *proficiency*, compared the two oral proficiency levels (high and low) previously assigned based on the EI scores. With the facet being positively oriented, participants with a *high* oral proficiency level appeared in the higher part of the column, which was indicative of their higher test performance. Participants who were designated as having a *low* oral proficiency level appeared in the lower part of the column, which represented their lower test performance than the *high* level participants. Visually, the logit difference seems to be substantial; the range was -1.01 logits (*high* level) to 1.01 logits (*low* level). The separation value between these two subgroups was 2.42 with a reliability of .98, which corroborated with the large spread of logits. The fixed chi-square value of 112.1 (df =1) was statistically significant ($p = .000$). Overall, it seems to be the case that raters were likely to rate participants relative to their oral proficiency level as evidenced from the EI test. The pre-determined two levels deemed to align with how these participants performed in the actual speaking test.

The sixth column, *planning condition*, displayed the comparison of test performance in terms of the three planning conditions. In alignment with the descriptive results in Table 12, the ratings did not differ to a significant extent across the planning conditions. Table 14 further reports on the summary statistics related to the planning condition. The observed average among the three conditions, which represents the raw observed score, had marginal difference. In logit scale (and with the negative orientation of the column), test performance under GT was the highest of -0.13 logits (*SE* = .16), followed by UG (*SE* = .17) and GW (*SE* = .17). Indeed, the separation index amongst the three conditions was 0 with a reliability of .90. The Fixed chi-square value for *planning condition* was 0.6 (*df* = 2), which was not statistically significant (*p* = .62); this indicated that the sub-conditions did not significantly differ from one another.

Table 14

Summary statistics of planning condition for Test Set A

Planning condition sub category	Observed Average	Fair (M) Average	Measure	Model *SE*	Infit mean square	Outfit mean square
UG	2.65	2.73	0.07	0.16	0.95	0.91
GW	2.69	2.77	0.05	0.17	1.01	1.05
GT	2.70	2.78	-0.13	0.16	0.96	0.98

The seventh column, *tasks*, demonstrated the test performance relative to the three test-task types. With the facet being positively oriented, tasks that received higher performance appeared in the higher part of the column. In alignment with Table 12, Table 15 demonstrated

that it was the two integrated tasks that received higher ratings than the IP task. IT-RL task and IT-L task each had values of -0.10 logits (*SE* = .16) and 0.03 logits (*SE* = .16), respectively. IP task had a value of 0.59 logits (*SE* = .16). While the differences in the logits seem to be not large, the separation index for the sub-categories was 2.33 with a reliability value of .84. This suggested that there were at least two distinct levels of categories that measurably separate participants in relation to their performance on the three test tasks. There was a statistically significant fixed chi-square value of 19.3 ($df = 2$, $p = .00$), which was also indicative of the possibility that the raters gave distinguishable ratings across tasks.

Table 15

Summary statistics of test-task types for Test Set A

Test-task types sub category	Observed Average	Fair (M) Average	Measure	Model *SE*	Infit mean square	Outfit mean square
IP	2.50	2.60	0.59	0.16	0.94	1.04
IT-RL	2.80	2.90	-0.10	0.16	1.01	1.03
IT-L	2.74	2.84	0.03	0.16	0.98	1.05

Overall, the findings for Test Set A supported the descriptive statistics reported in Table 12: Test performances for Test Set A marginally varied by planning condition while demonstrating evidence of an influence of test-task types. I further report on the follow-up analyses on pairwise comparisons amongst planning conditions and test-task types in section 3.1.3.

3.1.2.2 Test Set B

As can be seen in Figure 6, the variable map for Test Set B was quite similar to that of Test Set A (see Figure 5). Again, all six facets were positively oriented; the subcategory with higher ratings was positioned in the higher part of each corresponding column. Because of the similarity, I put more focus on reporting the results for the two facets of interest, *planning condition* and *test-task types*.

From the second column, *judge*, it is noticeable that the three raters did not have deviating patterns from when they scored speaking test performances. The differences in the level of severity were minimal, with a marginal spread in the logits amongst the raters (rater A = 0.75 logits, rater B = 0.58 logits, rater C = 0.98 logits). Raters also seemed to consistently conform to the rating scale as evidenced through their *Model S.E.*s and the *Infit* statistics, which fell within an acceptable range (*Model S.E.*: 0.16; *Infit*: 0.81 to 1.02).

The estimates for *examinee* in the third column spanned from -5.78 logits to 4.30 logits. As in Test Set A, the separation value of 2.37 with a reliability of .88 shows that there were at least two levels of categories distinguishing participants in the dataset. This was supported by the statistically significant fixed chi-square value of 475 ($df = 98$; $p = .00$).

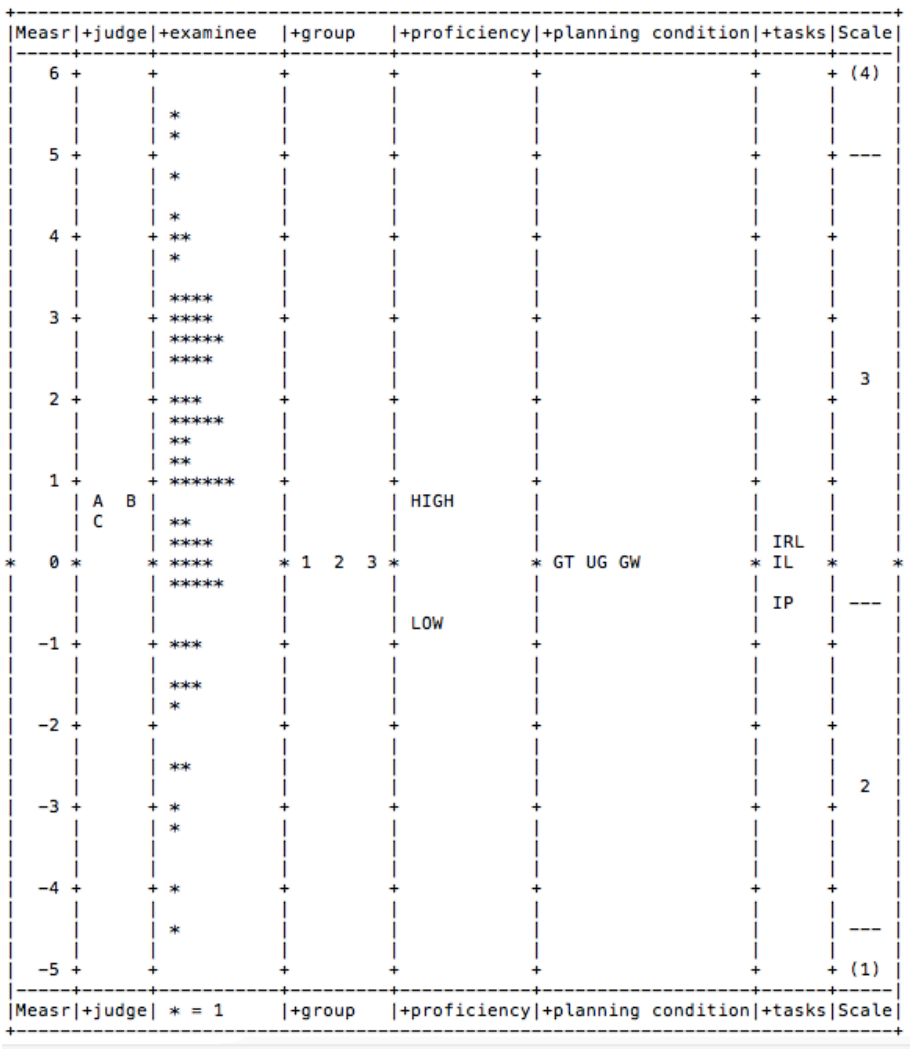


Figure 6 *Wright map for Test Set B*

As in the fourth column, *group*, there were no significant differences in performance by the three experimental groups, with the highest and the lowest performing group being 0.20 logits apart. As in Test Set A for *group*, the separation index was 0 with a reliability of .90, confirming that there were no distinct levels that distinguish among the three experimental groups in terms of test performance.

From the fifth column, *proficiency*, differences between participants with *high* and *low* oral proficiency was visually clear. The two sub-levels of oral proficiency were 1.48 logits apart

from one another, with participants of *high* oral proficiency being scored higher than participants with lower oral proficiency. These results were further supported by a separation index of 2.00 with a reliability of .96.

From the sixth column, *planning condition*, it was apparent that the test performance for Test Set B did not differ across the differing planning conditions. As in Table 16, observed values had no significant differences relative to the planning conditions. The differences were also marginal in the logits scale; the highest was GT of 2.69 logits (*SE* = 0.17), subsequently followed by GW of 0.05 logits (*SE* = 0.16) and UG of -0.12 logits (*SE* = 0.16). In addition, there was a separation value of 0 with a reliability of .90, which confirmed that test performance in relations to the differing planning conditions could not be separated into distinctive levels. That is, raters did not exercise distinguishable patterns of ratings relative to the three planning conditions. This was further corroborated by the non-significant fixed chi-square value of 0.9 (*df* = 2, *p* = .65).

Table 16

Summary statistics of planning condition for Test Set B

Planning condition sub category	Observed Average	Fair (M) Average	Measure	Model *SE*	Infit mean square	Outfit mean square
UG	2.66	2.72	-0.12	0.16	0.91	0.88
GW	2.67	2.77	0.05	0.16	1.01	1.06
GT	2.69	2.77	0.08	0.17	0.98	0.88

Finally, the seventh column in the variable map provides similar information as in Test Set A. As further depicted in Table 17, generally, test performance differed across independent and integrated tasks. Test performance was the highest for the IT-RL task with -0.15 logits (*SE* = 0.16) and the lowest for the IP task with 0.67 logits (*SE* = 0.17). The IT-L task was in between these two tasks with 0.02 logits (*SE* = 0.17). While the differences in logits seem not to be large between the IP and the integrated tasks, there was a noticeable separation index of 2.59 with a reliability of .90. This suggested that there were at least two levels of discernable subcategories of test performance in relations to test-task types in the dataset. This result was further confirmed with a statistically significant fixed chi-square value of 25.3 (*df* = 2, *p* = .00).

Table 17

Summary statistics of test-task type for Test Set B

Test-task type sub category	Observed Average	Fair (M) Average	Measure	Model *SE*	Infit mean square	Outfit mean square
IP	2.50	2.60	0.67	0.17	1.01	1.02
IT-RL	2.80	2.89	-0.15	0.16	0.90	0.88
IT-L	2.72	2.82	0.02	0.16	0.86	0.76

As noted above, similar findings from Test Set A could be drawn to Test Set B. Score variation was minimal with regards to which planning activities participants utilized before responding. Test-task types brought differences in test performance; participants were rated higher on the integrated tasks as opposed to the IP task.

3.1.2.3 Test Set C

When it comes to Test Set C, the variable map displayed in Figure 7 presents similar patterns of results as in the previous test sets. Thus, I briefly touch on the results pertaining to the following facets *judge*, *examinee*, *group*, and *proficiency*, and elaborate more on *planning condition* and *tasks*.

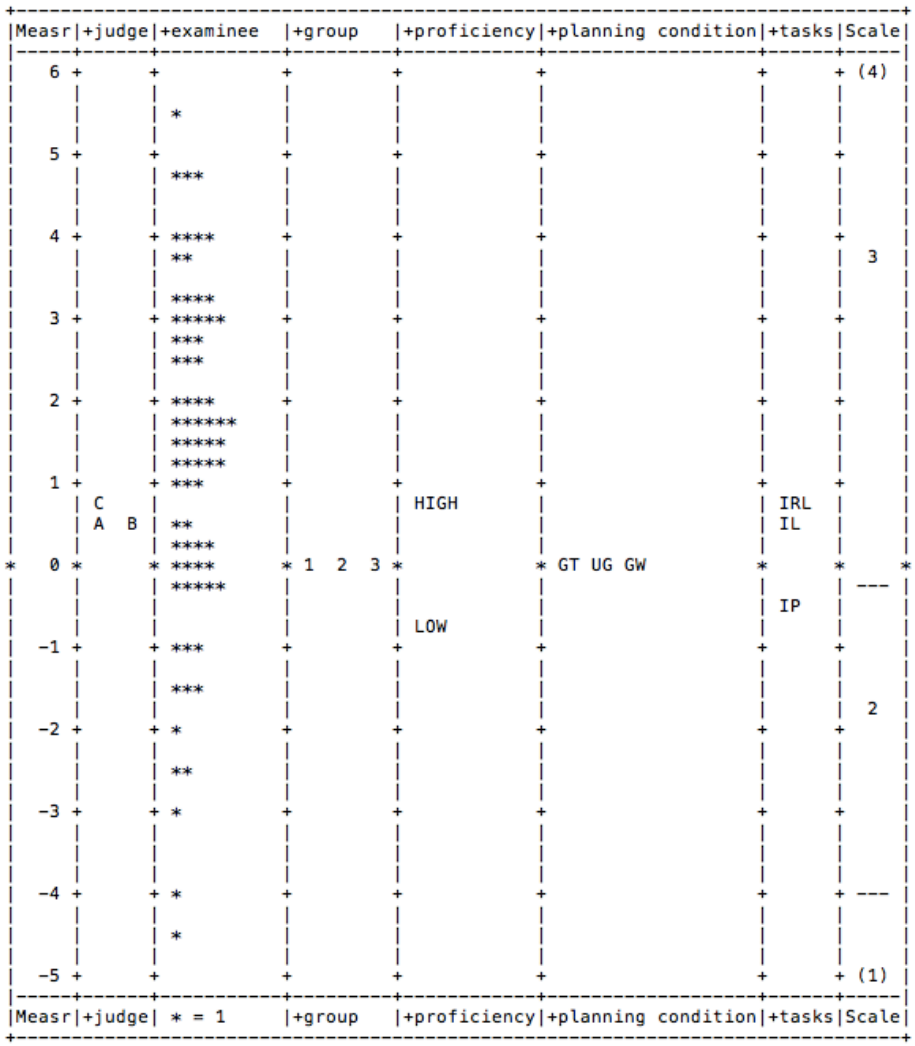


Figure 7 *Wright map for Test Set C*

The three raters, again, demonstrated fairly consistent ratings relative to one another. Their severity measures spanned from the lowest of 0.48 (Rater A) to the highest of 0.78 (Rater C). Reliability in their levels of severity were suggested through their *Model S.E.*s and *Infit* indices; all of the values maintained within the acceptable range of scale. For both *examinees* and *proficiency*, the separation indices were similarly suggesting 2 levels of test performance for each facet (*examinees*: separation index = 2.51, reliability = .84; *proficiency*: separation index = 2.82, reliability = .96). On the other hand, FACETS indicated comparability across the three sub-groups (separation index = 0, reliability = 0.94).

In terms of *planning condition*, no significant differences were noticeable in the variable map. The summary statistics for *planning condition* reported in Table 18 verified such comparability across the planning activities. The observed average values as well as the logit measures maintained a narrow range of values regardless of the planning condition. The separation index of 0 with a reliability of .90 further supported how test performance did not differ to a great extent in terms of the type of planning participants utilized. In addition, the fixed chi-square value of 0.1 was not statistically significant ($df = 2$, $p = .96$), which confirmed that variation in participants' test performance was kept to minimal.

Table 18

Summary statistics of planning condition for Test Set C

Planning condition sub category	Observed Average	Fair (M) Average	Measure	Model *SE*	Infit mean square	Outfit mean square

Table 18 (cont'd)

UG	2.74	2.77	0.01	0.19	0.96	0.95
GW	2.75	2.83	0.03	0.19	1.01	0.81
GT	2.74	2.81	0.01	0.17	0.97	0.91

The variable map as well as the summary statistics in Table 19 also showcased a similar tendency of differences in test performance relative to the test tasks. The observed average for the three tasks demonstrated a slight increase in ratings for Test Set C as opposed to the previous two test sets. In terms of logits, the IP task had the lowest value of 0.43 (*SE* = 0.18), while the IT-RL task had the highest value of -0.45 (*SE* = 0.18). The difference in logits was 0.85, which was the largest value amongst the three Test Sets. The separation index was 2.79 with a reliability of .84, implying that participants' test performance in terms of test-task types can be divided measurably into at least two distinct levels. Such a separation of test performance was supported by a significant fixed-chi square value of 19.3 (*df* =2, *p* = .00).

Table 19

Summary statistics of test-task type for Test Set C

Test-task type sub category	Observed Average	Fair (M) Average	Measure	Model *SE*	Infit mean square	Outfit mean square
IP	2.54	2.63	0.43	0.18	0.76	0.67
IT-RL	2.90	2.94	-0.45	0.18	1.08	0.90
IT-L	2.80	2.83	-0.32	0.18	1.10	1.11

Taken altogether, all three MFRM models applied to the three datasets commonly informed the following finding: participants' speaking test scores seemed to vary due to the type of test tasks, while the type of planning that they had employed had minimal impact. There were no significant differences on test performance depending on whether participants performed under unguided or guided planning conditions. Amongst the three test-task types, participants were likely to be rated lower on the IP task relative to the two integrated tasks.

3.1.3 Follow-up repeated measures ANOVA

To further elaborate on the MRFM analyses, I conducted a two-way repeated measures ANOVA (henceforth RM ANOVA). In this model, I entered *Test Set* (A, B, and C) and *Test-task type* (IP, IT-RL, and IT-L tasks) as the within-subjects variables, and *Planning Condition* (UG, GW, and GT) and *Proficiency Level* (oral proficiency levels designated from the EI test results) as the between-subjects variables. I specified the dependent variables as the speaking test scores from each test set and task types (e.g., Test Set A, IP task score).

The results from the RM ANOVA indicated that there were no statistically significant higher-level interactions between the variables (e.g., two-, three-, or four-way interaction between *Test Set, Test-task type*, *Planning Condition*, and *Proficiency Level*). However, the Greenhouse-Geisser statistics (for accounting for the violated sphericity assumptions in the data set; Field, 2009) informed that *Test-task type* ($F_{2, 182} = 13.116$, $p = .000$, $\eta p^2 = .126$) and *Proficiency Level* ($F_{1, 91} = 79.231$, $p = .000$, $\eta p^2 = .365$) were statistically significant factors impacting on how participants performed on the speaking tests, with moderate to high effect sizes. *Planning Condition*, on the other hand, appeared as a statistically non-significant factor on the overall test performance ($F_{2, 91} = 0.009$, $p = .991$, $\eta p^2 = .000$). In addition, it should be also noted that *Test Set* did not impact on how participants performed to a statistically significant

extent ($F_{2, 182} = 1.116$, $p = .351$, $\eta p^2 = .024$). These findings further support the MFRM models and outputs reported in the previous section.

Findings from the RM ANOVA analysis were additionally captured in the line graphs in Figure 8. Here, participants' test performance followed a similar pattern regardless of the three planning conditions. That is, participants' mean speaking scores were the lowest for the IP task and the highest for the IT-RL task in all three planning conditions.

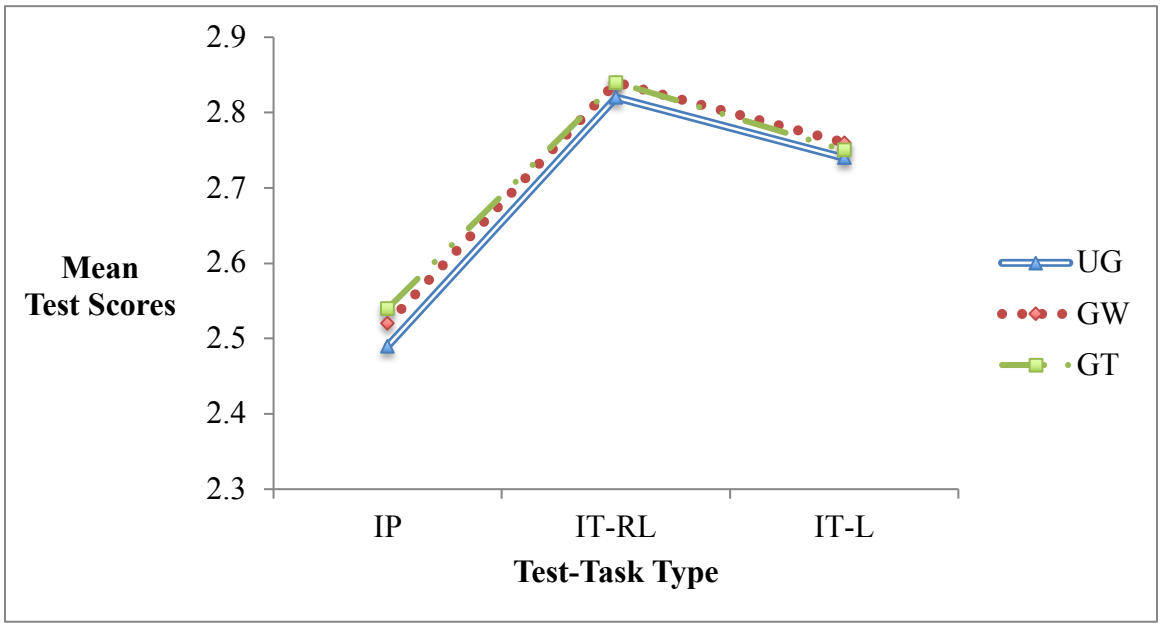


Figure 8 *Mean speaking test scores by test-task type and planning conditions*

To further illuminate on these results, I conducted a series of post-hoc pairwise comparisons of test performance across the three test tasks within each test administration (e.g., IP tasks versus IT-RL tasks in Test Set A). The primary comparison was between the IP task and the two integrated tasks. Table 20 reports the full summary statistics generated from the analysis. It could be seen that in all three test sets, test performance on the IP task was statistically different than both IT-RL and IT-L tasks. In Test Set A, the IP task was scored lower than both

the IT-RL task (mean difference = -0.29, 95% CI [-0.41, -0.19], p = .000, d = 0.50) and the IT-L task (mean difference = -0.26, 95% CI [-0.37, -0.15], p = .000, d = 0.42). In Test Set B, the IP task was scored the least compared to the IT-RL task (mean difference = -0.34, 95% CI [-0.45, -0.19], p = .000, d = 0.53) and the IT-L task (mean difference = -0.24, 95% CI [-0.35, -0.28], p = .000, d = 0.43). Likewise, in Test Set C, the IP task score differed significantly than the IT-RL task (mean difference = -0.38, 95% CI [-0.41, -0.17], p = .000, d = 0.75) and the IT-L task (mean difference = -0.30, 95% CI [-0.39, -0.19], p = .000, d = 0.50). Effect sizes for all pairwise comparisons were in the moderate range. On the other hand, score differences were statistically non-significant between the two integrated tasks across the three test sets.

Overall, findings from the RM ANOVA and the post-hoc comparisons suggest that participants' performance varied in terms of the type of test tasks, and not the type of planning activities. Amongst the test tasks, statistically significant score differences existed when comparing the IP task to the two integrated tasks. Even with small raw score differences, the interferential statistics and their effect sizes confirmed that participants scored higher on the two integrated tasks than the IP task.

As a final follow-up on the overall results, I conducted a generalized estimating equation (GEE) analysis to verify the predictive power that *planning condition* and *test-task types* each has on the level of test performance. I specified the following as the predictor factors: (1) *planning condition* as a categorical variable (with UG as the reference variable); (2) *test-task types* as continuous variables (with IT-L as the reference variable). The binary variable was participants' level of test performance (*low-* and *high-scorers*). The GEE results confirmed that *planning condition* was not a significant predictor of higher test scores (Wald χ^2 (2) = 0.49, p = .824). More specifically, neither the GW (b = 0.27, p = .524, 95% CI = [0.120, 14.32]) nor GT

planning conditions (b = 0.31, p = .452, 95% CI = [0.212, 19.215]) were associated to higher performance relative to the UG condition. On the other hand, there was a statistically significant effect of *test-task types* on predicting higher test performance (Wald χ^2 (2) = 4.958, p = .003). Especially, participants' higher ratings were likely to occur in the IT-RL task (b = 4.958, p = .000, 95% CI = [11.26, 20.71]) and the IT-L task (b = 4.994, p = .000, 95% CI = [5.32, 19.41]) relative to the baseline IP task. This is indicative of a possibility that the two integrated tasks were able to contribute to better test performance as opposed to the IP task condition.

Table 20

Summary statistics of pairwise comparisons for Test Sets A, B, and C

Pair	Test Set A				Test Set B				Test Set C			
	Mean diff.	*SD*	*t*	*p*	Mean diff.	*SD*	*t*	*p*	Mean diff.	*SD*	*t*	*p*
IP vs. IT-RL	-0.30	0.55	-5.333	.000	-0.34	0.55	-6.155	.000	-0.38	0.52	-7.385	.000
IP vs. IT-L	-0.26	0.55	-4.755	.000	-0.24	0.52	-4.300	.000	-0.30	0.54	-5.174	.000
IT-RL vs. IT-L	0.03	0.51	0.702	.484	0.04	0.43	0.820	.355	0.06	0.43	0.865	.342

Note. IP, IT-RL, and IT-L each refers to Independent, Integrated-Reading and Listening, and Integrated-Listening tasks, respectively. Mean diff. refers to the mean differences between the mean scores of the two test tasks.

3.2 Research question 2: Speech quality

3.2.1 Inter-coder reliability

Prior to examining the descriptive as well as inferential statistics on speech quality, I inspected the raters' inter-coder reliability by calculating the Intra-class Correlation Coefficients (ICCs). It should be noted that the variables included for the ICC analysis were the *raw coding* assigned by each coder (as opposed to the quantified values of speech quality measures that I subsequently generated; these include measures such as *speech rate* or *lexical errors per 100 words* that made use of the raw coding). I applied ICCs to the dataset that the two coders were assigned to commonly analyze. This corresponded to a subset of speech samples that comprises 30% of the entire dataset. Because there were two specific coders to be assessed, I selected a 2-way mixed-effects model with absolute agreement definition (Shrout & Fleiss, 1979). An absolute agreement in the current analysis would refer to the extent to which the two coders overlap in evaluating a particular quality dimension for an individual participant. In addition, an absolute agreement definition for the ICC model is generally used with repeated observations (as in the current study) in the dataset (Koo & Li, 2016).

Tables 21, 22, and 23 each summarizes the ICCs (and their 95% confidence intervals) with regard to the *fluency*, *accuracy*, and *complexity* dimensions generated by the two coders within each test set. Table 24 additionally reports on the ICCs between the two coders on their assessment based on the 4-point scale CAF rubric. For the basic descriptive statistics for all raw coding data, see Appendix H.

As the tables report, the agreement level between the two coders is moderate to high for most of the measures, within the .75 to .99 range (Cicchetti, 1994). In terms of the *fluency* indices, the ICCs were the lowest for *replacements* in Test Set B (ICC = .75, 95% CI [0.61,

0.83]), and the highest for the *filled pauses* measures in Test Set C (ICC = .99, 95% CI [0.99, 0.99]). Excluding certain measures having an ICC value lower than .80 in Test Sets A (e.g., *false starts*) and B (e.g., *reformulations*, *replacements*, and *false starts*), all of the measures had high ICC values in Test Set C. Yet it is noticeable that variations in coding were identified in the *repair fluency* measures.

Likewise, high levels of ICCs were found for both *accuracy* and *complexity* measures. For *accuracy*, *lexical errors* in Test Set B had the lowest ICC (ICC = .82, 95% CI [0.83, 0.87]) while *error-free clauses* in Test Set A had the highest ICC (ICC = .91, 95% CI [0.87, 0.94]). For *complexity*, *subordinate clauses* in Test Set C had the lowest ICC value (ICC = .85, 95% CI [0.80, 0.90]), and *AS-units* in Test Set A had the highest value (ICC = .95, 95% CI [0.91, 0.97]).

The ICCs for CAF ratings were also within the acceptable range of agreement level. The two coders agreed upon the least for *fluency rating* in Test Set A (ICC = .80, 95% CI [0.77, 0.84], and the most for *accuracy rating* in Test Set B (ICC = .92, 95% CI [0.90, 0.98]).

Table 21

Intra-class correlation coefficients for fluency measures

Fluency indices	Test Set A			Test Set B			Test Set C		
	ICC	95% Confidence Interval		ICC	95% Confidence Interval		ICC	95% Confidence Interval	
		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd
Filled Pauses (Num)	.96***	.94	.98	.96***	.94	.97	.99***	.99	.99
Unfilled Pauses (Num)	.93***	.90	.96	.99***	.97	.99	.99***	.98	.99
Reformulations (Num)	.85***	.77	.90	.77***	.70	.78	.80***	.76	.81
Repetitions (Num)	.84***	.76	.90	.91***	.86	.94	.87***	.74	.93
Replacements (Num)	.88***	.82	.90	.75***	.61	.83	.86***	.84	.86
Hesitations (Num)	.83***	.74	.89	.91***	.86	.94	.93***	.90	.98
False starts (Num)	.77***	.65	.85	.76***	.68	.78	.80***	.78	.87

Note. *** $p < .001$

Table 22

Intra-class correlation coefficients for accuracy measures

Accuracy indices	Test Set A			Test Set B			Test Set C		
	ICC	95% Confidence Interval		ICC	95% Confidence Interval		ICC	95% Confidence Interval	
		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd
Error-free clauses (Num)	.91***	.87	.94	.89***	.87	.90	.90***	.87	.94
Lexical errors (Num)	.86***	.78	.90	.82***	.83	.87	.88***	.84	.95

Note. *** $p < .001$

Table 23

Intra-class correlation coefficients for complexity measures

Complexity indices	Test Set A			Test Set B			Test Set C		
	ICC	95% Confidence Interval		ICC	95% Confidence Interval		ICC	95% Confidence Interval	
		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd
AS-units (Num)	.95***	.91	.97	.93***	.89	.96	.89***	.86	.90
Subordinate clauses (Num)	.92***	.88	.95	.88***	.77	.90	.85***	.80	.90

Note. *** $p < .001$

Table 24

Intra-class correlation coefficients for CAF ratings

CAF ratings	Test Set A			Test Set B			Test Set C		
	ICC	95% Confidence Interval		ICC	95% Confidence Interval		ICC	95% Confidence Interval	
		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd		Lower Bnd	Upper Bnd
Fluency	.80***	.77	.84	.88***	.85	.89	.84***	.78	.87
Accuracy	.83***	.80	.87	.92***	.90	.98	.89***	.87	.94
Complexity	.85***	.77	.90	.83***	.80	86	.85***	.80	.89

Note. *** $p < .001$. Ratings are based on Elder & Iwashita's (2005) CAF rubric.

3.2.2 Factor analysis

Prior to examining participants' oral performance in detail, I inspected how the qualitative measures represent distinct dimension of *fluency*, *accuracy*, and *complexity.* I collapsed the datasets from the three test sets for the current analysis as well as the analyses presented in the following sections. I conducted nine separate principal component analyses (PCA) for each planning condition (and further broken down by three test-task types) on the 19 CAF measures (for which I quantified the raw coding data). For all nine analyses, the Kaiser-Meyer-Olkin (KMO) measure ensured the sampling adequacy (i.e., sufficient number of observations for reliable analysis): the KMO values ranged from .649 (UG condition, IP task) to .728 (UG condition, IT-L task), which was above the threshold limit of .5 (Field, Miles, & Field, 2012). The Barlett's test of sphericity for all nine analyses were statistically significant ($p < .001$), demonstrating acceptable degree of correlations among variables for running PCAs.

The factor structures were markedly similar across all nine conditions; thus, I present here the results with the highest KMO value: UG condition, IT-L task. As in Table 25, Factor 1 mostly encompassed the *breakdown fluency* measures. Factor 2 showcased an interesting mix of features of all three dimensions. *Speed fluency* measures all clustered on Factor 2, with an addition of *mean length of run* displaying moderate strength of factor loading. All *accuracy* measures loaded on to Factor 2 with high factor loadings. Additionally, *subordinate clauses*, which is a sub-dimension of *syntactic complexity*, displayed moderate factor loadings to Factor 2. Factor 3 consisted of *repair fluency* measures, with a moderate factor loading of *time spent before articulation.* Factors 4 and 5 each included *syntactic complexity* and *lexical diversity* measures, respectively.

Overall, PCA informed that the measures adopted in the present study mostly conformed

to the overall CAF conceptualizations. At the same time, *accuracy* measures indicated a possible link within and across sub-dimensions.

Table 25

Factor analysis for IT-L task under UG planning condition

Measures	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Tot. Num. of Syllables		.876		.430	
Speech rate		.885			
Time spent before articulation		-.560	-.402		
Filled pauses per minute	-.607				
Unfilled pauses per minute	-.895				
Mean length of run	.810	.440			
Repetitions ratio			.681		
Replacements ratio			.707		
Reformulations ratio			.818		
Hesitations ratio			.709		
False starts ratio			.770		
Error free clauses		.773			
Lexical errors per 100 words		-.803			
AS-unit length				.726	
Subordinate clause ratio		.428		.841	
Tot. Num. of unique words		.524		.793	
Type-token ratio				.771	-.414
Conjunctions					.740
Sentence Linking					.835

Note. Bartlett's test of sphericity: $\chi^2(136) = 1344.76$, $p < .001$. Results are based on orthogonal rotation (varimax).

Factor loadings below .30 are not reported.

3.2.3 Descriptive statistics and comparison analysis

In this section, I report the basic descriptive statistics for the 19 CAF measures. I additionally provide comparison statistics (e.g., RM MANOVAs, pairwise comparisons) among the variables of interest.

3.2.3.1 Fluency

3.2.3.1.1 Speed fluency

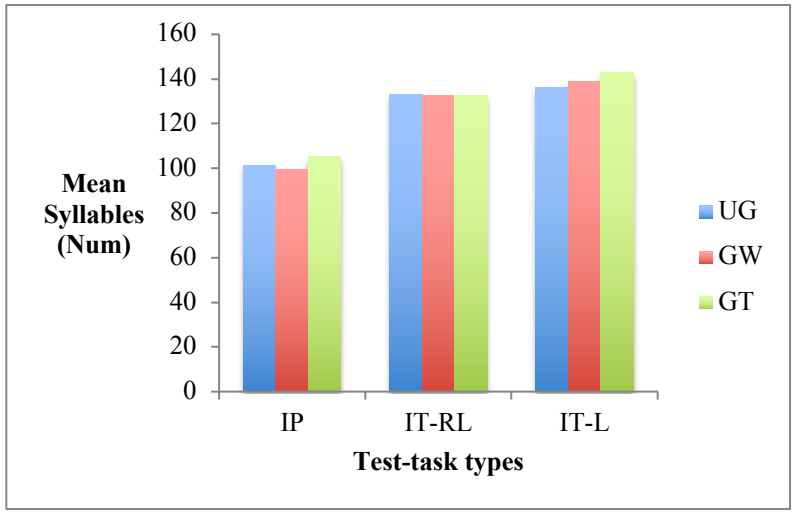
Speed fluency measures included: *total number of syllables per minute* produced by participants (within the response time), *speech rate* (total number of syllables divided per articulation time), and *time spent before articulation* (total amount of silence time before participants started to respond). A general tendency depicted in Table 26 was that across the three planning conditions, participants were able to produce substantially more syllables in the two integrated tasks; within the two integrated conditions, it was under the IT-L task (the integrated task with listening only) that participants spoke more. The speech rate seemed to be slightly higher in the IP task as opposed to the two integrated tasks; this suggests that participants generally had to speak faster in the IP task condition. Interestingly, in all planning conditions, *time spent before articulation* was slightly highest in the IP task condition, while the two integrated tasks did not show much difference between one another. The three bar graphs in Figure 9 additionally depict the trend found in the data: participants spoke more in the two integrated tasks, while taking some more time before responding for the IP task regardless of the three planning conditions.

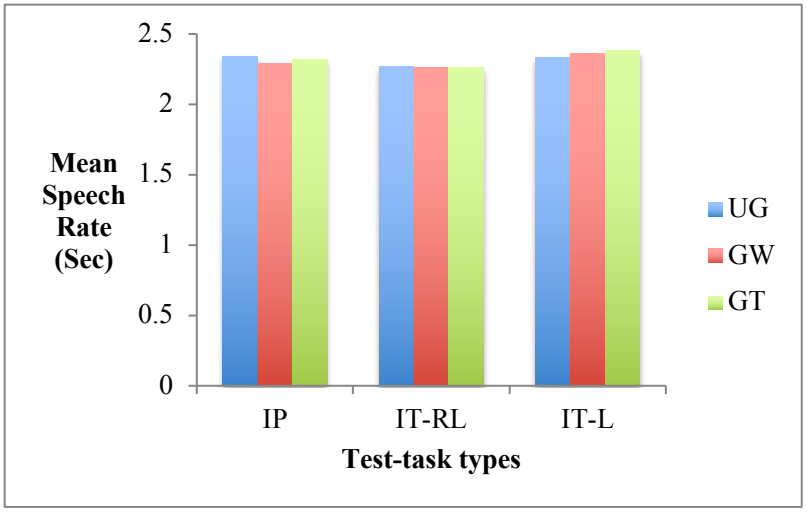
Table 26

Descriptive statistics for *speed fluency* by planning conditions and test tasks

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)			Pairwise significant differences	
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L	Planning types	Task types
Tot. Num. of syllables (Num)	101.46 (24.17)	133.24 (33.40)	136.27 (33.29)	99.47 (25.10)	132.43 (32.72)	138.66 (36.45)	105.13 (23.64)	132.82 (33.79)	142.84 (36.16)	ns	IP < IT-RL IP < IT-L IT-L < IT-RL
Speech rate (Sec)	2.34 (0.54)	2.27 (0.56)	2.33 (0.55)	2.30 (0.56)	2.26 (0.55)	2.36 (0.61)	2.38 (0.52)	2.26 (0.56)	2.32 (0.60)	ns	IT-RL < IP IT-RL < IT-L
Time spent before art. (sec)	1.70 (1.22)	1.33 (0.65)	1.54 (1.19)	1.75 (1.25)	1.44 (0.88)	1.41 (0.71)	1.78 (1.07)	1.33 (0.66)	1.45 (0.99)	ns	IT-RL < IP IT-L < IP

Note. Standard deviations are in parenthesis. Pairwise comparisons (subsequent analysis of RM MANOVA) are based on adjusted alpha level of 0.5/9. Ns indicates no statistically significant difference between paired variables.





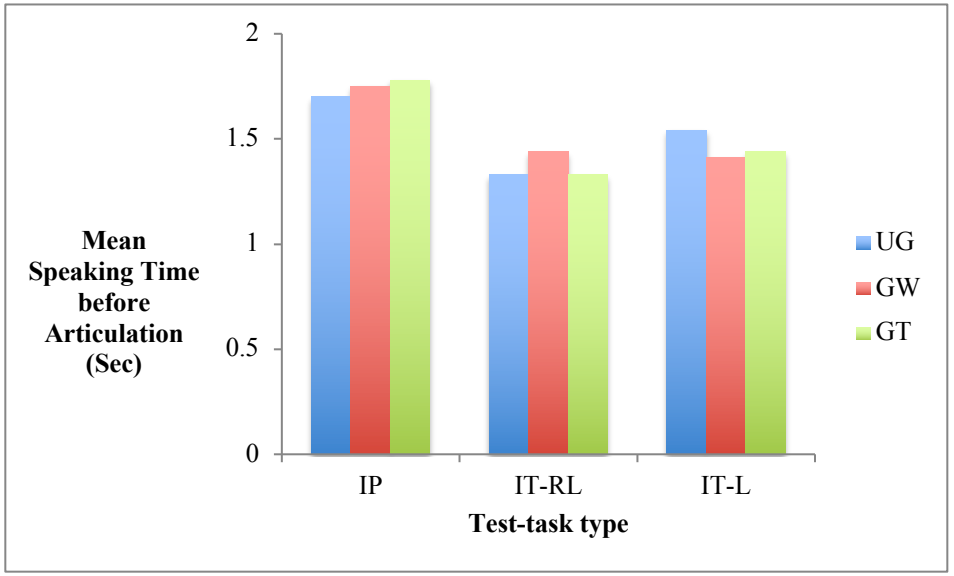


Figure 9 *Speed fluency measures*

A follow-up RM MANOVA revealed that there was no interaction effect between *planning condition* and *test-task types* for *speed fluency* (Greenhouse-Giesser corrections; *total number of syllables*: $F_{4, 582} = 1.641$, $p = .162$, $\eta p^2 = .011$; *time spent before articulation*: $F_{4, 582} = 0.670$, $p = .586$, $\eta p^2 = .005$), although *speech rate* seemed to exert a tendency towards significance ($F_{4, 582} = 2.035$, $p = .091$, $\eta p^2 = .014$). In terms of the main effects, *planning condition* again did not have statistical significance on affecting *speed fluency*. Table 26 reports that pairwise comparisons (using the Bonferroni corrections: adjusted alpha level = .05/9) amongst the three planning conditions in terms of *speed fluency* measures did not have statistically significant pairs. *Test-task types*, on the other hand, had a statistically significant main effect on all three measures (*total number of syllables*: $F_{2, 582} = 469.512$, $p = .000$, $\eta p^2 = .617$; *speech rate*: $F_{2, 582} = 11.516$, $p = .000$, $\eta p^2 = .038$; *time spent before articulation*: $F_{2, 582} = 14.531$, $p = .000$, $\eta p^2 = .048$).

Pairwise comparisons in Table 26 further revealed differences of *speed fluency* measures across the three test tasks, which supported the overall trends displayed in Figure 9. IT-L task exerted the highest *total number of syllables* relative to the IP task (mean difference = 37.310, 95% CI [33.968, .40.650], $p = .000$, $d = 1.23$) and the IT-RL task (mean difference = 6.425, 95% CI [3.503, 9.347], $p = .000$, $d = 0.18$). *Speech rate* was the least in IT-RL task relative to the IP task (mean difference = -.089, 95% CI [-.146, -.032], $p = .001$, $d = 0.11$) and the IT-L task (mean difference = -.113, 95% CI [-.162, -.063], $p = .000$, $d = 0.19$). This indicated that participants took more time in responding to IT-RL tasks. On the other hand, participants took the most time before breaking the silence to respond in the IP task more so than the IT-RL task (mean difference = .339, 95% CI [.187, .492], $p = .000$, $d = 0.35$) and the IT-L task (mean difference = .239, 95% CI [.052, .427], $p = .000$, $d = 0.22$).

3.2.3.1.2 Breakdown fluency

As in Table 27, *breakdown fluency* measures included: *filled pauses per minute*, *unfilled pauses per minute*, and *mean length of run* (total number of syllables divided by total number of unfilled pauses). Again, differences were marginal relative to planning conditions, but noticeable across test-task types. Participants produced relatively small amount of *filled pauses* and *unfilled pauses* per minute, with small differences noticed by planning condition (although the graphs in Figure 10 suggests that *unfilled pauses* seemed to be slightly greater in the GT condition). *Mean length of run* was the highest in the IT-L task; this suggests that participants produced the lengthiest utterance between pause boundaries in the IT-L task.

From a follow-up RM MANOVA, it was found that there was no interaction effect between planning condition and test-task type for *breakdown fluency* (Greenhouse-Giesser corrections; *filled pauses per minute*: $F_{4, 582} = 0.720$, $p = .578$, $\eta p^2 = .005$; *unfilled pauses per minute*: $F_{4, 582} = 0.238$, $p = .853$, $\eta p^2 = .002$; *mean length of run*: $F_{4, 582} = 0.561$, $p = .681$, $\eta p^2 = .004$). Likewise, planning condition was not a statistically significant factor affecting the three *breakdown fluency* measures. *Test-task types* had a statistically significant main effect for *unfilled pauses* ($F_{2, 582} = 45.440$, $p = .000$, $\eta p^2 = .136$) and *mean length of run* ($F_{2, 582} = 14.725$, $p = .000$, $\eta p^2 = .048$), but not for *filled pauses* ($F_{2, 582} = 1.603$, $p = .202$, $\eta p^2 = .006$).

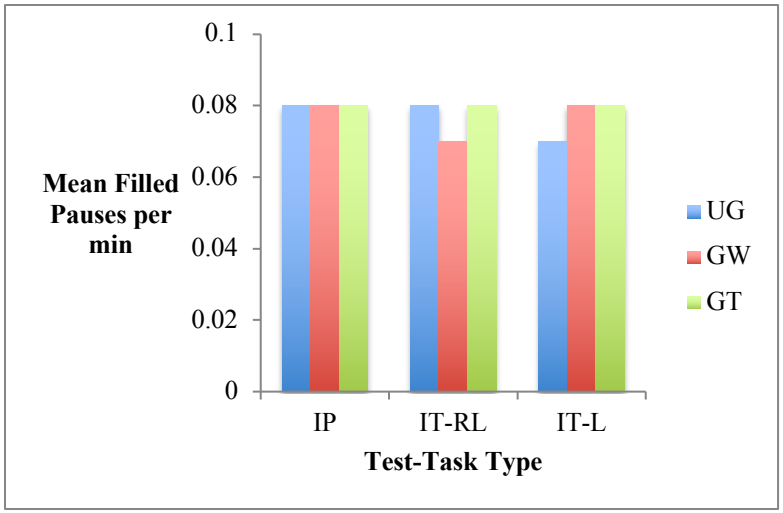
Pairwise comparisons (using the Bonferroni corrections) (see Table 27) within sub-levels of planning condition and test-task types further confirmed that it was in the IT-L task condition that participants produced denser utterance than the IP task (mean difference = 5.379, 95% CI [2.680, 8.078], $p = .000$, $d = 0.32$) and IT-RL task (mean difference = 3.556, 95% CI [1.322, .5.791], $p = .000$, $d = 0.18$).

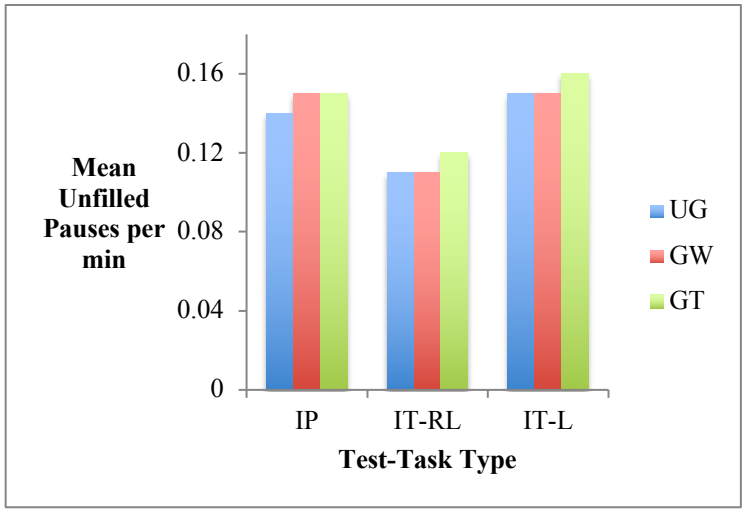
Table 27

Descriptive statistics for *breakdown fluency* by planning conditions and test tasks

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)			Pairwise significant differences	
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L	Planning types	Task type
Filled pauses per minute	0.08 (0.06)	0.08 (0.07)	0.07 (0.06)	0.08 (0.06)	0.07 (0.64)	0.08 (0.07)	0.08 (0.07)	0.08 (0.05)	0.08 (0.06)	ns	ns
Unfilled pauses per minute	0.14 (0.07)	0.11 (0.05)	0.15 (0.08)	0.15 (0.06)	0.11 (0.05)	0.15 (0.09)	0.15 (0.05)	0.12 (0.04)	0.16 (0.09)	ns	IT-RL < IP IP < IT-L
Mean length of run (in syllables)	17.10 (9.71)	19.78 (5.68)	23.32 (6.92)	17.86 (7.61)	19.37 (7.20)	21.47 (7.39)	16.47 (9.53)	17.70 (10.70)	22.85 (10.05)	ns	IP < IT-L IT-RL < IT-L

Note. Standard deviations are in parenthesis. Pairwise comparisons (subsequent analysis of RM MANOVA) are based on adjusted alpha level of 0.5/9. Ns indicates no statistically significant difference between paired variables.





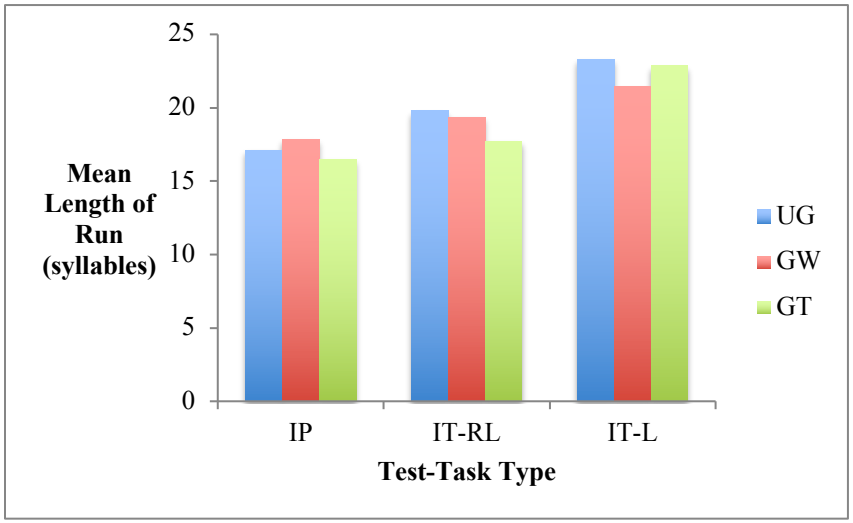


Figure 10 *Breakdown fluency measures*

3.2.3.1.3 Repair fluency

Five measures denoted for the two dimensions of *repair fluency*; namely, *verbatim repair* (*repetitions, replacements*; no linguistic modification on the repaired features) and *substitutive repair* (*reformulations*, *hesitations*, and *false starts*; linguistic modification made on repaired features). I derived the ratio of each repairing phenomenon to the total amount of articulation time for each individual participant. From Table 28, it was apparent that participants did not showcase a great extent of *repair fluency* in their speech; in fact, the values were close to 0, suggesting that participants' speech did not contain a number of repairing phenomenon.

Yet a follow-up RM MANOVA revealed a main effect of planning condition on *repetitions* ($F_{1, 289} = 8.122$, $p = .000$, $\eta p^2 = .053$); at the same time, there was also a borderline effect on *reformulations* ($F_{1, 289} = 2.954$, $p = .054$, $\eta p^2 = .020$). *Test-task types*, on the other hand, was not found to have statistically significant main effect on all *repair fluency* measures. As shown in Table 28, post-hoc pairwise comparisons (Bonferroni corrections applied) on the sub-levels of planning condition revealed that there were small yet significant differences in participants' production of *repetitions* and *reformulations.* In particular, participants tended to produce the most number of *repetitions* in the GT condition relative to the UG condition (mean difference = .008, 95% CI [.003, .013], $p = .000$, $d = 0.02$), and the GW condition (mean difference = .005, 95% CI [.000, .010], $p = .002$, $d = 0.02$). Likewise, participants produced more *reformulations* in the GT condition than the UG condition (mean difference = .004, 95% CI [.000, .007], $p = .047$, $d = 0.01$). On the other hand, *repair fluency* did not vary in accordance to the type of test tasks.

Overall, findings from the *fluency* dimensions denoted a stronger effect of test-task types than the type of planning activities on the amount of language produced and the pausing phenomena.

Table 28

Descriptive statistics for *repair fluency* by planning conditions and test tasks

Measure		UG (N = 98)			GW (N = 98)			GT (N = 98)			Pairwise significant differences	
		IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L	Planning	Task type
Verbatim	*Repetitions*	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.03 (0.02)	UG < GT GW < GT	ns
	Replacements	0.01 (0.02)	0.01 (0.02)	0.02 (0.02)	0.01 (0.02)	0.02 (0.02)	0.01 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	ns	ns
Substitutive	*Reformulations*	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.02)	0.02 (0.01)	0.02 (0.01)	0.02 (0.02)	0.02 (0.01)	0.02 (0.01)	UG < GT	ns
	Hesitations	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	0.01 (0.02)	ns	ns
	False starts	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	0.00 (0.01)	ns	ns

Note. Standard deviations are in parenthesis. Pairwise comparisons (subsequent analysis of RM MANOVA) are based on adjusted alpha level of 0.5/15. Ns indicates no statistically significant difference between paired variables.

3.2.3.2 Accuracy

Measures for *accuracy* included *error-free clauses per AS-unit* and *lexical errors per 100 words*. As seen in Table 29, although with slight differences, the ratio of *error-free clauses* to the total number of clauses was generally the highest in the GT condition. The differences were clearer between the IP task and the two integrated tasks; speech produced in the IP task condition contained the least *error-free clauses*. In terms of *lexical errors*, the trend was that participants produced fewer errors in the integrated tasks than the IP task. However, there was a sudden increase in the lexical errors in the IT-L task under the GT condition. Such a difference in the *lexical errors* is well illustrated in Figure 11.

Subsequently, a follow-up RM MANOVA confirmed a significant interaction effect between planning conditions and test-task types (Greenhouse Giesser corrections: $F_{4, 572} = 3.443$, $p = .006$, $\eta p^2 = .025$). Figure 12 further supports this result by depicting the increase in lexical errors in the IT-L task performed in the GT condition. I subsequently conducted a post-hoc one-way ANOVA with planning condition as an independent variable and *lexical errors per 100 words* for IT-L task as a dependent variable. The finding was that there was a main effect for planning condition ($F_{2, 290} = 9.676$, $p = .000$, $\eta p^2 = .007$). Furthermore, pairwise comparison (with Bonferonni correction) provided that participants made more lexical errors in the GT condition relative to the UG (mean difference = .451, 95% CI [.130, .774], $p = .002$, $d = 0.46$) and the GW conditions (mean difference = .551, 95% CI [.228, .873], $p = .000$, $d = 0.63$).

Table 29

Descriptive statistics for *accuracy* by planning conditions and test tasks

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)			Pairwise significant differences	
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L	Planning	Task type
Error-free clauses ratio	0.39 (0.76)	0.45 (0.67)	0.53 (0.73)	0.42 (0.76)	0.49 (0.57)	0.52 (0.69)	0.35 (0.85)	0.53 (0.72)	0.55 (0.65)	ns	IP < IT-RL IP < IT-L
Lexical errors per 100 words	1.10 (0.97)	1.08 (1.02)	1.15 (1.04)	1.13 (1.27)	1.17 (0.90)	1.22 (0.82)	1.08 (1.00)	1.17 (0.94)	1.47 (0.93)	Within IT-L: UG < GT GW < GT	IP < IT-L

Note. Standard deviations are in parenthesis. Pairwise comparisons (subsequent analysis of RM MANOVA) are based on adjusted alpha level of 0.5/6. Ns indicates no statistically significant difference between paired variables.

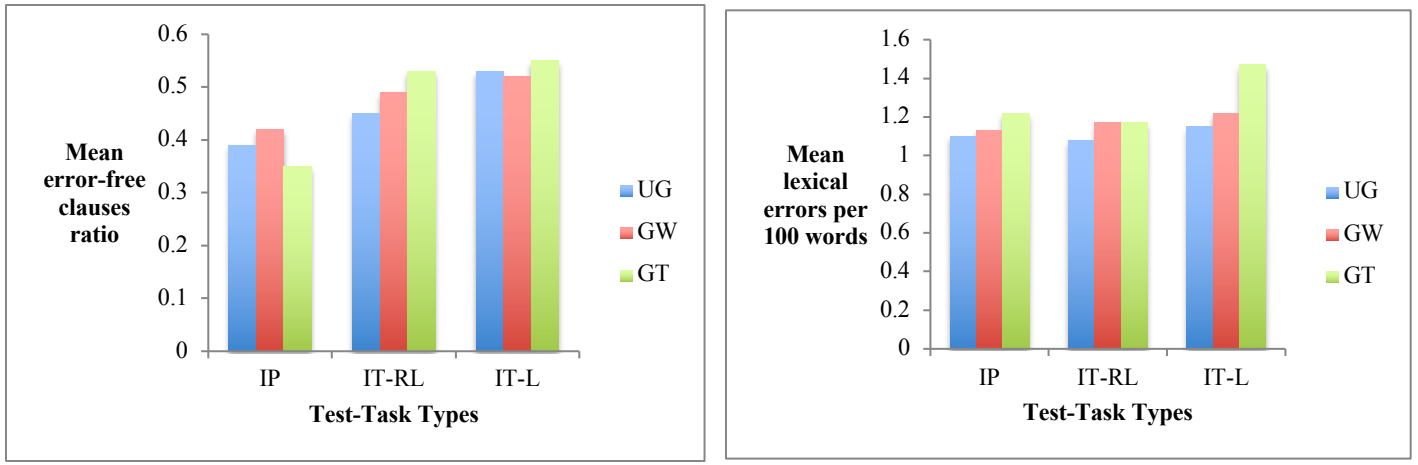


Figure 11 *Accuracy measures*

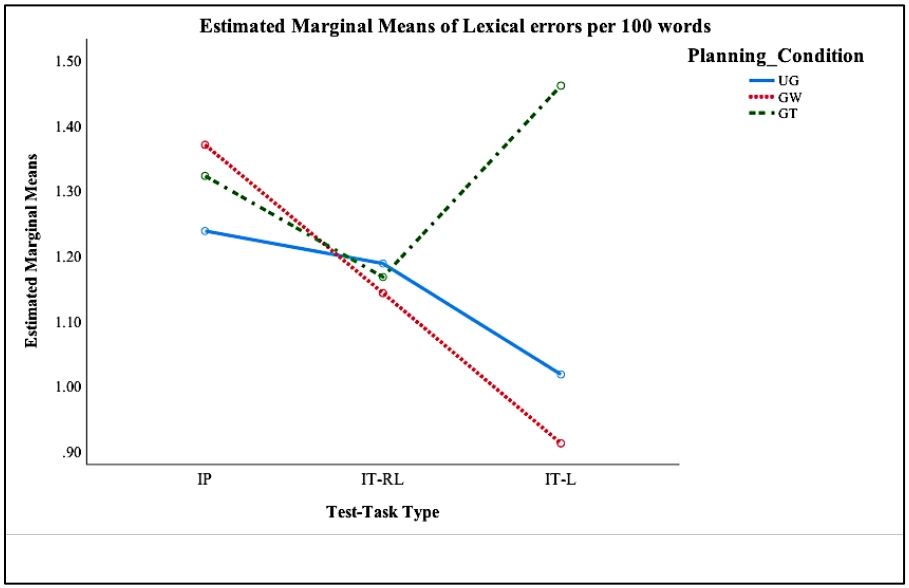


Figure 12 *Interaction effect on lexical error per 100 words*

3.2.3.2 Complexity

3.2.3.2.1 Syntactic complexity

For *syntactic complexity*, I examined *AS-unit length* (i.e., the number of syllables produced within an AS-unit) and *ratio of subordinate clause to AS-unit.* The results summarized in Table 30 seemed to suggest that the *AS-unit length* was the greatest in the GT condition, and the shortest in the GW condition. In particular, it was in the GW condition that the IT-RL task exerted the largest difference in the *AS-unit length* to that of the IT-L task (IT-RL task: M = 32.99, SD = 13.03; IT-L task: M = 38.67; SD = 21.43). In the remaining planning conditions, the differences between the two integrated tasks were less noticeable. The bar graphs depicted in Figure 13 for *AS-unit length* further provided that it was especially within the IT-RL task condition that the three planning conditions differed. There seemed to be a sudden drop in the *AS-unit length* for the GW condition within the IT-RL task. Variations within the three planning conditions in relations to the test-task types were more implied for the results for *subordinate clauses*. As in Table 30, the ratio did not vary to a great extent amongst the three test-task types in the GW condition, while variations were clearer in both UG and GT condition. Figure 13 depicted that similar to the trends in the *AS-unit length*, the differences amongst the planning conditions were identifiable within the integrated task conditions.

A follow-up RM MANOVA exhibited a statistically significant interaction effect between planning condition and test-task type only on *subordinate clauses* (Greenhouse-Giesser corrections: $F_{4, 572} = 2.266$, $p = .038$, $\eta p^2 = .018$). In terms of *AS-unit length*, the RM MANOVA exerted a statistically significant main effect for test-task type (Greenhouse-Giesser corrections: $F_{2, 578} = 12.232$, $p = .000$, $\eta p^2 = .041$). I subsequently performed a post-hoc one-way ANOVA with planning condition as an independent variable and *subordinate clauses* in each three test

tasks as the dependent variables. Planning condition had a statistically significant main effect on the IT-RL task ($F_{2, 293} = 5.893$, $p = .003$, $\eta p^2 = .004$). Pairwise comparisons (with Bonferonni correction) revealed that within the IT-RL task, subordinate clauses were the greatest in the UG condition relative to the GW condition (mean difference = .159, 95% CI [.041, .276], $p = .004$, $d = 0.49$) and the GT conditions (mean difference = .126, 95% CI [.009, .244], $p = .031$, $d = 0.35$). There was a borderline main effect of planning condition on the IT-L task ($F_{2, 293} = 3.050$, $p = .051$, $\eta p^2 = .002$); within the IT-L task, the GT condition facilitated more subordinate clauses than the GW condition (mean difference = .103, 95% CI [-.001, .206], $p = .053$, $d = 0.39$). Figure 14 corroborates the post-hoc analyses in that participants produced more subordinate clauses in the UG condition when responding to the IT-RL task as opposed to the two guided-planning conditions.

Table 30

Descriptive statistics for *syntactic complexity* by planning conditions and test tasks

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)			Pairwise significant differences	
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L	Planning	Task type
AS-Unit length	33.42 (18.97)	37.89 (22.79)	39.50 (19.21)	31.32 (19.16)	32.99 (13.03)	38.67 (21.43)	33.59 (21.01)	38.49 (19.16)	39.73 (16.71)	ns	IP < IT-RL IT -RL < IT-L
Subordinate clause per AS-unit	0.52 (0.53)	0.60 (0.42)	0.55 (0.33)	0.49 (0.46)	0.43 (0.26)	0.47 (0.25)	0.49 (0.42)	0.47 (0.50)	0.58 (0.31)	Within IT-RL: GW < UG GT < UG	ns

Note. Standard deviations are in parenthesis. Pairwise comparisons (subsequent analysis of RM MANOVA) are based on adjusted alpha level of 0.5/6. Ns indicates no statistically significant difference between paired variables.

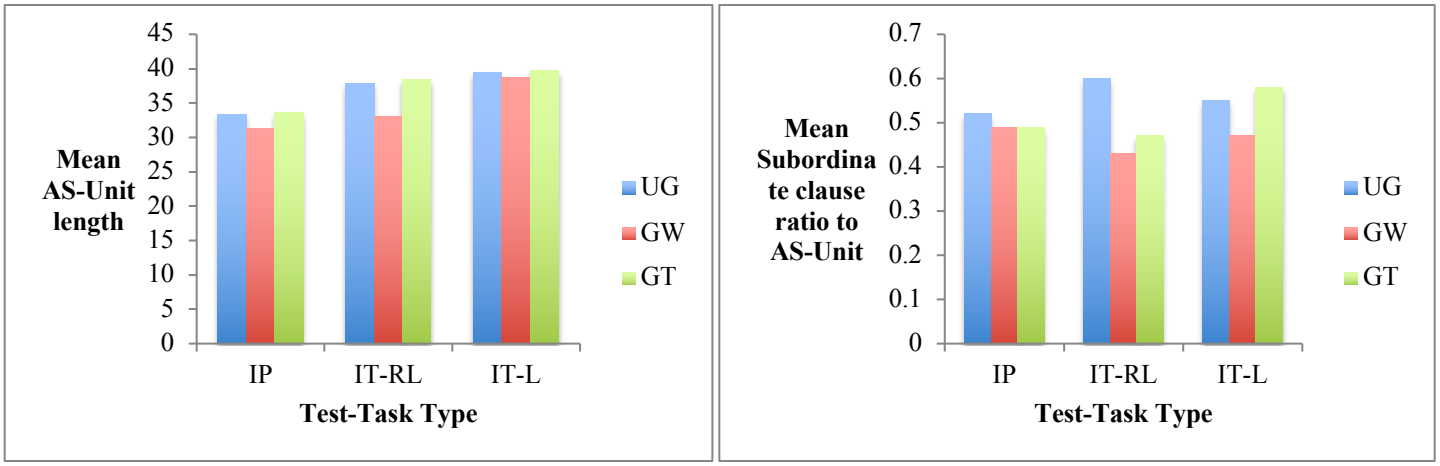


Figure 13 *Syntactic complexity measures*

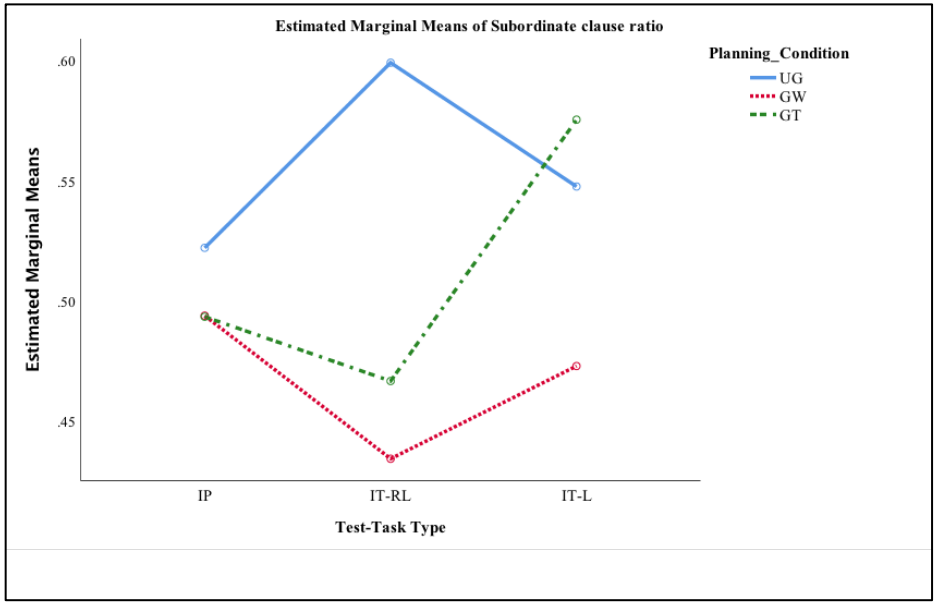


Figure 14 *Interaction effect on subordinate clauses*

3.2.3.2.2 Lexical diversity

For *lexical diversity*, I examined *the number of unique words* (type), *type-toke ratio* (TTR), and the percentage of cohesive devices such as *conjunctions* (e.g., *and*, *but*) and *sentence linking words* (e.g., *nonetheless*, *in other words*) in the speech samples. A general trend depicted in Table 31 and Figure 15 was that *the number of unique words* was the greatest in the IT-RL task regardless of planning conditions. On the other hand, TTR was the highest in the IP condition in all three planning conditions. Given the high number of unique words produced in the integrated tasks, the corresponding lower TTR values could suggest that there were more total number of words generated in these tasks. In terms of the cohesive devices, participants' speeches relatively contained more *sentence linking words* as opposed to *conjunctions*. Differences amongst the three planning conditions seemed to be less noticeable; yet as suggested in Figure 15, there were some extent of variations in IP task.

A follow-up RM MANOVA demonstrated a statistically significant interaction between planning condition and test-task type on *sentence linking words* (Greenhouse-Giesser corrections: $F_{4, 572} = 9.903$, $p = .000$, $\eta p^2 = .059$) (see Figure 16). For remaining variables, there was only a significant main effect of test-task type (*unique words*: $F_{2, 572} = 254.275$, $p = .000$, $\eta p^2 = .467$; *TTR*: $F_{2, 572} = 96.475$, $p = .000$, $\eta p^2 = .250$; *conjunctions*: $F_{2, 572} = 33.076$, $p = .000$, $\eta p^2 = .102$). I subsequently conducted a post-hoc one-way ANOVA with planning condition as independent variable and *sentence linking words* from each test-task type as dependent variables. It was found that a statistically significant main effect of planning condition was only on IP task condition ($F_{2, 291} = 13.421$, $p = .000$, $\eta p^2 = .303$). A pairwise comparison between planning conditions revealed that participants used significantly less *sentence linking words* in the UG

condition relative to the GW (mean difference = -.026, 95% CI [-.038, -.013], p = .000, d = 0.72) and the GT condition (mean difference = -.022, 95% CI [-.034, .-.009], p = .000, d = 0.49).

Table 31

Descriptive statistics for *lexical diversity* by planning conditions and test tasks

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)			Pairwise significant differences	
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L	Planning	Task type
Num. of Unique words	51.77 (10.48)	63.80 (12.31)	58.01 (10.20)	50.59 (9.86)	63.55 (12.10)	57.57 (10.16)	52.51 (9.08)	63.58 (12.59)	58.36 (12.15)	ns	IP < IT-RL IP < IT-L IT-L < IT-RL
Type-toke ratio (TTR)	0.54 (0.06)	0.52 (0.06)	0.49 (0.05)	0.55 (0.08)	0.52 (0.06)	0.49 (0.06)	0.54 (0.07)	0.52 (0.06)	0.48 (0.05)	ns	IT-RL < IP IT-L < IP IT-L < IT-RL
Conjunctions	0.04 (0.02)	0.04 (0.02)	0.04 (0.02)	0.03 (0.02)	0.03 (0.02)	0.04 (0.02)	0.04 (0.19)	0.03 (0.02)	0.04 (0.02)	ns	IP < IT-RL IT-RL < IT-L
Sentence linking devices	0.07 (0.05)	0.08 (0.03)	0.09 (0.02)	0.10 (0.03)	0.08 (0.03)	0.09 (0.02)	0.09 (0.03)	0.08 (0.03)	0.09 (0.02)	UG < GW UG < GT	IT-RL < IP IT-RL < IT-L

Note. Standard deviations are in parenthesis. Pairwise comparisons (subsequent analysis of RM MANOVA) are based on adjusted alpha level of 0.5/6. Ns indicates no statistically significant difference between paired variables.

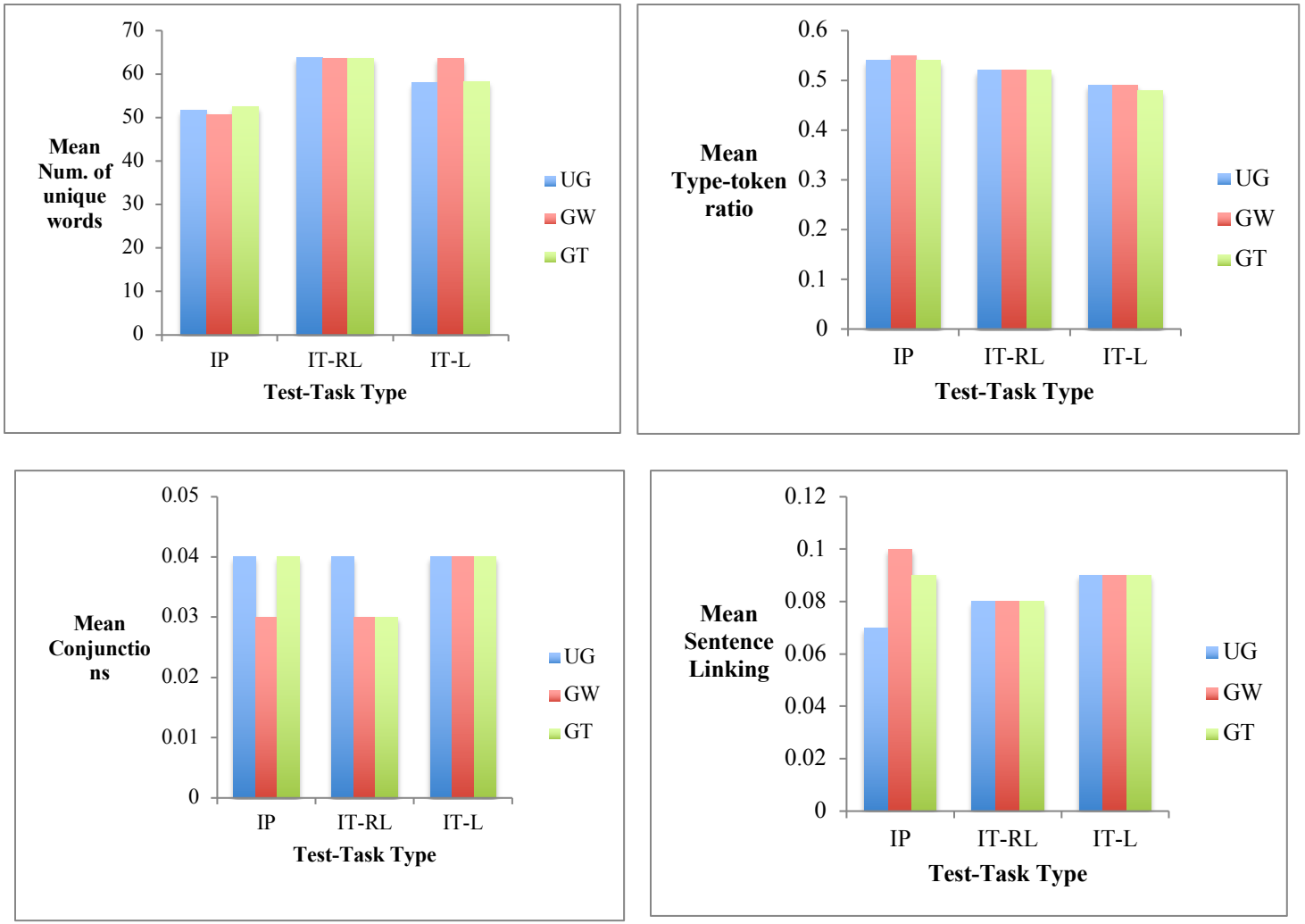


Figure 15 *Lexical diversity measures*

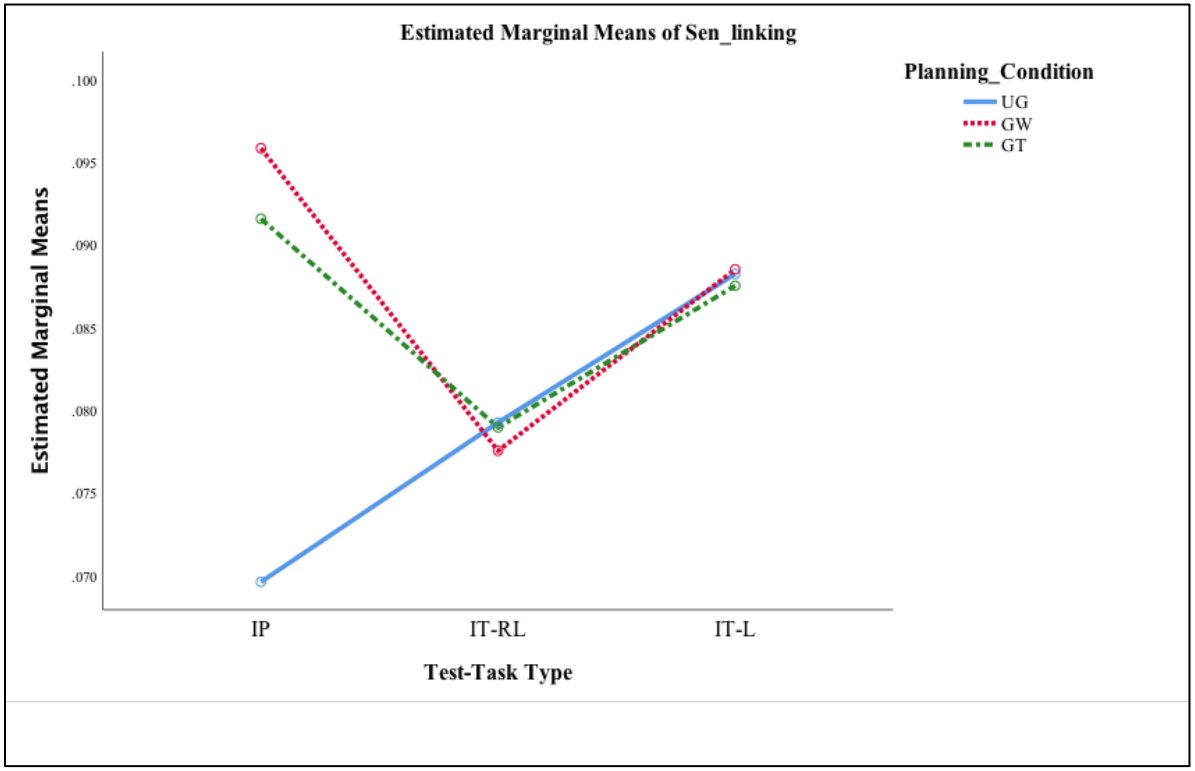


Figure 16 *Interaction effect on sentence linking devices*

3.2.4 Relationship between test scores and speech quality

To explore whether speech quality from each planning condition predicted test performance, I conducted three separate GEE analyses with 19 CAF measures generated from each planning condition as independent variables and level of test performance as binary dependent variable (*low-* and *high-scorers*). Results for *fluency*, *accuracy*, and *complexity* measures are reported in Table 32, 33, and 34.

In terms of *fluency* (see Table 32), *speech rate* measured for three planning conditions consistently contributed to higher test performance to a significant extent (UG: $b = 2.83$, $p = .000$; GW: $b = 4.74$, $p = .014$; GT: $b = 2.77$, $p = .000$); the positive b values suggested that *higher-scorers* produced fast and fluid speech regardless of planning types. In addition, two measures appeared to be significant factors for higher test performance in UG and GT condition; these included *time before articulation* (UG: $b = -0.73$, $p = .047$; GT: $b = -0.31$, $p = .050$) and *mean length of utterance* (UG: $b = 0.08$, $p = .020$; GT: $b = 0.08$, $p = .012$). The negative b value for *time before articulation* indicated that *high-scorers* were less likely to take time before responding. As per *mean length of utterance*, *high-scorers* tended to produce more syllables within an utterance boundary. *Repair fluency* measures were not significant factors leading to higher test scores.

In terms of *accuracy* (see Table 33), *lexical errors per 100 words* predicted higher test scores in all planning conditions (UG: $b = -0.82$, $p = .000$; GW: $b = -0.47$, $p = .010$; GT: $b = -0.46$, $p = .012$). Again, the negative b values in all cases suggested that higher scorers on the speaking test made fewer lexical errors in their speech, and this association appeared to be statistically significant regardless of planning types.

For *complexity* (see Table 34), it was *unique words* that consistently made statistically

significant contribution to test performance (UG: $b = 0.10$, $p = .000$; GW: $b = 0.12$ $p = .000$; GT: $b = 0.10$, $p = .000$). The positive b values implied that higher scorers were likely to use produce more unique words in their speech across all three planning conditions. Notably, it was only within the GW condition that *AS-unit length* appeared as a significant factor impacting on test performance ($b = 0.02$; $p = .024$).

All in all, there were at least one indices within each performance dimension that indicated a trend in significantly predicting test performance regardless of planning conditions: these included *speech rate* from *fluency*, *lexical errors per 100 words* from *accuracy*, and *unique words* from *complexity*.

Table 32

Summary of GEE statistics for *fluency* measures

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)		
	B	*Wald*	*p*	*B*	*Wald*	*p*	*B*	*Wald*	*p*
Tot. Num. of syllables	-0.06	1.75	.186	-0.02	1.46	.227	-0.01	2.28	.131
Speech rate	**2.83**	**23.24**	**.000**	**4.74**	**6.00**	**.014**	**2.77**	**35.23**	**.000**
Time spent before art.	**-0.73**	**3.51**	**.047**	-0.12	1.34	.248	**-0.31**	**3.30**	**.050**
Filled pauses per min.	1.77	0.32	.571	-4.27	2.15	.142	-2.49	0.74	.390
Unfilled pauses per min.	7.83	3.34	.060	-2.48	1.53	.216	-5.84	1.67	.196
Mean length of utterance	**0.08**	**5.39**	**.020**	0.05	2.60	.107	**0.08**	**6.24**	**.012**
Repetitions	1.56	0.03	.864	7.12	0.69	.405	-1.83	0.07	.788
Replacements	-14.40	2.32	.129	-13.85	1.63	.202	-10.32	1.52	.218
Reformulations	-12.16	1.65	.199	-8.94	0.67	.413	-10.21	0.94	.332
Hesitations	9.44	0.73	.392	3.60	0.07	.791	-0.19	0.00	.990
False starts	21.34	2.91	.088	10.15	0.53	.466	-2.27	0.02	.884

Note. Goodness of fit for UG, GW, and GT models are based on the Quasi Likelihood under Independence Model Criterion (QIC). QIC values were 425.04, 423.75, and 427.78 for UG, GW, and GT models, respectively.

Table 33

Summary of GEE statistics for *accuracy* measures

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)		
	B	*Wald*	*p*	*B*	*Wald*	*p*	*B*	*Wald*	*p*
Error-free clauses	0.78	0.91	.763	0.27	0.77	.381	0.48	2.60	.107
Lexical errors per 100 words	**-0.82**	**17.55**	**.000**	**-0.47**	**6.69**	**.010**	**-0.46**	**6.27**	**.012**

Note. Goodness of fit for UG, GW, and GT models are based on the Quasi Likelihood under Independence Model Criterion (QIC). QIC values were 402.13, 410.84, and 414.06 for UG, GW, and GT models, respectively.

Table 34

Summary of GEE statistics for *complexity* measures

Measure	UG (N = 98)			GW (N = 98)			GT (N = 98)		
	B	*Wald*	*p*	*B*	*Wald*	*p*	*B*	*Wald*	*p*
AS-Unit length	0.02	0.06	.805	**0.02**	**5.07**	**.024**	0.02	2.34	.126
Subordinate clause per AS-unit	0.40	1.10	.295	0.69	2.58	.108	0.95	0.92	.338
Unique words	**0.10**	**51.34**	**.000**	**0.12**	**57.54**	**.000**	**0.10**	**42.44**	**.000**
Type-token ratio	-1.24	0.22	.889	0.51	1.50	.221	-1.48	0.53	.465
Conjunctions	0.95	0.02	.035	5.23	0.74	.391	2.25	0.67	.797
Sentence linking	-1.38	0.15	.697	8.32	1.73	.188	-0.67	0.01	.909

Note. Goodness of fit for UG, GW, and GT models are based on the Quasi Likelihood under Independence Model Criterion (QIC). QIC values were 418.05, 418.65, and 415.09 for UG, GW, and GT models, respectively.

3.3 Research question 3: Test-takers' survey responses

For research question 3, I looked at how participants' survey responses differed by planning conditions and test-task types. These responses concerned (a) the participants' confidence in performance in different conditions (*confidence*); (b) the participants' perceptions on the appropriateness of the length of planning time by different test tasks (*appropriateness of planning time*); and (c) the participants evaluation of the effectiveness of the type of planning condition on different test tasks (*effectiveness of type of planning*). I further broke down the results in terms of levels of test performance (*low-scorers*: $N = 49$; *high-scorers*: $N = 50$).

3.3.1 Confidence in performance

Figures 17, 18, and 19 graphically summarize the results for *confidence* (rated on a 5-point scale with 1 being "completely unconfident" to 5 being "completely confident"). Each bar represents frequency counts for a scale category. For the IP task, 40.8% of *low-scorers* felt fairly ($N = 18$) or completely confident ($N = 2$) that they had performed well in the GW condition, while 65.3% of *low-scorers* in the GT condition were fairly ($N = 28$) and completely confident ($N = 4$) in their performance. On the other hand, *high-scorers* only gave the highest confidence rating in the GW condition as opposed to *low-scorers.* For the IT-RL task, the majority gave moderate to high ratings. One difference was that a subset of *low-scorers* tended to give lower ratings in the UG condition ($N = 14$; 28.6%) while such phenomenon for *high-scorers* was noticeable in the GT condition ($N = 13$; 26%). For the IT-L task, *low-scorers* exerted similar patterns across planning conditions. Yet *high-scorers* gave contrasting response patterns for the GW and GT conditions. While vast majority of them strongly felt they performed well in the GW condition ($N = 32$; 64%), their responses were more scattered in the GT condition; their responses for fairly unconfident (rating category "2") were the highest of all planning conditions.

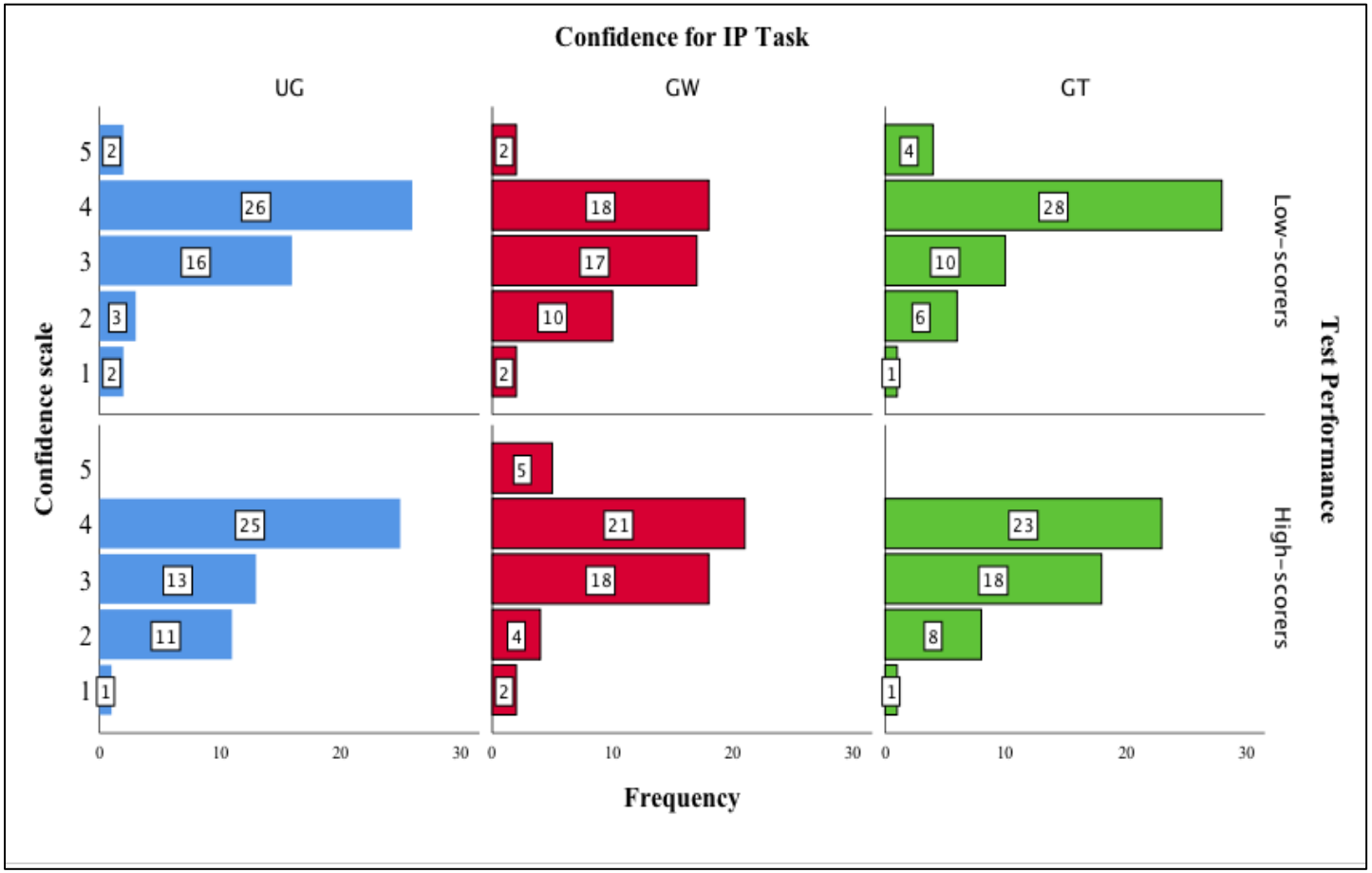


Figure 17 *Confidence ratings for IP task*

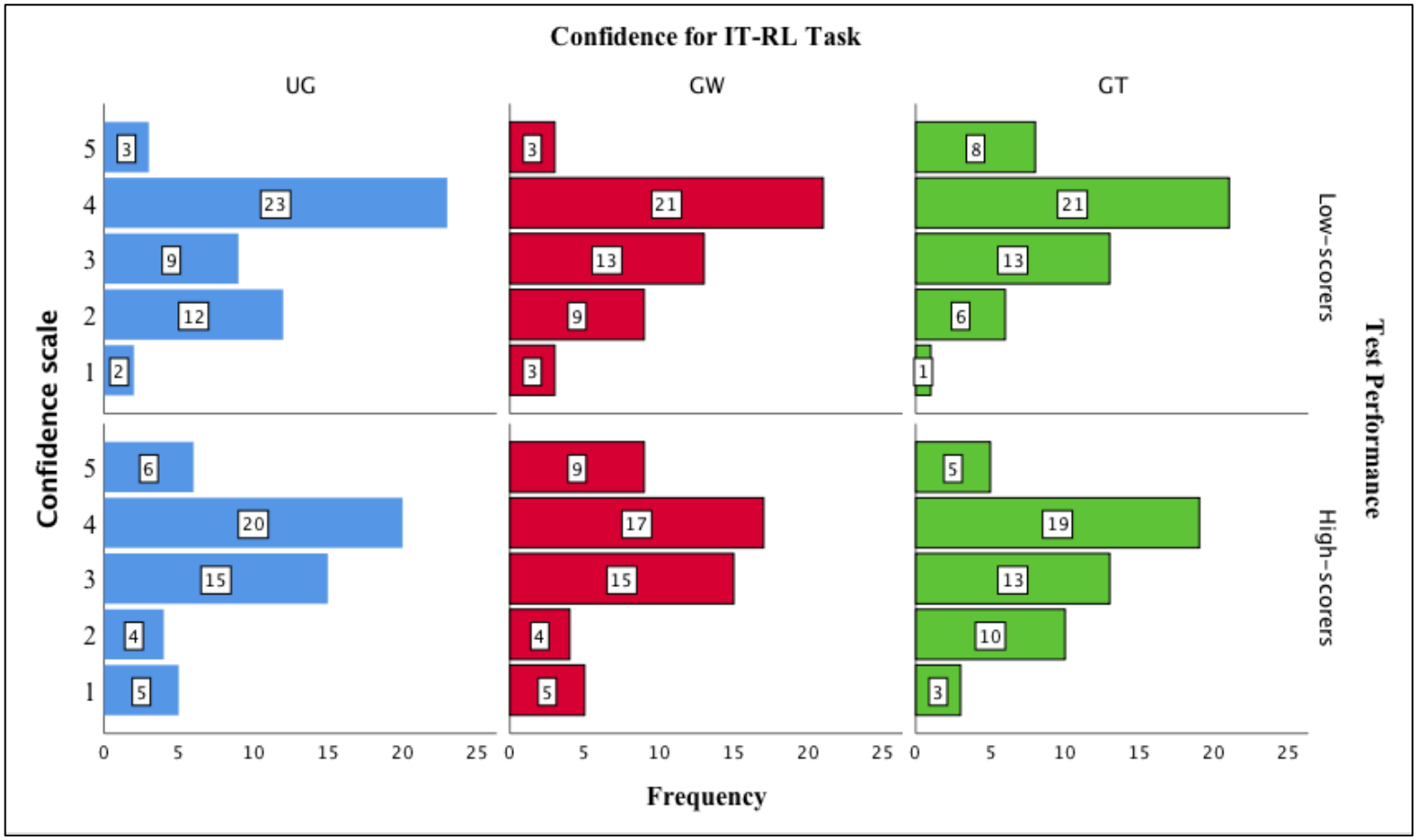


Figure 18 *Confidence ratings for IT-RL task*

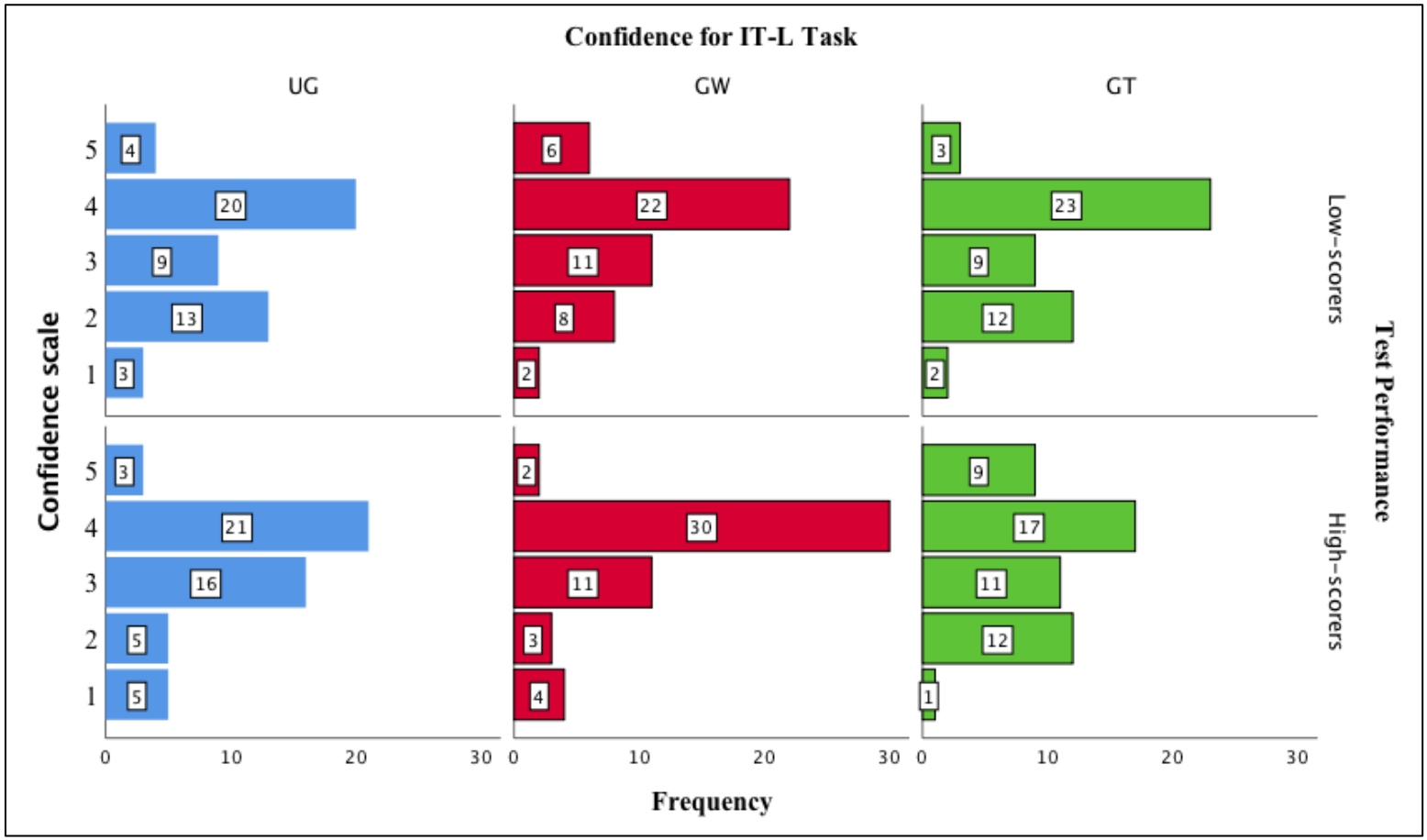


Figure 19 *Confidence ratings for IT-L task*

3.3.2 Appropriateness of planning time

In terms of *appropriateness of planning time*, Figures 21 and 22 demonstrated that participants' responses generally clustered around higher ratings of appropriateness for the two integrated tasks. On the other hand, as in Figure 20, responses were more likely to disperse for the IP task; more specifically, increased responses on fairly inappropriate ("2" rating) to neutral ("3" rating) in the GW condition were noticeable for both *low-* and *high-scorers.* This indicated that both groups of participants felt that planning time for the IP task was not sufficient especially under the GW condition. For *low-scorers*, stark differences in perceptions were identifiable in the GT condition; these participants thought that planning time in the IP task is both fairly appropriate (N = 21; 42.9%) and fairly inappropriate (N = 21; 42.9%). For the IT-RL task, the majority of *high-scorers* strongly agreed in the GW condition that planning time was fairly sufficient (N = 30; 60%). In fact, this trend of tight clustering of higher ratings was found from both groups in both GW and GT conditions. On the other hand, *low-scorers* tended to give diverse ratings under the UG condition; they gave contrasting responses that planning time was both fairly inappropriate (N = 17; 34.7%) and fairly appropriate (N = 20; 40.8%). For the IT-L task, response patterns were quite similar from both groups across planning condition; yet *low-scorers* strongly felt that planning time felt sufficient under the GW condition (N = 30; 61.2%).

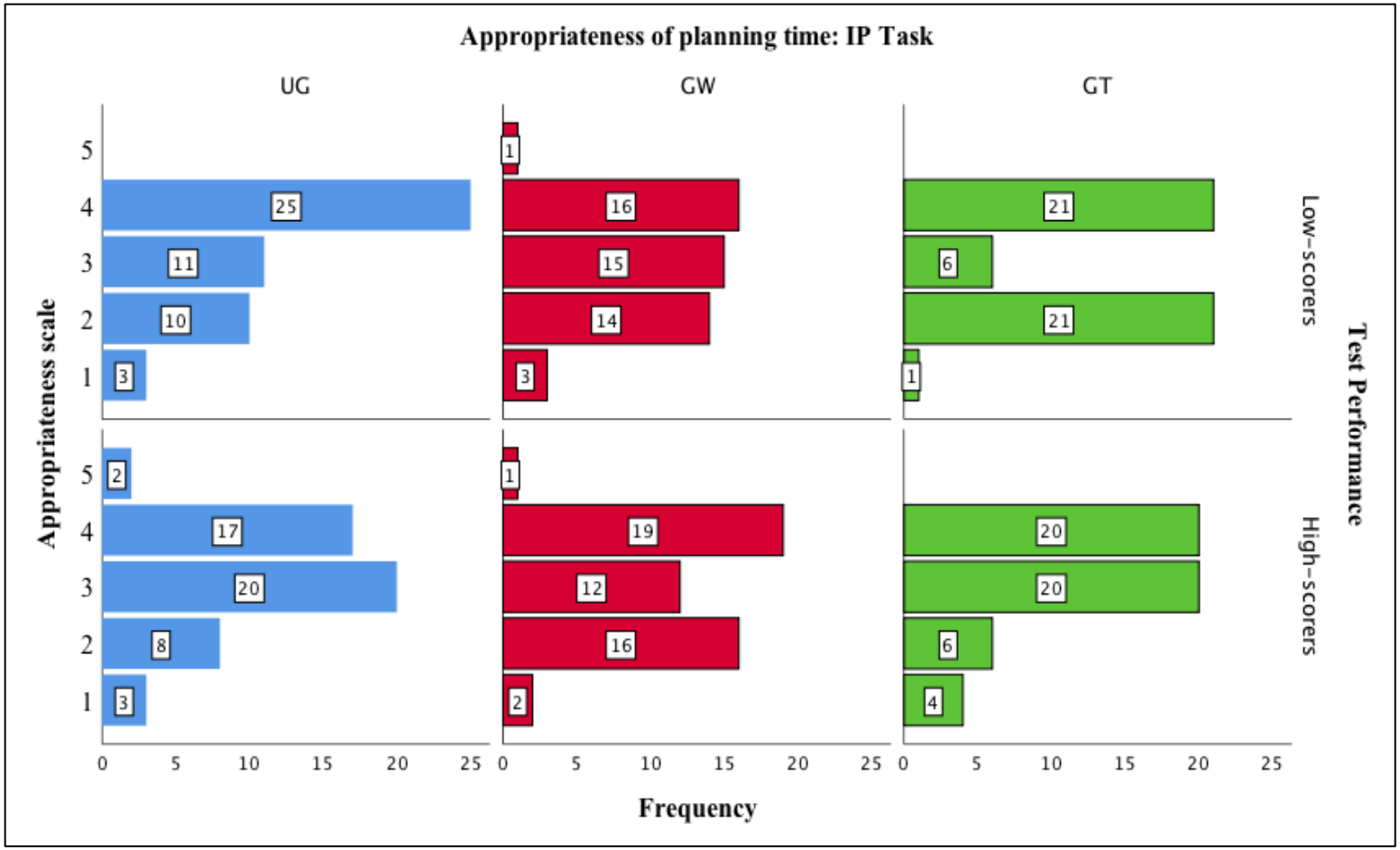


Figure 20 *Appropriateness of planning time in IP task*

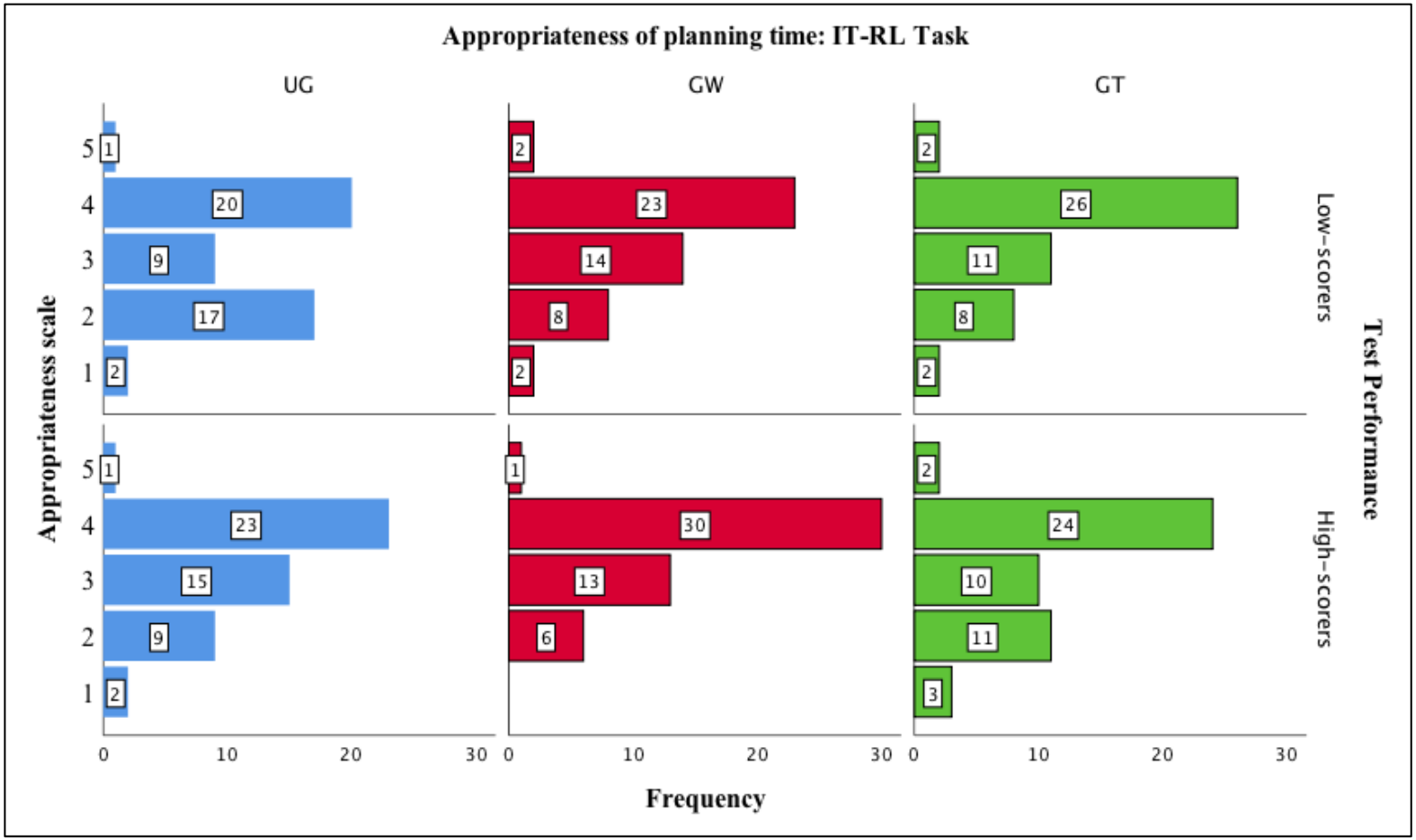


Figure 21 *Appropriateness of planning time in IT-RL task*

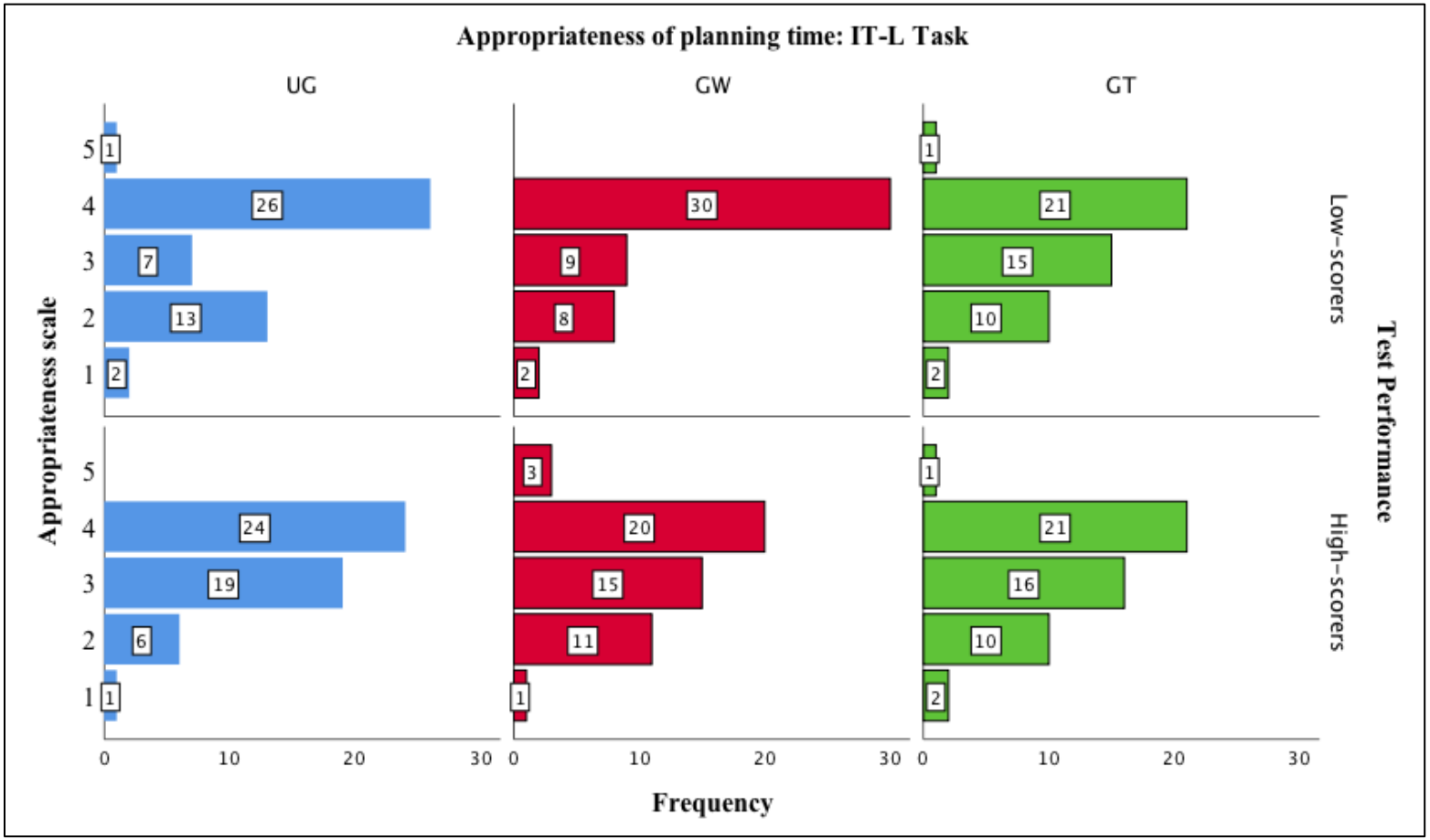


Figure 22 *Appropriateness of planning time in IT-L task*

3.3.3 Effectiveness of types of planning

As depicted in Figures 23, 24, and 25, participants' response patterns for effectiveness of types of planning were generally similar across test-task types. In most cases, participants gave moderate to high ratings of effectiveness regardless of planning conditions. Interestingly, high-scorers' ratings on "5" slightly increased from the IP (N = 10; 20%) to the IT-L (N = 16; 32%) task for GW condition. At the same time, relatively fewer high-scorers seemed to be completely satisfied by GT condition when performing the IT-RL task (N = 6; 12%), while more of them favored the condition when performing the IP (N = 13; 26%) and IT-L task (N = 11; 22%).

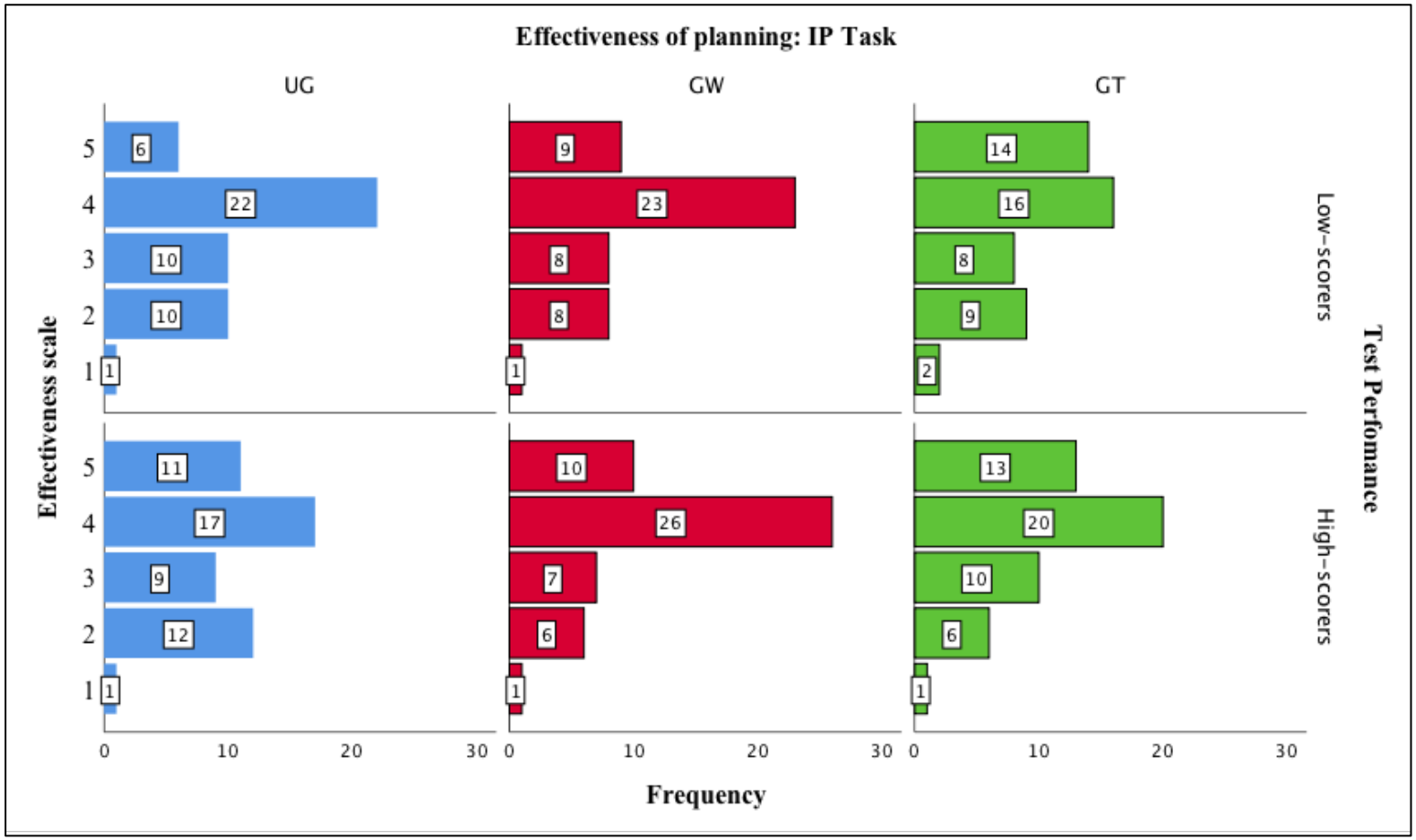


Figure 23 *Effectiveness of planning for IP task*

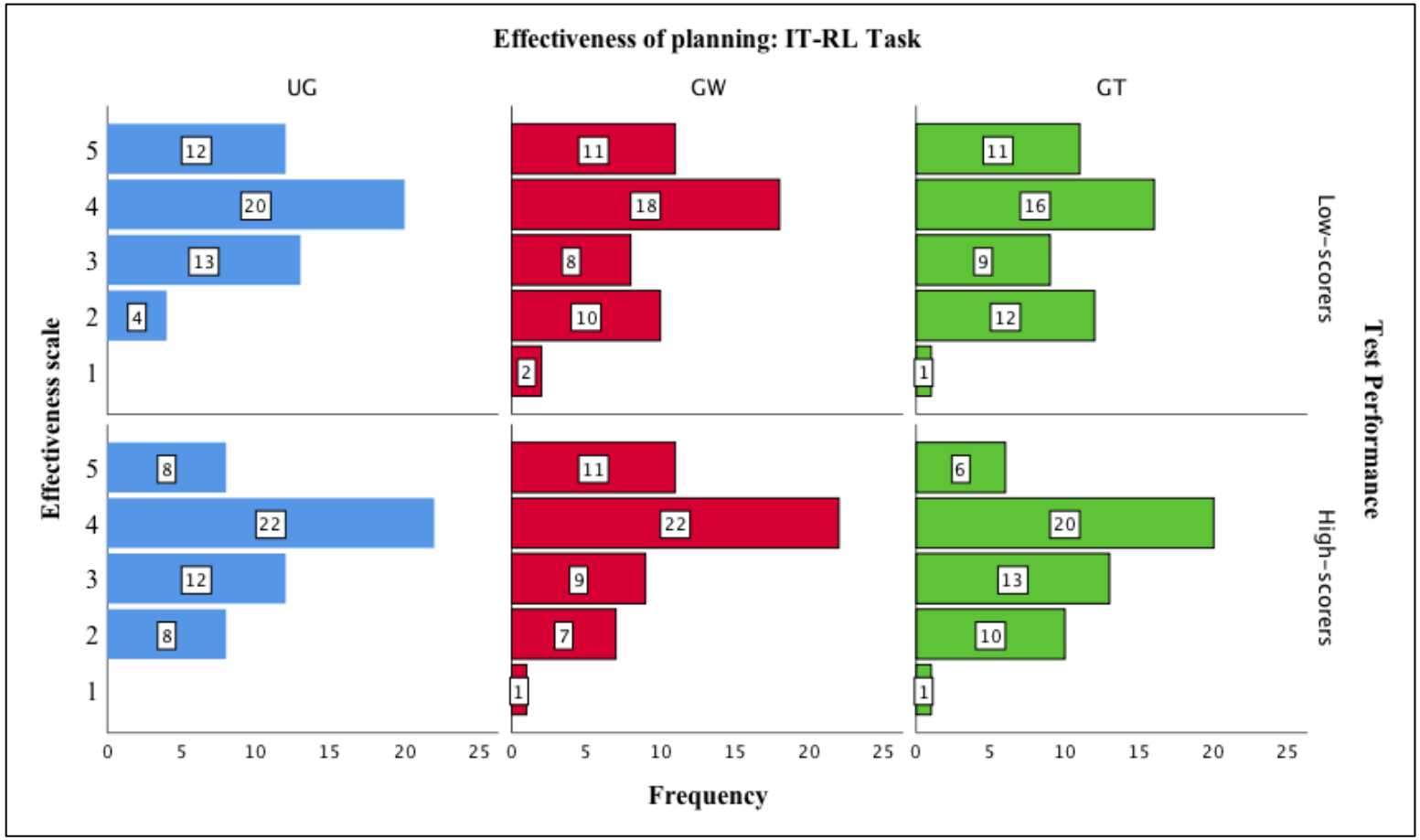


Figure 24 *Effectiveness of planning for IT-RL task*

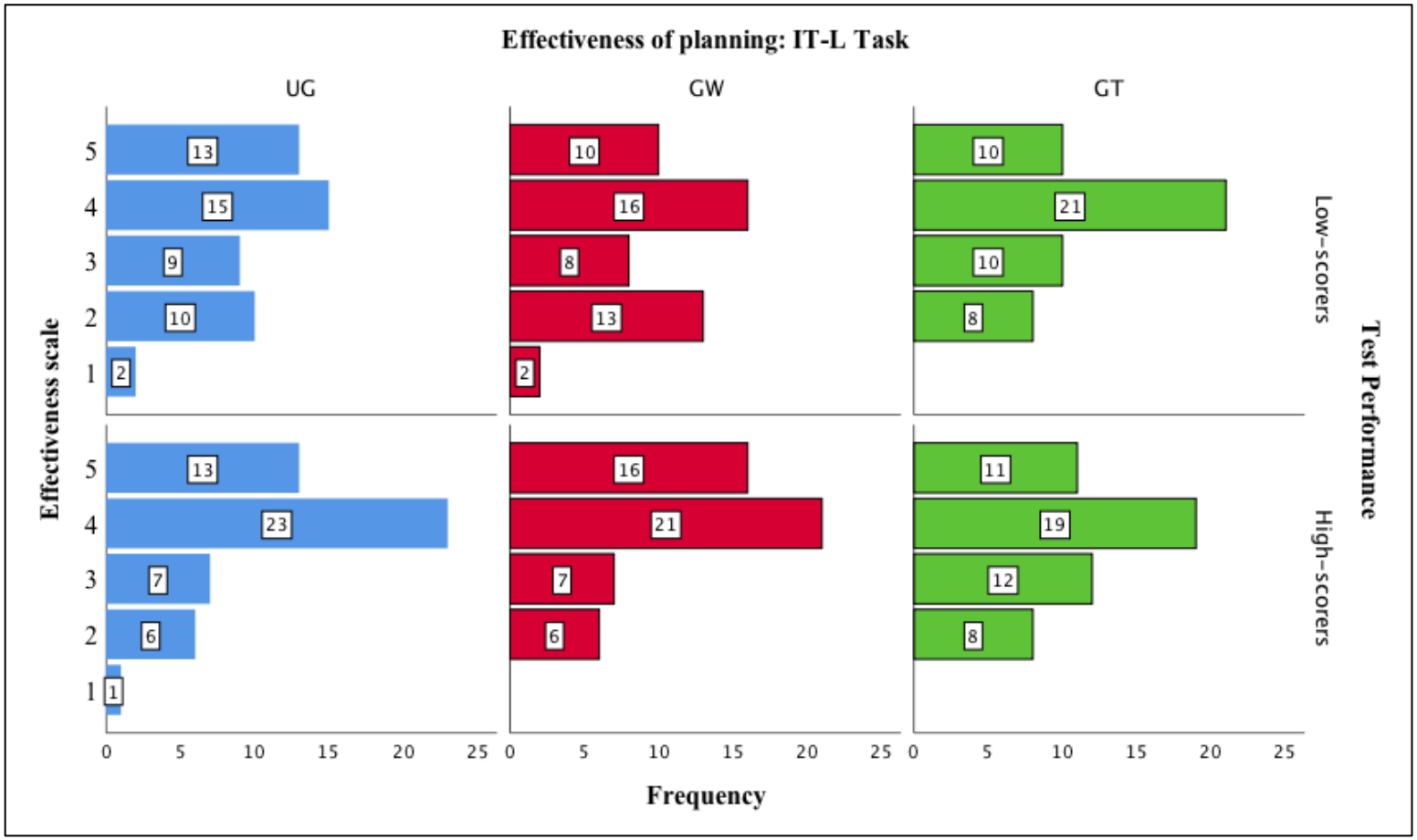


Figure 25 *Effectiveness of planning for IT-L task*

3.3.4 Perceptual differences by planning and task type

To further reveal differences in perceptions by testing conditions, I conducted a repeated measures (RM) MANOVA with three survey response categories as dependent variables, test-task type as within-subjects variable, and planning condition and test performance level as between-subjects variables. There was a statistically significant interaction effect between planning condition and test-task type on *appropriateness of planning time* (Greenhouse-Geisser correction: $F_{2, 538} = 2.84$, $p = .027$, $\eta p^2 = .016$) and a borderline effect on *effectives of type of planning* ($F_{2, 538} = 2.30$, $p = .060$, $\eta p^2 = .015$). On the other hand, there were no factors contributing significantly to *confidence* ratings.

To follow-up on the main effect found from RM MANOVA results, I conducted a series of Wilcoxon paired samples tests for *appropriateness of planning* and *effectiveness of planning*, while comparing the results by planning condition. In terms of *appropriateness of planning*, the finding was that it was only within the two guided planning conditions that participants had differences in perceptions. Upon performing within the GW condition, participants regarded the planning time on the IP task was relatively less sufficiently given than that of the IT-RL task ($Z = -3.56$, $p = .000$) and the IT-L task ($Z = -2.58$, $p = .010$). Similarly, within the GT condition, *appropriateness* rating for planning time of the IP task was lower than that of the IT-RL task ($Z = -2.33$, $p = .020$) and the IT-L task ($Z = -2.15$, $p = .025$). There were no differences between the two integrated tasks on participants' perceptions. The results corroborated the graphical results illustrated in Figures 21 and 22.

For *effectiveness of planning*, it was only within the GT condition that participants had different perceptions with regard to how the planning condition effectively contributed to the performance of different test-task types. For instance, participants thought that the GT condition

was more helpful when performing the IP task (Z = -2.47 p = .013) and the IT-L task (Z = -2.17, p = .030) relative to the IT-RL task. This result aligned with the trend depicted in Figures 23 and 25. Participants' perceptions on *effectiveness of planning* did not differ across test-task types for both UG and GW conditions.

Overall, the results indicate that participants (both low and high-scorers) had similar opinions across the three categories of survey questions. They were mostly confident in their performance and agreed on a moderate to high effectiveness of all planning conditions. In addition, they considered that IP tasks were in need of more planning time than the integrated tasks.

3.4 Research question 4: Test-takers' interview responses

To address research question 4, I individually interviewed participants for gauging their perceptions on the effectiveness of the planning conditions (type and given amount of time) in association with test-task types. In this section, I focus on reporting the response patterns pertaining to two interview questions from which the most unexpected, yet interesting responses were generated. I further break down the results by the two sub-groups of high- and low-test performance. It should be noted that the results are reported with 90 participants' interview responses (due to missing audio files of 4 participants, and inaudible quality of audio files of 5 participants).

3.4.1 Interview question 1: Which type of planning was helpful for you when responding?

With Interview question 1, I opted for gauging participants' thoughts on how each planning condition was helpful for them during test taking. As reported in Table 35, 57.8% of all participants (N = 52), which included the most of both *low*- and *high-scorers*, favored the GW

condition. On the other hand, neither UG nor GT conditions were mentioned greatly by most of the participants. Interestingly, about a quarter of all participants did not opt for a specific planning condition. About 24% of participants (N = 22) agreed that planning condition is effective when used for a particular test-task type, but not useful for another. A small subset of participants had contrasting opinions: 7.78% of participants (N = 7) thought that all three planning types were helpful while 6.64% of participants (N = 6) thought none were effective.

Table 35

Frequency of comments on helpful planning conditions

Category	Number of comments		
	Low-scorers (N = 44)	High-scorers (N = 46)	Total
GW condition	24	28	52
Depends on test-task type	12	10	22
All were helpful	2	5	7
None were helpful	4	2	6
UG/GT condition	1	1	2
Total	44	46	90

Participants favoring the GW condition mentioned that GW allowed them to use writing strategies that they were mostly familiar with or had been using in most testing contexts. Although participants did not have extensive experiences in taking oral language tests, GW condition made it possible for them to transfer their strategies in non-testing situations (e.g.,

regular studying, reading, etc.) as well as conventional paper-based testing settings. This is well illustrated in participant 2077's response below:

> [Participant 2077, female, *low-scorer*]
> When I am taking reading tests or just generally any paper-based tests, I like to write down key words or paraphrased sentences of the given text. This gives me a complete sense of understanding the text. I think I was trying to do the same thing in this speaking test, which helped me organize the information given from the question.

In a similar vein, because most participants were used to scribbling or making notes in non-testing situations, they found it rather difficult *not* to do so in the GT condition. Participant 2068's response gave a good contrast of the two conditions. He mentioned that GT is only helpful for retrieving information given in the tasks, while GW facilitates language production.

> [Participant 2068, male, *high-scorer*]
> I personally keep a diary and like to organize my thoughts by writing them out. I think this made me kind of not like the GT condition. Although GT was helpful in tracing back to the information I was given from the task, I don't think it helped me be prepared of the type of language I would use for responding. Recalling the information does not necessarily lead to fluent speaking, I would say. In GW, it was helpful to read off of the notes I took during planning time when responding. Even just briefly looking at the key words I wrote down helped me think of what I would say next.

Yet some participants assessed the effectiveness of planning by certain test-task types. For instance, for Participant 2065, GW was not effective in the IP task condition due to the short amount of time given. For him, there was not sufficient time to make written planning; instead, GT was found more useful.

> [Participant 2065, male, *high-scorer*]
> There was just simply not enough time to write down what I would like to say. I think there must be more effective ways to make notes, or it may be that I am just a slow writer. This is why I think I did better in the GT condition for the IP task. It helped me better to get at the two points I would like to raise.

Participant 2013 thought that the abstractness of the statements given in the IP task makes it difficult to come up with concrete written planning.

> [Participant 2013, Female, *high-scorer*]
> I think the information given in the IT-RL or IT-L tasks are pretty much structured: statement of the problem, and the two interlocutors' opinions on that. IP task is not structured in any way because it deals with abstract statements that you have not thought of before; almost like whether you like your mom or dad better. Because of this nature, you need to think hard about your own opinion on that, and even good supporting examples or reasons, which is quite difficult to achieve within 15 seconds. By the time you are ready to jot down some ideas or even key words for reference, the planning time is up.

In a similar vein, some participants discussed about the effectiveness of GW and GT with regard to the availability of the input resources in the IP tasks versus the two integrated tasks. For instance, GW aligned to the goal of the two integrated tasks: to be able to make use of the key words and to rightfully integrate the information in the response. For that, a certain extent of writing was needed to accurately capture the key information from the input source. For Participant 2083, GW was more useful for IT-RL task in making a visual map that structured the given information. The graphic information helped him on efficiently conveying a good chunk of information within the limited time.

> [Participant 2083, male, *high-scorer*]
>
> Because you get a lot of information from IT-RL task, I always found it better to first re-structure the information in the following order: the issue at hand posed by the University, and the woman's (or the man's) opinion on that. I found it useful to use graphic organizers for that. When responding, that graphic information made it easy for me to give a structured response. I would not be able to achieve that by merely outlining and summarizing the information in my head.

On the other hand, participants who had mixed feelings toward the planning conditions discussed about the provision of guidance for planning. Those who favored the guided conditions thought that it is especially helpful for novice test takers. These participants, as in the responses of Participants 2087 and 2031, seemed to appreciate even a brief statement given to them before taking the test.

[Participant 2087, male, *high-scorer*]

I personally never prepared for TOEFL or any equivalent test. I think this makes it more appreciative to the fact that there were some outlines given for us for planning. I don't know if it had actually improved my speaking or not, but it did give me a sense of feeling that the test developers are actually caring about the test takers. I always thought that they would want us to fail to some extent so that we could keep on taking the tests.

[Participant 2031, female, *high-scorer*]

Overall, the guidance was helpful for me to realize certain strategies when taking speaking tests. I never thought of thinking silently would help better organize the scrambled information I have in my head. I always thought I needed to grab a piece of paper and make a script out of what I would say, and read it off from it; especially for presenting something in English. But it did actually work when I had tried it.

Yet there were others that were not affected by the guidance provided for planning. Participant 2055, for instance, thought that it would only affect novice test takers, and not more experienced ones. Ultimately, she thought that test performance is not a matter of planning the response well or not; it is simply a matter of whether one has higher proficiency in English speaking.

[Participant 2055, female, *low-scorer*]

You would have already internalized a strategy of your own regardless of elaborated instruction on planning. Of course, I personally would benefit from lengthened time to think before I start speaking, but I would think those who are experienced and are proficient would not even need so much time either. They would be capable in making something up as they go.

Overall, as opposed to the results reported for research questions 1 and 2, writing as the planning seemed to be favored by participants; yet still a good number of participants thought the effectiveness of a particular planning is enhanced when used for a certain test-task type. Again, this result coincided with the findings reported in previous sections that directs to the overall impact of test-task characteristics on both test performance and speech quality.

3.4.2 Interview question 2: Were the times allotted for planning sufficient to you?

From interview question 2, participants overwhelmingly came to a consensus with which the findings in the previous section coincided: there was a strong influence of test-task characteristics. Figure 26 demonstrates such a trend (the number of responses are accounted for overlapping responses; *low-scorers*: $N = 44$; *high-scorers*: $N = 46$). In general, more than half of the two sub-groups of participants agreed that integrated tasks are given sufficient, or perhaps given too much, planning time; this was slightly more geared towards the IT-RL task ($N = 53$; 58.9% of participants). On the other hand, the majority of participants claimed that planning times were substantially lacking for the IP task ($N = 65$; 72.2% of participants). When making the sub-group comparison, *low-scorers* ($N = 23$; 52.3% of *low-scorer* participants) found IT-L task given the longest planning time relative to *high-scorers* ($N = 14$; 30.4% of *high-scorer* participants). *Low-scorers* ($N = 36$; 81.8% of *low-scorer* participants) were more likely to think

that IP task lacked planning time than *high-scorers* (*N* = 29; 63% of *high-scorer* participants). For these *low-scorers*, IT-L task shared some similarities with IP task in that it asks their opinions about the discussed issue in the listening dialogue; therefore, they thought more planning time should be allotted to IT-L task than IT-RL task, which does not require any elaboration on one's opinion. Overall, it was apparent that participants were agreeing that IP tasks need more planning time, while integrated tasks could be given lesser planning time.

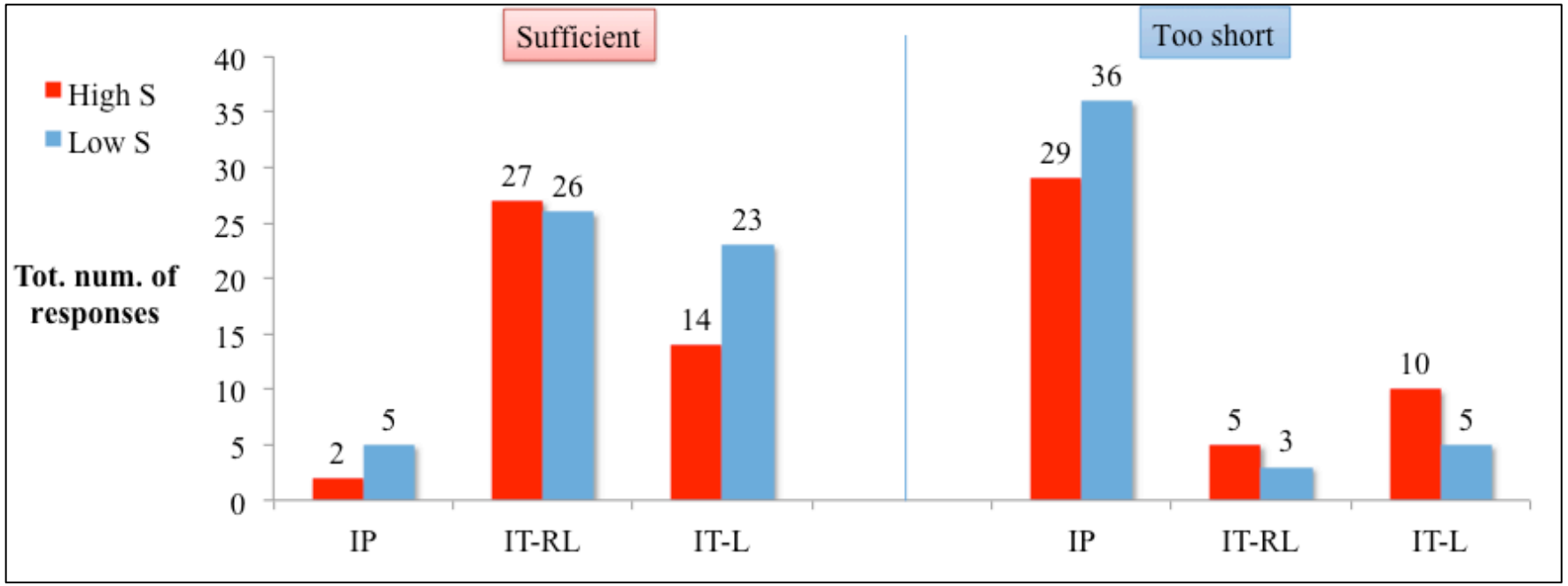


Figure 26 *Frequency of responses: Sufficiency and lack of planning time across test tasks*

Participants further gave reasons for the claims they had made. Figure 27 demonstrates the response patterns for why corresponding participants perceived integrated tasks were given sufficient planning time. Interestingly, the responses ultimately revealed what participants thought that each test-task type measured.

The majority of participants pointed out that the integrated tasks provided structured input sources that can be directly incorporated in the responses. For theses participants, restructuring or summarizing contents given in the reading passage and listening dialogue made it rather unnecessary to have a long buffering time in between. Participants 2007 and 2029 touched on this very matter. In so doing, they further made connections between the length of planning time and how the integrated tasks are not necessarily testing one's speaking ability.

> [Participant 2007, female, *low-scorer*]
> Because integrated tasks require the extraction of detailed information from reading and listening, you have nothing to prepare if you have not done a good job in reading and listening, or *vice versa*. Either way, you just sit there the whole time staring at the monitor. I would say integrated tasks evaluate your reading and listening ability, not your speaking ability.

> [Participant 2029, female, *high-scorer*]
> The goal is to repeat what was given in the prompts, not produce original language and content. If one remembers well or took quality notes, then 30 seconds are way too much of a time to spend for planning. In a way, integrated tasks test one's memorizing ability and skills. If you managed to jot down a lot of information, then you are good.

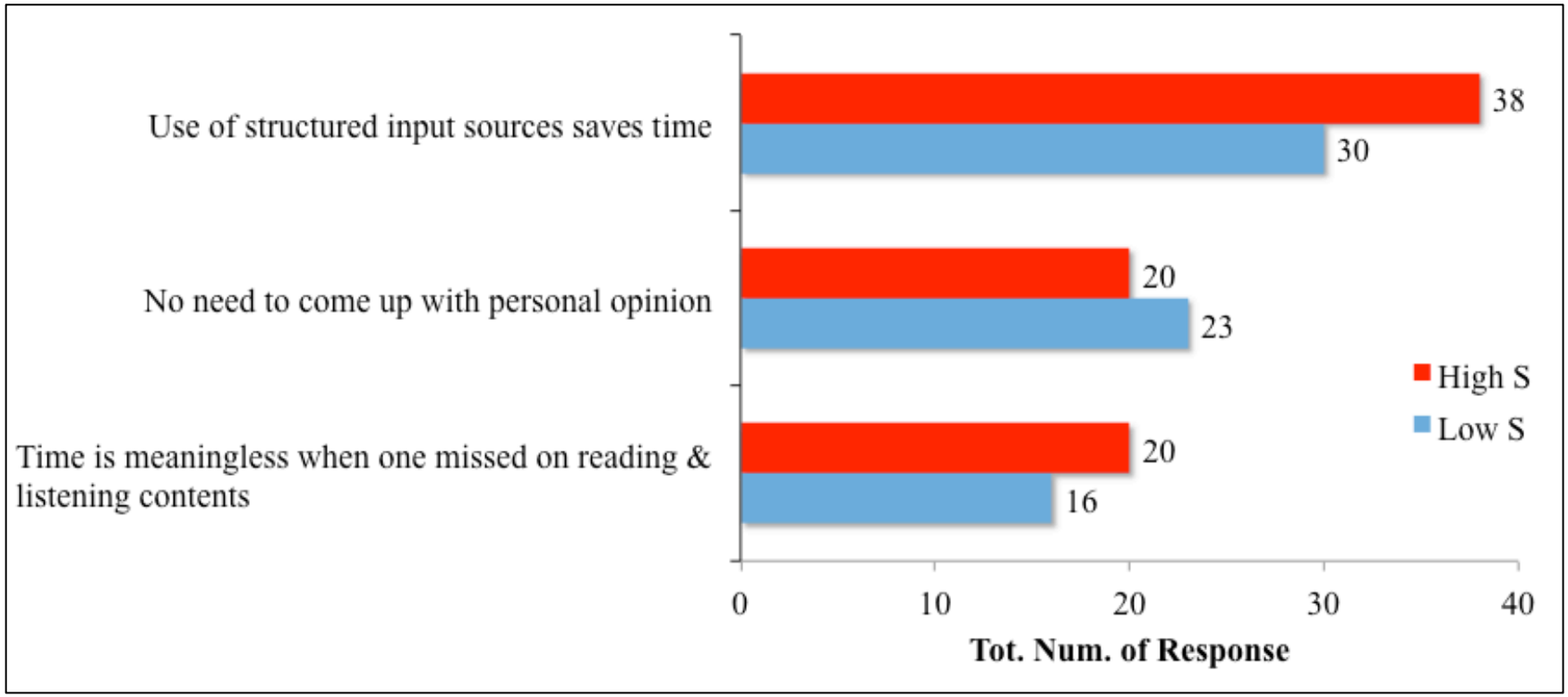


Figure 27 *Reasons given for the sufficiency of planning time in integrated tasks*

It is interesting to note how participants perceived that the appropriateness of planning time is linked with what they feel is being measured by the tasks. From Figure 27 and the comments above, it is apparent that those who were not successful in reading and listening the input sources were also not able to use the planning time usefully; they did not have sufficient information to structure their responses during the given time. In addition to language abilities such as reading or listening, it is striking that participants perceived that memory skills are at the core of the performance of integrated tasks.

Figure 28 summarizes the reasons for why participants felt IP tasks lacked sufficient planning time. Participants' concerns with IP tasks primarily lied in the heavier burden on constructing original responses to conceptual statements relative to the integrated tasks in which such burden is much reduced with input sources for reference (e.g., reading passages and listening dialogues) and concrete requirements of task fulfillment.

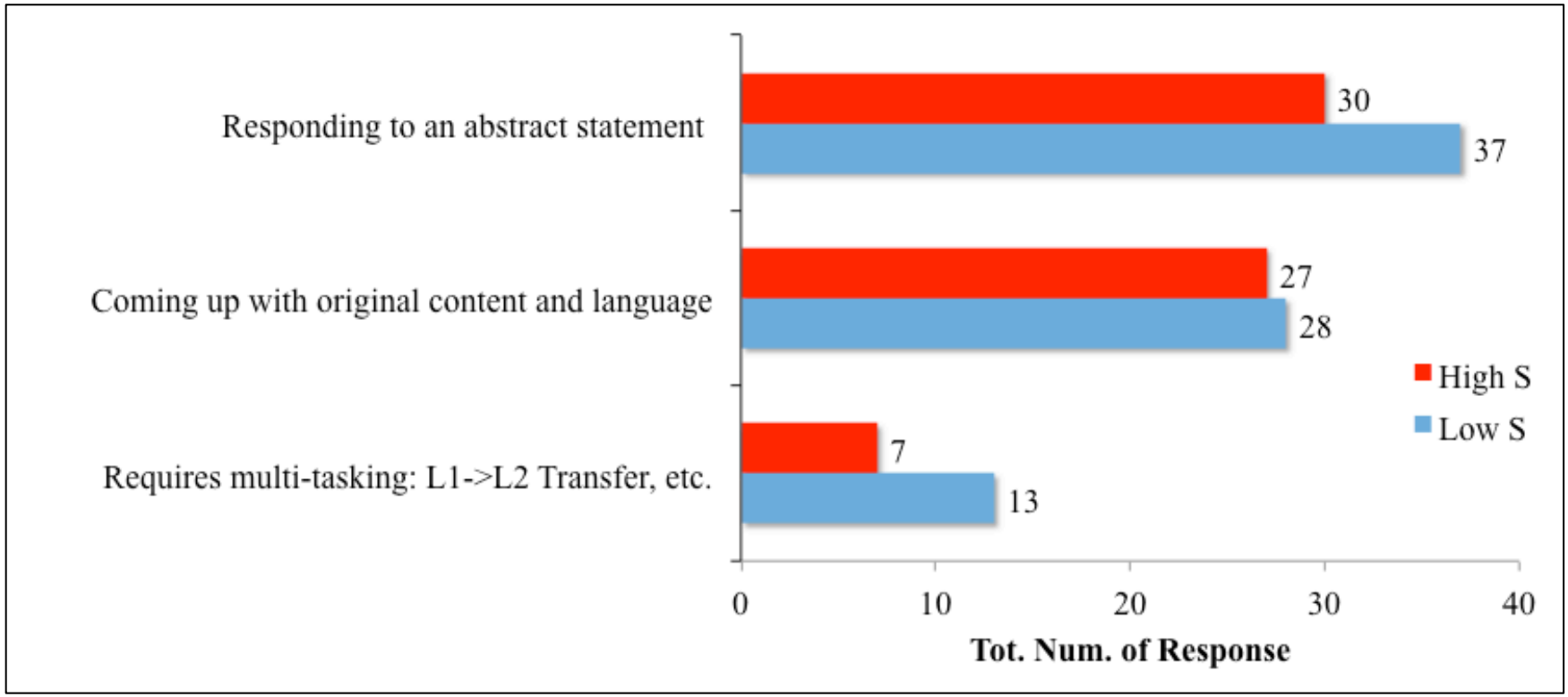


Figure 28 *Reasons given for the lack of planning time in IP tasks*

Participant 2025 well demonstrated on the trend depicted in Figure 28.

[Participant 2025, female, *high-scorer*]

Giving 15 seconds to plan for an abstract statement that I have never thought of before is quite harsh. What do you prefer reading: newspaper, books, or magazines? Who asks you this in real life? It's hard to think about my preference and supporting examples for such a topic, even though I know it's a test and I can just fake my answers. I would also have hard time instantly giving an answer to it in Korean.

The need to construct an original response was a burden for participants who went through the language transfer process (from Korean to English) before responding in English. For them, 15 seconds of planning time was insufficient for taking multiple steps in constructing a response, which encompassed understanding the statement, coming up with ideas, and transferring the ideas from Korean to English. Participant 2099 commented about this matter.

[Participant 2099, male, *low-scorer*]

I feel like I can do so much better if I was a native speaker of English. These are the type of questions that you would be better off if you were the speaker of that language, not even just being good at speaking it. You can't take anything away from the test prompt, it's just your ability to create an original response. And there's always that buffering time to transfer your thoughts in Korean first then to English. 15 seconds are way too short for that.

Due to the reasons above, some participants further commented on the possibility of long buffering or carried-over planning in the actual response. Because participants lacked with time for effective planning in IP tasks, they were not able to give immediate responses after being cued. Participants 2044 and 2081 gave such striking comments.

> [Participant 2044, female, *low-scorer*]
> For integrated tasks, your job is to structure your response according to the given information. Maybe adding a bit more of your own language, but nothing new. With IP tasks, you need to come up with a new story, which could be fake, but still, you need to start from scratch. Fifteen seconds are definitely too short to accomplish that. That's why I was not able to begin speaking immediately after the beep. I feel like I was quite numb, and that I kept thinking about what to say even after being prompted.

> [Participant 2081, female, *low-scorer*]
> I need to first think what I'll say in Korean, and then transfer it to English, but time only allows me to get to the Korean part. I need sometime before I speak. Even with that, I have to make my response as I go. At times I thought: what is the point of having planning time?

The spilled-over planning time essentially coincided with the *time spent before articulation* measured for *speed fluency.* This suggested that short planning time in IP task led some participants to take some time even during the actual response time, which further may have impacted on his or her *speed fluency.*

Table 36 summarizes the test-task characteristics in conjunction to what participants had commented on the associated planning time. From participants' perspectives, IP tasks were equipped with [+time pressure], [-input resources], and [+prompt abstractness], which led to participants feel less prepared for responding. Participants generally required more time to "jump start"; consequently, they may have used some of their actual responding time for planning. On the other hand, integrated tasks exerted features such as [-time pressure], [+input resources], and [+prompt concreteness] that may have nurtured a sense of preparedness among participants. Yet participants also claimed on how non-oral skills were essential for successfully integrated tasks.

Table 36

Properties of test-task characteristics in relations to planning time

Short planning time: IP tasks	Extensive planning time: Integrated tasks
+ Time pressure - Input resources + Prompt abstractness	- Time pressure + Input resources + Prompt concreteness
Less preparedness: Need more time to ***jump start*** Spilled over planning during response time	Enhanced sense of preparedness All or nothing: less efficient use of planning if reading and listening was not successful

CHAPTER 4: DISCUSSION

In this chapter, I discuss the major research results of the current study in light of previous literature on pre-task planning and task complexity; I specifically draw onto related perspectives from both conventional SLA and language assessment research. I organize the current chapter into four sub-sections in accordance to the four research questions guiding the present study.

4.1 Research question 1

For research question 1, I explored whether participants' test scores on the speaking tests are affected by planning conditions as well as test-task types. The major result from the MFRM analyses was that planning conditions had null effects on test performance across all three test sets. More specifically, test scores did not differ depending on whether participants were provided with detailed guidance of planning or not (i.e., Unguided vs. Guided planning conditions). Likewise, test scores did not vary within the two guided planning conditions either (i.e., GW vs. GT planning conditions). Such insignificant effects of planning conditions in all three test sets are in partial alignment with the results found in most of the language assessment literature on planning time (Elder & Iwashita, 2005; Elder et al., 2002; Wigglesworth, 1997; Wigglesworth & Elder, 2010). I note that it is partial alignment due to the fact that these studies compared the effect of planning in terms of the availability of planning ([- planning] versus [+ planning] conditions). In the current study, it was the *type of planning* that was manipulated across testing conditions; The planning itself was always made available. Therefore, it is uncertain whether the study results would have differed had the study design followed that of the previous literature.

Nevertheless, a robust consensus among the previous line of studies is that regardless of different study contexts, [+ planning] condition *per se* does not make a substantial impact on test score variations within testing environments (Ellis, 2009); In fact, this lack of a substantial impact was consistently observed across divergent planning conditions that varied on the length of planning time from 1 minutes of planning (Wigglesworth, 1997, 2001) to 3 minutes of planning (Elder & Iwashita, 2005; Wigglesworth & Elder, 2010). If this finding was similarly applied to the current study settings, it could be that the insignificance of varied planning conditions simply stems from the limited effects of [+ planning] in testing contexts in general. It should be noted that the same conclusion cannot be drawn from classroom- or laboratory-based SLA studies in which *performance* is primarily captured through discourse measures of CAF, and not quantified by subjective ratings. Yet as informed from the current study results (which will be discussed further in the following section), and from the bulk of relevant literature in language assessment, planning conditions demonstrate limited impact on overall speech quality measured by CAF dimensions as well (*cf.* see Nitta & Nakatsuhara, 2014, on how paired dynamics change between paired candidates in accordance to planning conditions in the contexts of paired oral assessment). This might be because there is a potentiality of unique factors underlying testing contexts that render the limited effects of [+ planning] condition on overall oral performance.

Some scholars responded to the lack of a [+ planning] effect in the testing environment as stemming from the impressionistic indicators of oral proficiency (Elder & Iwashita, 2005; Nitta & Nakatsuhara, 2014; Skehan, 2015; Wigglesworth & Elder, 2010); that is, they blamed the subjective nature of the score assignment itself. Wigglesworth and Elder (2010), for instance, noted that with memory constraints, test takers might be able to sustain the benefits of planning

only within the first few utterances of their speech (see Ortega, 2005, on how even in non-testing contexts, and hence with lengthier planning time, planned contents are limited in being completely transferred to real-time performance). Raters, on the other hand, may tend to formulate a final impression based on the overall speech or towards the end of the speech. Raters' score assignments, therefore, are averaged from the entire stretch of speech, and presumably not informed from a couple of planned utterance, if any (Wigglesworth & Elder, 2010). Thus, the assumption is that even if differential planning time differentially affects speech, it may not affect it for a sustained amount of time, and thus the effects may not be perceived in test scores. This is because the scores are rendered after the raters listen to sustained speech, long after the effect has, in essence, melted away.

On a related note, raters may resort to a general 'first impression' during the real-time rating processes. Although difference in modes does not warrant complete generalization, research from writing assessment informs that raters are less likely to make concurrent judgments during their live rating processes (Crisp, 2012; Grainger, Purnell, & Kipf, 2008; Lumely, 2002; Wolfe, 1997). Rather, they may make a rapid, intuitive evaluation based on their mental representation of the performance, and not necessarily refer back to a specific part of language output or scoring criteria for deriving a final judgment (Wolfe, 1997). Researchers from educational measurement have further analogously put forth that with cognitive demands and memory constraints in real-time rating procedures, raters, especially those with less experience, may rely on a general impression of one's performance (Lance, LaPointe, & Stewart, 1994). At times, these rating behaviors extend to a phenomenon widely coined as the *halo effect* (Fisicaro & Lance, 1990), with raters assigning similar score bands for the same individual candidate across test items that are administered under different conditions. The relatively less

experienced raters in the current study, therefore, may have been less sensitive in capturing qualitative differences in test takers' speech, if any (in the current study contexts, it would be planned versus unplanned utterances). However, the picture is still complicated in that as will be discussed shortly, results suggest that raters were indeed able to discern differences from IP tasks to integrated tasks in terms of score assignment. Presumably, it may be that saliency in what affects the discrimination of planned versus unplanned speech in the current study was insufficient to generate significant score discriminations, relative to how test-task types influenced the construction of speech.

From the test taker's perspective, this speculation is also plausible when considering how stringent the amount of planning time given in testing contexts is (*cf.* SLA researchers have put forth that planning is effective with at least 10 minutes of time; Mehnert, 1998; Ellis, 2009). With such timed natures, test takers might only be allowed to plan for one or two major points, addressing the task at hand (which they manage to actually put into words in the earlier part of their responses), and consequently they may make up the rest of the content during their actual response time. Recall again that such a pattern of incomplete planning was also observed with the current study's participants as evidenced through their interview responses (see Chapter 3). If this was truly the case in the current study, planned speech might have not constituted the entire speech sample of an individual participant to a great extent, thereby not weighing substantially in raters' judgments to render significant variations in scores. In relations to the timed nature of testing contexts, researchers have also noted a possibility of test takers being less skillful in using such short amount of time for planning (Elder & Iwashita, 2005; Wigglesworth & Elder, 2010). Test takers may be fairly familiar with planning or preparing for academic speech performance (e.g., presentations in non-testing contexts), but not within pressured, timed testing conditions.

Similarly, although instructions for planning were given to test takers in the current study (e.g., GW and GT conditions), their inexperience of taking TOEFL iBT-type tests (or generally any timed oral proficiency tests) might have confounded the effects of planning. In this sense, the notion of *authenticity* of employing planning time in testing contexts may be called into question (Elder & Iwashita, 2005). The premise of this line of thought is that the option of planning in testing contexts corresponds to how learners prepare themselves for performing academic speech in classrooms. However, although *authenticity* in testing does not always entail close resemblance of language practices in non-testing environment (Bachman, 2002; Splosky, 1985), it seems that the type of planning performed between testing and non-testing contexts exerts significant discrepancies.

A related speculation is whether the planning time itself was ever sufficient enough to generate meaningful contributions to improved speech or if the planning time was an effective use of time. As will be discussed shortly, this speculation is partially coming from the results of the current study, which showed a strong influence of test-task characteristics on whether participants benefitted from shorter or lengthier planning times. Even with the IP tasks in which participants generally claimed that they were disadvantaged by the short amount of planning time, participants made connections to such limits with the unique nature of the IP tasks. Therefore, as previous researchers uniformly concluded (Elder & Iwashita, 2005; Ortega, 2005; Wigglesworth, 1997; Wigglesworth & Elder, 2005), the length of planning time is not the sole factor explaining the underlying limited effects of [+ planning] conditions in testing contexts in general.

Going back to the discussion of rater's score assignment, previous researchers have interestingly suggested that a discrepancy may exist between the focus of what test takers take

into consideration when planning and what is being prioritized by raters when they judge (Ellis, 2005; Nitta & Nakatsuhara, 2014; Skehan & Foster, 1997, 1998; Wigglesworth & Elder, 2010). The demanding nature of testing contexts likely direct test takers to accurate language output, or perhaps helps them focus on forms during their planning. Indeed, Ortega (2005) noted that in non-testing situations, learners tend to not fixate on accuracy. On the other hand, if raters were able to adequately discriminate given speech samples according to the sub-dimensions of the rating scale, they would have globally looked into a number of qualitative measures (e.g., fluency, pronunciation, topic development, etc.) in addition to accuracy of language. In terms of the current study, if test-takers' awareness of being situated in a testing situation took over the effects of planning conditions, they may have consciously (or unconsciously) exerted selective behaviors in planning.

The above-mentioned possibilities are yet to be proven without further investigations in rater cognition as well as what test takers had planned for during the actual planning processes (e.g., analysis on the written planning of participants). In fact, what appeared to be more transparent in eliciting score variations from the MFRM analyses (and the GEE analysis) was the influence of test-task types. Regardless of the planning conditions, participants performed relatively poorly in the impromptu, IP tasks than the input-based, integrated tasks. This result could be somewhat counter-intuitive to the widely attested complexity and difficulty of integrated tasks, which are said to be difficult because they have dual-demands in skill application (more elements are required to be accomplished) as opposed to speaking-only test-tasks (e.g., Brown et al., 1984; Iwashita, McNamara, & Elder, 2001; Skehan,1998). Lengthened planning and responding times given in operational integrated tasks may have accordingly resulted from such attested task characteristics (see Plakans, 2010, for similar claims made in the

contexts of L2 writing assessment in which reading-writing integrated tasks are compared to writing-only independent tasks). However, speaking-only IP tasks are equally capable in eliciting not only perceived difficulty of the tasks, but also an enhanced sense of test anxiety during task performance among test takers (Huang & Hung, 2010; Hong et al., 2016). This finding in addition to the current study's results seem to point to the possibility that inherent test-task characteristics uniquely contribute to an over-ride of whichever planning activity was used (if any).

Especially within the context of IP task condition, there is a possibility of *within-task planning* (i.e., planning that takes place during real-time performance) taking over *pre-task* planning (Elder & Iwashita, 2005; Nitta & Nakatsuhara, 2014; Wigglesworth & Elder, 2010). Also termed as *online planning,* previous researchers argued that when pressured (by time limits or other cognitive capacities), speakers are likely to prioritize different levels of planning processes (Ellis & Yuan, 2003, 2005; Skehan & Foster, 2005). In such a context, speakers may not have enough time during pre-task planning to access necessary linguistic information in their working memory to facilitate language production. Likewise, if the stringent amount of planning time (15 seconds) and the IP-specific task features (e.g., abstractness in test prompts; see Chapter 3 for supporting interview responses) indeed had impacted on test-taking, participants in the current study may have not been granted the opportunity to fully exploit the given amount of time to plan for their responses. That is, the underlying task condition may even have forced participants to make use of constant short-term planning and monitoring of their speech during the actual response times. Recall again that in Chapter 3, the existence of *time spent before articulation* and within-task planning was revealed in the results of speech quality and test takers' interview responses on the effectiveness of the planning time in IP tasks. In testing

contexts specific to the IP-type tasks, therefore, test takers might benefit from ongoing planning, while pre-task planning are less likely to be used efficiently; in such cases, *pre-task* planning activities, whether they were supported with guidance or not, would have less significance.

But did participants similarly benefit from within-task planning for integrated tasks? One can speculate, but the survey and interview responses from participants demonstrated that the participants were less likely to feel the same extent of pressure from the length of planning time for integrated tasks that might lead to the extensive use of online planning. In fact, the majority thought that planning times in integrated tasks were sufficient, or at times, excessive. Note the relatively small difference in the length of planning times from IT-L and IT-RL tasks to IP tasks (5 and 10 seconds of difference from IP tasks). What might have caused these perceived differences (even with such small differences in length of planning time across tasks), and also the score variation between IP and integrated tasks? One plausible explanation could be that for integrated tasks, the use of *note-taking* during reading and listening might have been sufficed for constructing responses. Although Ellis (2005) pointed out that note-taking is not a core feature of what constitutes a "testing context," most high-stakes oral proficiency tests indeed provide participants with the options to take notes during testing (see Cublio & Winke, 2010); thus, it is not a feature only pertaining to "classroom, laboratory studies" of planning (Ellis, 2005, p. 218). In fact, the nature of information given in TOEFL iBT-type integrated tasks is highly demanding in terms of quantity and quality for restoring in and retrieving back from working memory; for instance, IT-RL presents both a paragraph-long reading text and a listening dialogue between two interlocutors. Upon such a condition, the tasks require a certain portion of information to be incorporated in responses, which hypothetically, would put more demand in the quality of immediate recording of the presented information. But because integrated tasks do not require

substantial *transformation of language* in responses (i.e., summarization of contents are required than originality of responses) (Prahbu, 1987; Skehan & Foster, 1997; Skehan, 1998), notes of key words or contents would suffice to construct a response (see Huang & Hung, 2010 and Hong et al., 2016 for similar suggestions). On the contrary, as some participants noted in the interview, if reading and listening was not successful (resulting in unsuccessful uptake of required information in note-taking), then they might not have had *enough* information to make use of during the given planning time. In such cases, pre-task planning would not have exerted a substantial impact on subsequent performance. Presumably, this precise feature of integration in a variety of levels (e.g., integration of both language skills and given information) seems to prompt participants to prioritize instant recording of information, then to retrieve information after few seconds later. If note-taking of the given input sources has operated as one variety of pre-task planning (and hence provided a sense of preparedness in responding), participants might have felt planning time was repetitive, something they had already engaged in.

Overall, the findings from research question 1 corresponds to Ellis (2009): Planning alone does not contribute to the underlying factors of task performance, but it should be understood in conjunctions to the fundamental characteristics of tasks. In the current study I found the lack of any significant findings of [+ planning] to be associated with how planning times were manipulated across the test-tasks, and how the test-tasks each required different approaches for preparing for responses.

4.2 Research question 2

For research question 2, I analyzed the quality of speech samples generated in each planning condition in terms of the CAF dimensions. In terms of the preliminary factor analysis,

the factor loadings of all dimensions were partially in line with previous researchers' findings (Skehan & Foster, 2005; Tavakoli & Skehan, 2005). For instance, the researchers also found that different measures across the three dimensions loaded onto the same factor. In the present study, *error-free clauses* loaded on to the same factor with *fluency* measures (e.g., *mean length of run*) and *complexity* measures. Yet the high factor loadings of the *accuracy* measures are an interesting finding.

In terms of the three performance dimensions, there was an alignment with the study results from Research question 1. The overall finding was that the effects of planning conditions were mostly over-ridden by test-task characteristics yet there were still interesting interaction effects between the two variables in a few measures. *Fluency* was generally influenced by the type of test-tasks that participants performed. Participants produced lengthier discourse in integrated tasks, while demonstrating less immediacy in actually articulating their responses in IP tasks. Such an effect of test-tasks was not revealed for *repair fluency*; however, there were hardly any instances of repairs in the raw coding (see Appendix H) to render any subsequent significance. Instead, planning conditions had significant main effects on the number of *repetitions* produced especially in the GT condition. Yet the small effect sizes and mean difference found amongst planning conditions within *repetitions* (GT vs UG: mean difference = .008, 95% CI [.003, .013], p = .000, d = 0.02; GT vs GW: mean difference = .005, 95% CI [.000, .010], p = .002, d = 0.02) are indicative of that the effect of planning conditions might be in effect not meaningful (Field, 2012). Thus, for *fluency*, I conclude that there were statistically significant main effects of test-task types on *speed* and *breakdown fluency*, which generally took over the effects of pre-task planning.

For *accuracy*, there was a significant interaction between planning conditions and test-task types on *lexical errors per 100 words*; the effects were revealed within the IT-L task in which lexical errors were made the most under GT condition. In terms of *grammatical errors*, on the other hand, there was only a main effect of test-task types; specifically, the integrated tasks contained fewer clauses embedding grammatical errors than the IP task. Therefore, the results are indicating that the GT condition had an effect on *accuracy*.

In terms of *complexity*, most *lexical diversity* measures demonstrated clear effects of test-task types; for instance, more *unique words* and *conjunctions* were produced in the integrated tasks than the IP tasks. On the other hand, IP tasks generated higher *TTR*, which is indicative of the measure's association with the length of the produced language across tasks. It was only from *subordination* for *syntactic complexity* that planning conditions interacted with test-task types; specifically, *subordination of clauses* was the greatest in the UG condition compared to GW and GT conditions within the IT-RL task. A main effect of planning was also only found from *sentence linking* devices for *lexical diversity*. In all cases, the two guided conditions contained more *sentence linking* devices than UG.

The overall finding departs from an extensive line of SLA research regarding the role of [+ planning] leading to greater *fluency* in spoken performance and mixed effects rendered on *accuracy* and *complexity* (Crookes, 1989; Foster & Skehan, 1996; Gilabert, 2007; Skehan & Foster, 2005; Tavokoli & Skehan, 2005; Yuan & Ellis, 2003). The current study is also unique in that such a null effect of planning on *fluency* was found in the contexts in which planning times were always made available. Moreover, it was the test-task types steadily influencing both *speed* and *breakdown fluency* in the absence of an effect of pre-task planning varieties. Several features of the study design that differs from the previous line of studies need to be discussed. First of all,

the previous studies were likely to focus on discerning the effects of planning itself while controlling for other variables such as task types (e.g., Elder & Iwashita, 2005; Kawauchi, 2005; Ortega, 1999; Mochizuki & Ortega, 2008; Rutherford, 2001; Skehan & Foster, 1997; Tajima, 2003; Wendel, 1997; Yuan & Ellis, 2003). In such cases, the same type of narrative or picture-description tasks were used across [+ planning] and [- planning] conditions. However, the positive effects of planning on *fluency* have been also discovered in few studies conducted by Foster and Skehan where task types had been manipulated in the study design (e.g., Foster, 1996; Foster & Skehan, 1996; Skehan & Foster, 2005). Then another possibility is that the type of *fluency* measures employed in the current study were fundamentally capturing a different aspect of *fluency* as conceptualized in previous literature (see Tavakoli & Skehan, 2005, for a discussion on the analytic measures adopted by SLA researchers for investigating the construct of *fluency*). Even if this is true, *fluency* itself has been inevitably operationalized through countless measures amongst the SLA literature; it is essentially a multi-faceted construct (Freed, 2000; Koponen & Riggenbach, 2000). Yet the fact that *fluency* has been continuously uncovered to make significant difference on oral performance with the aid of [+ planning] in SLA literature is suggestive of the potentiality that the current study can raise on either (1) differences in how planning operates across contexts (testing versus non-testing environment) (Ellis, 2009) and/or (2) the intrinsic characteristics of test-task types prevailing over pre-task planning varieties (and many other possibilities).

In terms of (1), previous researchers (Ellis, 2005, 2009; Elder et al., 2002; Nitta & Nakatsuhara, 2014) asserted that the high-stakes nature of testing contexts leads test takers to attend to the accuracy of the outcome of speech relative to its delivery. That is, based on their previous test-taking experience, test takers might have unconsciously (or consciously) developed

an internalized conception that answers to tests are essentially dichotomous in nature (e.g., "right" or "wrong") (Kohn, 2000). If such a perception governs an individual test taker during the test-taking process, he or she would hope to provide a "right" answer and attend less to how it is expressed even in constructed-response tests (recall that *accuracy* is also operationalized as "a concern to avoid error"; Skehan, 2009, p. 510). For instance, test takers might produce errors in suprasegmental features in speech (e.g., speech rate, intonation, word stress, accents that have them appear less engaging) yet strive to deliver error-free speech. Subsequently, such test-takers' sensitivity towards providing accurate responses would lead them to engage in careful online monitoring of their speech articulation even during actual response time (Ellis, 2005). In fact, the results of certain sub-measures of the CAF dimensions potentially speak to the task-specific effects of *within-task planning* raised in the previous section. Although different allotments in response times across test-tasks should be taken into consideration, the tendency in the IP tasks was that participants generally spoke less within a shorter stretch of articulation time (reflected by the increased *time spent before articulation*, *speech rate*, and decreased *mean length of run*) and demonstrated pausing or hesitating phenomena more frequently (suggested by the number of *unfilled pauses* and the *time spent before articulation*). On the other hand, they produced fewer *lexical errors* while generating more cohesive devices (such as *sentence linking*) for the IP tasks than for the IT-RL tasks. Within the integrated task conditions, however, the CAF dimensions showed a potential case of parallel gains. The differentiated output for *fluency* and *accuracy* in accordance to test-task types suggests that participants may had difficulty in directing equivalent and simultaneous attentional resources for *fluency* and *accuracy* particularly in the IP tasks; the high-stakes nature of testing may have imposed more emphasis on language usage during live performance so that participants engaged more in pausing for monitoring (e.g., *unfilled pauses*)

or buying time to plan the language for what will be said (e.g., *time spent before articulation*) (Skehan & Foster, 2005) at the expense of fluidity in speech.

In particular, the significant increase in the *time spent before articulation* in the IP tasks relative to the two integrated tasks necessitates further exploration. It is an empirical question as to what such measure generally signifies: is it a reflection of test anxiety specific to the timed nature of speaking tests, with an added pressure stemming from the unfamiliarity of talking onto a computer machine (see Lee & Winke, 2018, for discussions on how computerized test features such as a timing device can affect test taker's cognitive operations for performing on oral test-tasks)?[2] Drawing upon its factor loading with *speed fluency* measures (see Table 25), the increased *time spent before articulation* specific to IP tasks seems to at least indicate a decreased sense of immediacy or promptness (i.e., hesitance) in providing instant responses. When further probing into participants' interview responses, the measure also implies the carried-over, incomplete pre-task planning. In fact, seminary researchers in the field of cognitive psychology in spontaneous speech have long put forth a relevant proposition describing the very phenomenon (Butterworth, 1975; Goldman-Esler, 1968; Griffin & Bock, 2000; Tannenbaum et al., 1965). Goldman-Eisler (1968) claimed that spontaneous speech is both cyclic and incremental in nature; that is, although with divergent variations in duration and length, speech is consisted with a *hesitant phase* (chiefly consisting of silence; Butterworth, 1975) proceeding a *fluent phase* of speech (i.e., ongoing utterance). For instance, even in real-time speech settings, speakers tend to take at least a brief second or two to respond back to their interlocutors' questions (see Griffin & Bock, 2000, in how speakers, even gazing at relevant objects, take 500

[2] This is less likely as participants did not show the same pattern of behaviors for the integrated tasks. Novelty effect on tasks does not seem to be the answer either since all task conditions (orders of test-tasks, planning conditions) were counter-balanced across study groups.

milliseconds of time before properly naming the objects in their L1). Within such contexts, the researchers posited that the preceding *hesitant phase* function as the cognitive processing time, or the planning time necessary for the subsequent *fluent phase* to take place (Butterworth, 1975). Therefore, the silence during the *hesitant phase* is psychological in nature in that by delaying the onset of speech, speakers are either consciously or unconsciously striving to retrieve sufficient information from their working memory and prepare for the upcoming articulation (Butterworth, 1975; Goldman-Eisler, 1968). An extensive amount of time taken during the *hesitant phase* then illustrates that speakers are in need of longer *buffering time* (Levelt, 1989). In the current study, the buffered time used in addition to the given planning time would suggest that participants had to engage in extra planning due to the inefficient and ineffective use of original planning time given for the IP task (this will be discussed further in the following section). While *time spent for articulation* reflects the cyclical processes in naturalized speaking contexts, it might function differently in time-pressured settings such as testing contexts. In fact, the amount of such buffered time was negatively associated with test performance, indicating that the longer spent for "jump starting," the lower the test scores had gotten. This suggests that at least for the lower scorers, planning may not have been complete or was inefficiently carried out during the original pre-task planning time, which subsequently affected these individuals to carve out a certain portion of time out of the actual responding time for an additional preparation time. This may have caused them to engage more frequently in inner speech monitoring during speech and have them eventually run out of time for providing complete responses. While further investigation is needed, results from factor analysis that *time spent for articulation* is negatively related to some *fluency* and *syntactic complexity* measures is also potentially suggestive of its unfavorable effect on subsequent speech production (yet the directionality of the effects on different measures

should be further warranted). Overall, regardless of whichever pre-task planning conditions participants were under, the effects could be limited due to the inevitable split-over planning as well as *within-task planning* taking place.

Going back to the general trend appearing in the data, the findings seem to lend moderate support regarding the Limited Capacity Hypothesis (otherwise coined as the Trade-off Hypothesis; Bygate, 1999; Skehan, 1998; Skehan & Foster, 1999) because there was a trade-off between *fluency* and *accuracy* in oral performance, and generally there was a competition among the three performance dimensions. However, the fact that such phenomenon was only observed within the IP tasks and not uniformly across test-tasks or planning conditions calls for speculation. Alternatively, the results could be understood from the finer categorization of task characteristics and conditions (Skehan, 2001). As previous researchers have advocated, the following task features are likely to advantage both *accuracy* and *fluency*: (1) tasks presenting concrete or familiar information; and (2) tasks containing clear structure (Skehan, 2009). In addition, tasks requiring information manipulation (e.g., summarizing or describing pictures) lead to higher *complexity*. From what participants had elaborated, IP tasks are quite the opposite from these descriptions owing to the abstractness in the statements and the absence of cued information; rather, it seems that such features pertain to the integrated tasks. If this was the case, then the integrated tasks may be equipped with characteristics that basically free up limited working memory constraints during test taking and eventually advantage all or the majority of the CAF dimensions simultaneously (which is connected to the higher test scores participants received on these tasks). These may include [+ concreteness] of test prompts, [+ input resources] by presenting structured information for reference, and [- time pressure] with increased planning and responding time (see Table 36 in Chapter 3). Additionally, the aid of note-taking operating

as pre-task planning for integrated tasks could have boosted the effects. On the other hand, IP tasks may inherently challenge test takers to concurrently improve on all three CAF constructs. The elements constituting IP tasks may have contributed to higher extents of constraints in one's working memory (Huang & Hung, 2010), which led to commitment of attention to one area of speech quality (Skehan, 2009). Subsequently, this result may indicate that the IP tasks and the integrated tasks are not only tapping onto different processing constraints for test taking, but also mapping onto different constructs of spoken performance (this will be discussed in the following section in relations to participants' perceived differences in the effectiveness of planning conditions inherent in IP versus integrated tasks) (*cf.* see Brown et al., 2005, on how the two types of tasks did not exert significant qualitative differences; but see Kyle et al., 2016, on their take on the two task types by means of a variety of linguistic characteristics).

Meanwhile, the results from the integrated tasks may partially corroborate Robinson's Cognition Hypothesis (Robinson, 2001, 2002) concerning the simultaneous improvement of the CAF dimensions. The premise of Robinson's claim is that *accuracy* and *complexity*, in particular, increase within the contexts of increased task complexity; that is, the complex nature of the tasks advances speakers to generate enhanced quality of speech (Swain, 1993). Based on Robinson's taxonomy of task complexity (2001, p. 294), integrated tasks do indeed possess several elements contributing to the added layer of complexity of the task conditions such as the extra demands in language ability (e.g., reading, listening, then speaking) as well as integrating relevant information into the performance. However, it turns out that in the present study, such inherent elements of the integrated tasks were *perceived* by participants to function positively on language production. That is, the input resources (e.g., reading passage, listening dialogue) had served as references for participants. Although it is unknown to what exact extent did the

participants made use of the input sources directly into their responses, the potentiality of such text integration taking place could explain the concurrent enhancement of *accuracy* and *complexity* in the integrated task performance. Recently, Crossley et al. (2014) found evidence as to a positive influence of text integration practices on overall speech quality in the context of IT-L task condition. They discovered that the amount of the words in the source texts being integrated into the response was the strongest predictor of test takers' overall speaking test scores on the IT-L task. Implicit in Crossley et al. (2014) (and also in the extensive line of research on text integration in integrated writing tasks; see, for instance, Plakans & Gebril, 2016) is that through integration and repetition of source-text features, test takers are able to reproduce key contents of the sources in their responses, which in turn lead to the construction of an overall coherent, rich response. While Crossley et al. (2014) made their observations upon usages of individual content words, it might also be the case that test takers go beyond word-level integration to exploit further grammatical and lexical properties of the language used in the source texts (e.g., verb phrases, collocations). To some extent, the responses to integrated tasks then might be summaries or paraphrases of the presented source texts at best (*cf.* IT-L tasks require a summary as well as the test taker's personal evaluation of the given problem in the task). If the *originality of language* is not a major concern of evaluation (as implied in the TOEFL iBT Speaking rubric for integrated tasks), then such text-integration practices might actually be one of the key factors what make the overall performance to appear as upgraded in dimensions such as *accuracy* and/or *complexity* (*cf.* see Crossley et al., 2014, for their concerns on the construct validity of integrated tasks owing to the active text-integration practices of test takers). IP tasks, on the other hand, may appear to adhere to the kind of simpler monologue tasks (as what previous researchers had identified; Foster & Skehan, 1996; Robinson, 2001), but in

effect, the elements they inherit may impose higher cognitive demand on test takers; hence, resulting in lower test performance and imbalanced development of the speech quality dimensions. While empirical investigation is further needed to confirm these speculations, participants' way of uniformly discerning task characteristics specific to IP and integrated tasks suggest that the widely-held task complexity dichotomy in the field (e.g., Robinson, 2001) may not be well applied to the current study's context, or generally to the testing contexts as the TOEFL iBT. Although task participants' perceptions (also defined as *task difficulty*; Robinson, 2001) cannot be relied upon entirely, the current study found plausible connections as to the results pertaining to the test performance, speech quality, and to the *perceived differences* of test-taking conditions across test-task types.

The final point that should be addressed is the small, yet significant effects of planning conditions found on three specific measures; namely, *lexical errors per 100 words*, *subordination of clauses*, and *sentence linking devices.* Especially, the two guided conditions lend mixed results; for instance, while GT condition seemed to render *lexical errors* the most (within the IT-L task condition), both GW and GT conditions generated more cohesive features in speech than the UG condition. The UG condition, on the other hand, facilitated subordination of clauses more so than the two guided conditions. In terms of the results pertaining to *lexical errors* and cohesive features in the speech, Foster and Skehan (1996) found similar results in their study comparing the effects of detailed and undetailed planning conditions (which basically corresponds to the guided and unguided planning conditions in the present study). In their study, the number of *error-free clauses* (which encompassed both grammatical and lexical errors) was least likely to occur in the undetailed planning condition. On the other hand, *complexity* measured through clause-to-C-unit ratio was greater in the detailed planning condition. The

researchers hence suggested that when explicit suggestions or guidance directing to how speakers should use the given planning time or relevant pre-task planning strategies, task participants may direct their attention to the content of the message (e.g., coherence of the speech) to be expressed rather than language. When also drawing onto the interview responses, the GT condition in the current study may have caused selective channeling of resources in responses, facilitating global idea generation at the expense of the retrieval of appropriate language forms. Conversely, the condition that simply gives planning time and no directions as to tasks might have allowed participants to speak more freely. Especially when considering that the effects were found on the integrated tasks, the inherent task characteristics such as [- planning time pressure] and [+ input resources] may have equipped participants with enhanced sense of preparedness (or confidence in speaking), which in turn facilitated them to make more attempts in elaborating; hence, leading them to produce more lengthened subordinate clauses. It could also be the case that the text integration practices (as mentioned above) enabled in IT-RL task conditions may have given sufficient source of information for participants to "play around" with the language and eventually push them to try out more.

4.3 Research question 3

For research question 3, I analyzed the survey response data in terms of participants' *confidence* in their performance per test-task type, participants' perceptions of *appropriateness of planning time* per test-task type, and participants' evaluation on the *effectiveness of type of planning*. In accordance to the previous studies conducted within testing contexts, neither the participants' self-assessment of task difficulty (indirectly gauged as *confidence* in task performance in the current study) or the effectiveness of [+ planning] was found to significantly

differ per planning conditions (e.g., Elder & Iwashita, 2005; Elder et al., 2002; Wigglesworth & Elder, 2010). Researchers considered such results generally stemming from consistent null effect of [+ planning] found in testing conditions. From various reasons given in the previous sections (over-riding of *within-planning*, note-taking, and test-task characteristics), the current study's finding concurs with what the previous researchers have asserted.

However, the results from the RM MANOVA with the survey responses revealed that participants' perceptions differed statistically significant on the *appropriateness of planning time* per test-task types across planning conditions (Greenhouse-Geisser correction: $F_{2, 538} = 2.84$, $p = .027$, $\eta p^2 = .016$). From the follow-up post-hoc tests, IP tasks' planning time was considered by participants to be insufficiently given than the other two integrated tasks, especially only under the two guided planning conditions. In other words, the planning time inherent in the IP tasks was perceived to be shorter when participants were given guidance in planning (either by writing for planning or thinking for planning). This might indicate that neither GW or GT operated positively in reducing the time pressure participants may have been imposed on particularly under the IP task condition.

Two possibilities can be raised from this result: (1) The provision of guidance did not help participants to use the planning time in the IP tasks efficiently; and (2) participants are likely to be less pressured even by the short amount of planning time in the IP tasks when not given any guidance in planning. In terms of (1), the physical act of writing (as pointed out subsequently by participants in the interview), although with its advantages, takes up quite some time of the planning time; in such cases, it might not function efficiently within the most time-pressured planning condition such as the IP tasks. The GT condition, similarly, may have restrained participants from freely resorting to their own strategies for planning. This connects to

point (2), in that the instructions given for planning have operated as an additional layer of pressure for participants; presumably they might have felt like they are obliged to follow the instructions, and not resort to their own strategies for planning. Especially within the unique contexts of testing (enhance time pressure, high-stakes nature of tests, etc.), test takers are inclined to develop and use whichever strategy works best for them (Wigglesworth & Elder, 2010). In real-life settings, if test-takers have prepared extensively for such high-stakes tests, they are likely to have internalized a certain set of strategies that are biased for their best (Swain, 1984). Thus, unlike the non-testing contexts where students are eager to learn and receive support from their teachers, careful guidance in planning may not be effective, or even be distractive to test-takers who may have already developed their own way of responding to the test-tasks.

4.4 Research question 4

The quantified performance data observed in answering research questions 1 and 2 indirectly touched on the possibility of the influence of test-task characteristics on how participants managed the given planning times per test-task type. However, the interview data directly addressed the possible discrepancy between test developers and test takers in terms of the immediate impact of test-task characteristics on real-time test-taking and performance (Ockey, Koyama, & Setoguchi, 2013).

There were essentially three recurring themes consistently maintained throughout the interview data. First, in line with the survey responses, participants felt that they had benefitted the most under GW planning condition. This was due to their familiarity with the act of writing in both testing and non-testing situations for preparing for a certain task. With the GT planning

condition, on the other hand, some participants agreed that the GT planning condition had hampered their effective planning. It is quite interesting to note that in reality, participants' test performance did not differ significantly between either of the planning condition. It could be that the familiarity factor inherent in the GW planning condition may have masked some of the advantageous features pertaining to the GT planning condition. This is plausible because some participants were indeed able to point out the efficiency of GT in certain instances. For instance, participants noted that GW is directly helpful in retrieving linguistic information while GT is useful in idea generation. In addition, they also pointed out that GT is helpful for test-tasks that are given shorter amount of planning time, which is indicative of a possible interaction effect between planning conditions and test-task types. Perhaps such an effect may be realized with participants with more diverse profiles (e.g., low or intermediate level of English proficiency). Kawauchi (2005) is certainly an example of this case: learners with lower levels of English proficiency in this study preferred and benefitted from the option of reviewing reading resources over making notes for planning. Such diverging patterns between learner groups were not clearly observable in the current study owing to the background of participants, which is a limitation in the current study design that should be addressed by future studies.

Second, participants put forth a uniform statement that the planning times should be reversely offered between the IP and integrated test-task types. This result further suggests that the IP test-tasks are not in any way 'simpler' than integrated test-tasks; perhaps, it could even be that test takers perceive the IP test-tasks to be more difficult owing to how the tasks are designed. Presumably, the perceptions of different groups of stakeholders hold in terms of task complexity and difficulty vary considerably. Although it is not clearly articulated in the TOEFL iBT testing manual, the task-based language testing scholarship (or largely the language testing field) has

shaped an assumption that integrated test-tasks requiring an added demand in language skills are inherently complex (e.g., Kyle et al., 2015; Plakan, 2010; Wigglesworth & Foster, 2016). Yet according to the participants, the precise nature of dual demands in language skills was what assisted their test performance. On the other hand, the IP test-tasks were equipped with task characteristics signifying higher degree of complexity (Skehan, 1998); [+ time pressure] from the short amount of planning time, [+ abstractness] from the test prompt, and [- input resources] that they can refer to. On the basis of these categorization, participants' responses further reflected on how they thought about the construct of the test-tasks are. Interestingly, they held diverging views on what the IP and integrated test-tasks are each measuring.

Third, participants' responses validated the possibility of the influence of a spilled-over planning on prompts performance of IP test-tasks. In addition to the disadvantage of the short amount of planning time, participants provided that how they process and decode the test prompts are the reason why their *pre-task* planning is carried over the actual responding time. If this is indeed a valid phenomenon, the short amount of planning time, in some cases, might be destructively operating in subsequent responding processes. In fact, the trend found in research question 2 regarding *time spent before articulation* was that lower-scorers involved in more silence time before they began to articulate their responses. To obtain a more comprehensive account of what really *time spent before articulation* entails, introspective investigations (e.g., retrospective think-aloud protocols or exploring the written planning sheets of participants) could be carried out in future research. Yet the negative relationship between test performance and increased amount of carried-over time could at least cast a concern on the effectiveness of the short planning time given in IP test-tasks.

This study result, therefore, indicate two further points. First, observing how test takers used the given planning times can further tap onto the inherent test-task characteristics and their influence on test performance. Second, test takers are capable of providing insights on the properties of test-task conditions and characteristics, and hence, reveal their perceptions of the validity of test constructs. This suggests the need to regularly monitor the widely-practiced task implementation methods and conditions as well as to revisit test developer's assumptions on how test takers *would* perform on a given task.

CHAPTER 5: CONCLUSION

5.1 Implication

Task-based performance testing has its significance in promoting authenticity and practicality in assessment with the aim to replicate the kinds of activities as well as the language ability which candidates are likely to (and are needed to) demonstrate in the real-world contexts (Long, 2015; Wigglesworth & Elder, 2010). However, often such merits disguise the hard reality that "as a test method…it remains one of the most expensive approaches to assessment and, in terms of development and delivery, one of the most complex" (Wigglesworth & Foster, 2017, p. 129). The fact that it is the test-task that is the centralized unit underlying a number of relevant practices such as test development, scoring, and language pedagogy, lends implications in different areas of task-based language assessment and pedagogy scholarship (Norris, 2016). In concluding the dissertation, I first draw on the purpose and the results of the current study to address its precise contribution to the existing and growing body of task-based language testing research, and to further elaborate on the broader implications this research has for language pedagogy (Long, 2015; Van Gorp & Deygers, 2013).

First and foremost, the study adds to the literature of task-based language testing research by undertaking a joint investigation of both test-task implementation condition (e.g., planning conditions) and test-task characteristics (e.g., IP and integrated test-task types) in the context closely resembling the operational testing of TOEFL iBT. As a result, the study findings could be interpreted to address the theoretical and empirical underpinning of an existing assessment not to mention the ecological validity in research. Studies of planning time thus far have been carried out in closed laboratory settings (e.g., Wigglesworth, 1997) or in real testing settings where planning times are originally not provided (e.g., Elder & Iwashita, 2005; Elder et al., 2002;

Tavakoli & Skehan, 2005; Wigglesworth & Elder, 2010); in such contexts, it may be often difficult to discern whether the effects are rightfully pertaining to the speech-processing mechanisms specific to the unique environment of test-taking. In this dissertation, the null effects of planning were consistent with the previous planning literature in language testing (Wigglesworth & Elder, 2010). However, the reasons as to that precise discrepancy from conventional task-based research finding was discussed in light of the mediating influence of test-task types, which fundamentally stemmed from the unique contribution of the testing environment of the TOEFL iBT speaking section. Therefore, the study results not only addressed why there are divergent effects of planning in testing research, which in turn could also inform the existing task-based research (Ellis, 2009; Skehan, 2016; Tavakoli & Skehan, 2005), but also tapped onto the test-takers' processing constraints stemming from the imposed conditions and characteristics specific to the TOEFL iBT test-task types. Such an investigation has the potential to raise either concerns or validation of how the employed test-tasks relate to the construct of interest, as well as which factors may potentially obstruct valid and reliable testing practices (Wigglesworth & Foster, 2017).

In this sense, the study can also be seen as an effort to bridge the two lines of research entities: namely, the more theoretically-underpinned general SLA and language assessment. One aspect of the current study that adds to such link is the study's establishment on the theoretical framework of conventional task-based research in addition to employing methodologies used in the language testing research. In this dissertation, I attempted to triangulate multiple data sources by conducting analyses linking the quantified scoring data to the CAF measures (e.g., GEE analysis of the relationship between the test performance and the quality of speech as measured through CAF measures). This was to account for the arbitrariness of the three-way categorization

of performance constructs, which is a critical concern from a measurement standpoint. Two relevant points can be raised. First of all, task-based research has mostly, and solely conceptualized *performance* as to the extent to which higher and lower degrees of CAF measures are identified; that is, higher degree of the three dimensions (e.g., high *accuracy*, *complexity*, and *fluency* reflected in speech) generally imply greater spoken performance. However, *performance* as understood from the perspective of language assessment is that it is a wholesome concept that needs to be interpreted based on multiple yardsticks and evidence. Validation of an assessment product is essentially a holy grail of collecting evidence as to whether the implemented task condition or test-task features are leading up to demonstrating/tapping onto one's true ability. Second, the distinctions of the three constructs are not so much of a concern in language assessment. Often times scoring rubrics may or may not treat each dimension as separate entities; holistic rubrics, for instance, may have raters assign a broad level band of performance that basically compromises the varying degrees of performance dimensions (which might have been the case of the current study as discussed above in Chapter 4). Therefore, lower degrees of a particular dimension are not seen as detrimental to overall test performance. In addition, it is likely that raters are trained or inclined to assign a score out of their global impression of a test-taker's speech and paying less attention to specific aspects of performance.

But at the same time, language testing research has not fully exploited what the three performance dimensions (CAF) can offer in illuminating performance differences across planning and task characteristics (Wigglesworth & Frost, 2017). Presumably, the three CAF constructs are the most researched and supported 'watchdogs' of language development from both theoretical and empirical standpoint. They account for the *processing competence* (Skehan, 1998), or the mechanisms related to the *response processes* (AERA, APA, NCME, 2014;

Standards for Educational and Psychological Testing), which is an under-researched domain in language testing, but it is very much vital to comprehend the multi-faceted setting of testing. Thus, the CAF framework could give valuable insight as to both test development and research practices in terms of refining the rating scales and rubric on a narrower level yet ultimately the test construct on a broader level.

The study result also pinpoints the discrepancy between speaking test developers and speaking test takers (Ockey et al., 2013). Often times, test taker's voices operate as a supplemental source of reference of test validation (Cheng & DeLuca, 2011; Hamp-Lyons, 2000). However, the current study's findings provide validation evidence for the TOEFL iBT test-task types from the test takers owing to their evaluation of "test constructs and the interaction between these interpretations...and test design" (Fox & Cheng, 2007, p. 9). The test takers in this study shed new insights on a number of task conditions and characteristics that may or may not have been clearly articulated in terms of their theoretical and practical considerations. This demonstrated that test-takers' perceptions and orientations are equally valid sources that prompt test developers to revisit their hunches and assumptions on what have been granted and practiced for an extensive period of time (Moss, Girard, & Haniford, 2006).

On a pedagogical standpoint, task-based testing research (as in the current study) can help teachers in making informed decisions about their own assessments and instructional practices. As a starting point, the intertwined effects of task conditions and inherent characteristics can be interpreted in syllabus design, teaching material development, and generally teacher's professional development. But on a fundamental level, the findings pertaining to the discrepancy found in the planning effects between the conventional task-based and testing research can facilitate the thinking of connecting assessment and instructional practices in classrooms.

Teachers can first draw from research, such as the current study, on studying what is unique to the testing settings. They can further make revisions on how they administer and design test-tasks for assessment purposes in their classrooms. For instance, they could try out a variety of lengths of planning time or planning activities for different in-class and assessment tasks (e.g., presentations, pair/group discussion, one-on-on interview), and see in which condition students are able to demonstrate their spoken abilities better. Teachers would greatly benefit from exploring the evidence-based research findings that hint at a possible intertwined effect between task design and language performance on enhancing their assessment literacy, and the awareness of properly gauging student performance in classrooms.

5.2 Limitation and future research

There are a number of limitations of the current study that could be addressed by future research. First of all, the study design could be refined to better ascertain the effects of planning time and test-task types. In the current study, the test-task conditions were employed directly from the operational testing setting; but the findings could differ in a more tightly controlled environment where variables are manipulated in a variety of ways. For instance, it could be investigated whether test takers truly benefit more from a shorter planning time for the integrated test-tasks and a longer time for IP test-tasks. Likewise, the planning conditions devised in the present study might not have been divergent enough from one another to generate significant differences in test performance. Yet given that participants have distinctive orientation toward a particular planning condition (e.g., GW) especially when performing a particular test-task, future studies can look at whether such effects hold with different type of test-tasks (e.g., decision-making tasks).

Second, there is a possibility that participants performed better on integrated tasks owing to note-taking before the planning time had begun. In this sense, the effect of pre-task planning might have been masked to some extent because of a pre-planning activity. Although allowing note-taking was to simulate the actual operational testing, future studies might consider not providing the option of note-taking to tease out the effect of pre-task planning.

Third, the current study collected data from speakers with specific profiles; that is, they were relatively proficient speakers of English and they did not have extensive experience of taking the TOEFL iBT test (or other types of English oral proficiency tests). The rationale behind recruiting such participants were two folds: first, to ensure that a meaningful unit of speech is elicited for data analysis, and second, not to confound participants' reactions to the employed test-task condition from previous test-taking or test-preparation practices. However, the findings could differ with test takers with more diverse profiles; particularly, the limited effects of planning found in the current study could be reversed when involving speakers with lower level of English proficiency (e.g., Kawauchi, 2005).

Finally, the study results pertaining to the CAF indices were based on subjective, manual coding. Although the coding procedures involved multiple steps and intervention sessions to maintain rigorousness and reliability in the data, the labor-intensive nature of manual coding could have had a certain extent of influence on how coders interpreted the data. Future studies could make use of the automatic text-analysis tools (e.g., TAACO; Crossley et al., 2016; Kyle et al., 2015) to avoid such potential impact of subjectivity (and fatigue) of coders imposed on speech data. However, the use of the automatic tools should be cautiously carried out as they are primarily designed to analyze written texts, and hence, may not be sensible to catch certain speech phenomena (e.g., pausing). A certain extent of manual coding in such cases could be a

beneficiary supplement.

APPENDICES

Appendix A

Language learning and test-taking background questionnaire (in English)

Thank you for participating in this survey. This survey is distributed as a part of a larger study conducted by Shinhye Lee. Please send her an email at leeshin2@msu.edu if you have any questions. This will take 10 to 15 minutes in total.

General Information

1. Name: _____
2. Gender: ☐ Female ☐ Male
3. Date and year of birth:
4. Name of university that you are enrolled in now:
5. Year in college: ☐ Freshman ☐ Sophomore ☐ Junior ☐ Senior ☐ MA/Ph.D.
6. Major field of study: _____
7. What is the main language you speak at home? _____
8. What other languages do you speak at home? _____

Language Learning Background

9. At what age did you first started to study English? _____
10. How long have you been studying English? _____(years) _____(months)

11. In which contexts/situations did you study English? Check all that apply.
    - ☐ At home (from parents, caregivers)
    - ☐ At school (Primary, secondary, high school)
    - ☐ At private institutions
    - ☐ After immigrating to the English-speaking countries
    - ☐ At language courses during my study abroad in the English-speaking countries
    - ☐ Other (specify): _____

12. How often are you engaged in the following activities for English?

Activities	Daily	Weekly	A few times a month	Once a month or less
Listening to news broadcasts or music	☐	☐	☐	☐
Watching TV or movies	☐	☐	☐	☐
Reading books/magazines	☐	☐	☐	☐
Writing emails	☐	☐	☐	☐
Speaking with friends outside the class	☐	☐	☐	☐

13. Please rate on a scale of 1-6 your current ability on English reading, writing, and listening (circle the number below). (1= *Very poor*; 2= *Poor*; 3= *Fair*; 4= *Good*; 5= *Very good*; 6= *Native-like*)

Reading	Writing	Speaking	Listening
1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4 5 6

14. Please rate on a scale of 1-6 your interest in studying English (circle the number below). (1=*Not interested* to 6=*Strongly interested*)

Strongly interested					Not interested
1	2	3	4	5	6

Test-Taking Background

15. Please mark your relevance of the listed English language proficiency tests below.

Test	Awareness of the tests			Test Preparation	
	I don't know the test at all	I am somewhat familiar with the test	I am very familiar with the test	I am not preparing for the test	I am currently preparing for the test
TOEIC	☐	☐	☐	☐	☐
TOEIC Speaking	☐	☐	☐	☐	☐
TEPS	☐	☐	☐	☐	☐
TOEFL iBT	☐	☐	☐	☐	☐
IELTS	☐	☐	☐	☐	☐
OPIc	☐	☐	☐	☐	☐

16. Please mark the relevant box below regarding your test-taking experiences on the listed English language proficiency tests.

Test	Test-taking experience		Number of test-taking
	I have never taken the test	I have taken the test	
TOEIC	☐	☐	☐ 1-3 times ☐ 4-6 times ☐ 7-10 times

Table for question 16 (cont'd)

TOEIC Speaking	☐	☐	☐ 1-3 times ☐ 4-6 times ☐ 7-10 times
TEPS	☐	☐	☐ 1-3 times ☐ 4-6 times ☐ 7-10 times
TOEFL iBT	☐	☐	☐ 1-3 times ☐ 4-6 times ☐ 7-10 times
IELTS	☐	☐	☐ 1-3 times ☐ 4-6 times ☐ 7-10 times
OPIc	☐	☐	☐ 1-3 times ☐ 4-6 times ☐ 7-10 times

17. Please indicate your score on the relevant English language proficiency test below. If you have an OPIc score, please indicate the band level you were assigned.

Test	Scores/Band level
TOEIC	
TOEIC Speaking	
TEPS	
TOEFL iBT	
IELTS	
OPIc	

18. In what way did you prepare for the test(s) that you have indicated above?

Test	Test preparation method			
	A week – A month	A month – 3 months	3 months – 6 months	6 months – 1 year
Test preparation courses offered at my university	☐	☐	☐	☐

Table for question 18 (cont'd)

Test preparation courses at a private academy/institute	☐	☐	☐	☐
Private study groups	☐	☐	☐	☐
Online resources	☐	☐	☐	☐
Self-study	☐	☐	☐	☐
Other (specify): _____	☐	☐	☐	☐

19. If you have answered items on "test preparation courses," which type of test-taking strategies were you instructed on? Please mark everything that applies below.
    - ☐ Taking tests on computer
    - ☐ Constructing a key template of speaking response
    - ☐ Constructing a key template of speaking response
    - ☐ Note-taking strategies
    - ☐ Time management strategies (e.g., strategic use of responding & planning time)
    - ☐ Other (specify): _____

Appendix B

Elicited imitation task

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October 3-6). *An investigation of elicited imitation tasks in crosslinguistic SLA research.* Paper presented at the Second Language Research Forum, Toronto.

Instruction (given in Korean):

In this task, you'll be asked to repeat some sentences in Korean and some sentences in English. Please follow the instructions carefully. Please do not take any notes during this exercise. Now let's begin.
이 시험에서는 한국어 문장과 영어 문장을 듣고 문장을 소리내어 따라하는 능력을 테스트합니다. 주어진 설명을 잘 듣고 따라해 보세요. 시험을 치는 동안 노트필기는 할 수 없습니다.

You are going to hear several sentences in Korean. After each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, DON'T START REPEATING THE SENTENCE UNTIL YOU HEAR THE TONE SOUND {TONE}. Now let's begin.
지금부터 한국어 문장 6개를 듣게 됩니다. 각 문장이 끝난 후, 삐소리가 나면 들었던 문장을 따라 말해보세요. 문장을 부분적으로라도 최대한 많이 따라하는 것이 중요합니다. 문장을 반복할 시간은 삐소리 후 충분히 주어집니다. 삐소리를 듣기 전 문장을 반복하면 안됩니다. 지금부터 시험을 시작합니다.

나는 꽃이 좋다. (6 syllables) 3.9 seconds pause

[translation: I like flowers]

나는 편지를 쓴다. (7 syllables) 4.1 s

[translation: I write a letter]

나는 큰 차가 필요하다. (9 syllables) 5.6 s

[translation: I need a big car]

비가와서 밖에 안 나간다. (10 syllables) 6 s

[translation: As it is raining, I don't go out]

여자 아이는 넘어져서 다쳤다. (12 syllables) 7.2 s

[translation: The girl fell down and got hurt]

나는 집에 돌아오자마자 밥을 먹었다. (15 syllables) 9.5 s

[translation: As soon as I returned home, I ate meal]

That was the last Korean sentence

한국어 문장은 여기까지 입니다.

Now, you are going to hear 30 sentences in English. Once again, after each sentence, there will be a short pause, followed by a tone sound {TONE}. Your task is to try to repeat exactly what you hear in English. You will be given sufficient time after the tone to repeat the sentence. Repeat as much as you can. Remember, DON'T START REPEATING THE SENTENCE UNTIL YOU HEAR THE TONE SOUND {TONE}. Now let's begin.
지금부터는 30개의 영어문장을 듣게 됩니다. 이전과 마찬가지로 각 문장을 들은 후 삐소리가 나면, 들었던 영어 문장을 따라 말해 보세요. 문장을 부분적으로라도 최대한 많이 따라하는 것이 중요합니다. 삐소리를 듣기 전 문장을 반복하면 안됩니다. 지금부터 시험을 시작합니다.

1. I have to get a haircut (7) 3.6 seconds pause
2. The red book is on the table (8) 4.55 s
3. The streets in this city are wide (8) 4.8 s
4. He takes a shower every morning (9) 5.37 s
5. What did you say you were doing today? (10) 5.45 s
6. I doubt that he knows how to drive that well (10) 6.2 s
7. After dinner I had a long, peaceful nap (11) 6.8 s
8. It is possible that it will rain tomorrow (12) 6.74 s
9. I enjoy movies which have a happy ending (12) 7.22 s
10. The houses are very nice but too expensive (12) 7.92 s
11. The little boy whose kitten died yesterday is sad (13) 8.6 s
12. That restaurant is supposed to have very good food (13) 8.1 s
13. I want a nice, big house in which my animals can live (14) 9.3 s
14. You really enjoy listening to country music, don't you (14) 8.7 s
15. She just finished painting the inside of her apartment (14) 8.7 s
16. Cross the street at the light and then just continue straight ahead (15) 9.85 s
17. The person I'm dating has a wonderful sense of humor (15) 9.1 s
18. She only orders meat dishes and never eats vegetables (15/16) 10.1 s
19. I wish the price of town houses would become affordable (15) 9.4 s

20. I hope it will get warmer sooner this year than it did last year (16) 10.1 s
21. A good friend of mine always takes care of my neighbor's three children (16) 10.6 s
22. The black cat that you fed yesterday was the one chased by the dog(16) 10.8 s
23. Before he can go outside, he has to finish cleaning his room (16) 10.4 s
24. The most fun I've ever had was when we went to the opera (16) 10.2 s
25. The terrible thief whom the police caught was very tall and thin (17) 12.5 s
26. Would you be so kind as to hand me the book which is on the table? (17) 10.8 s
27. The number of people who smoke cigars is increasing every year (17/18) 11.1 s
28. I don't know if the 11:30 train has left the station yet (18) 11.2 s
29. The exam wasn't nearly as difficult as you told me it would be (18) 11.2 s
30. There are a lot of people who don't eat anything at all in the morning (19) 12 s

This is the end of the repetition task. Thank you.

문장반복시험이 끝났습니다. 감사합니다.

Scoring guidelines for Elicited Imitation task

SCORE 0

Criteria	Examples
• Nothing (Silence)	
• Garbled (unintelligible, usually transcribed as XXX)	
• Minimal repetition, then item abandoned: - Only 1 word repeated - Only 1 content word plus function word(s) - Only 1 content word plus function word(s) plus extraneous words that weren't in the original stimulus - Only function word(s) repeated **NOTE: with** only, just, yet (meaningful adverbs), score 1	- The- the street in... in... street... hmm (16/#2) - I wish... comfta-portable (19/#1) - I watch a movie (9/#22) - You don't... don't you? (14/#1) - He just finished (15/#23) (Closed word + Adv + lexical word) (score 1)

SCORE 1

Criteria	Examples
• When only about half of idea units are represented in the string but a lot of important information in the original stimulus is left out	- Cross the cross--cross the street ahead and. (16/#4) - I don't have nap (7/#1) - I ...the last year (20/#4) - I have to hair-haircu (1/#24) - Would you... the book on the table (26/#7)
• When barely half of lexical words get repeated and meaningful content results that is unrelated (or opposed) to stimulus, frequently with hesitation markers	- I wonder... why he... drive... well (6/#9) - He just finished painting... inside the park (15/#11)
• Or when string doesn't in itself constitute a self-standing sentence with some (targetlike or nontargetlike) meaning (This may happen more often with shorter items, where if only 2 of 3 content words are repeated and no grammatical relation between them is attempted, then score 1)	- I enjoy movie what shew have a... have a (9/#3) - She only eats vegetables and have xx- never eat vegetables (18/#4) - I want to big nice house.(13/#25) - A good frien of my take a good my chilren (21/#25) - I wannata ..... animalslive (13/#26) - Zu book .... table (2/#26) - I doubt he how to drive (6/#25)
• Also when half of a long stimulus is left out, and the sentence produced is incomplete	-The little boy the kitten... no.. is sad... I can't remember (11/#8) - Before... before he can go outside for (23/#11)

SCORE 2

Criteria	Examples
• When content of string preserves at least more than half of the idea units in the original stimulus; string in meaningful, and the meaning is close or related to original, but it departs from it in some slight changes in content, which makes content inexact, incomplete, or ambiguous	- The gooda friend take care o- chi- children (left out that it was the neighbor's children, and that they were three) (21/#1) - After dinner I have a long piece [peace?] of a nap (<a long, peaceful nap) (7/#4) - She just finished painting the seaside her apartment (<inside of) (15/#4) - The restaurant was supposed to have ve- good food (<is supposed; meaning changed to past) (12/#4) - I want to big house which... in which... animal can live (left out 'nice' 'my' and made animal into singular) (13/#4) - Would you hand me... the books which are on the table (<book; meaning changed to plural) (26/#4) - It is possible to day tomorrow (from pronunciation problem, it is ambiguous whether 'rain' has been understood, but it is possible) (8/#1)

SCORE 3

Criteria	Examples
• Original, complete meaning is preserved as in the stimulus. Strings which are quite ungrammatical can get a 3 score, as long as exact meaning is preserved. Some synonymous substitutions are acceptable.	- It is possible... the rain tomorrow (8/#11) - That restaurant ah.... supposed to... ah... very good food (12/#14)
• Examples of acceptable substitutions (SCORE 3): hand/give/pass are acceptable synonyms for item 26. Substitutions of and & but are acceptable. A lot of = many, etc. • Anything with or without 'very' can be considered synonymous.	- Would you pass me the book on the table (26/#21)(Score 3) - Would you be so kind...to bring... the book...on the table (26/#13)(Score 3) - The rest-restaurant is supposed to have **good food** (12/#11)(Score 3)
• Examples of unacceptable substitutions or omissions (SCORE 2): - cigar smoking> smoking - apartment >house/room - he<>she - sense of humor> humor - finished cleaning>cleaned - order> eat - nice,big > big	- The number of people who **smoke** ...um is increasing every year (27/#10)(Score 2) - **He** just finished painting... inside of a **his** house (15/#5)(Score 2) - She finished **a** painting... inside her apartment (15/#7)(Score 2) - The person I'm dading is ...**wonderful... humour** (17/#11)(Score 2)

- AUX cannot be omitted (can go> go) - a lot of Noun> 0 Noun -too Adj > 0 Adj	- Before he get outside...he must **clean** his room (23/#9)(Score 2) - She always **eat**...meat...nev-never eat vegetable (18/#5)(Score 2)
• Changes in grammar that don't affect meaning should be scored as 3. For instance, failure to supply past tense (had>have) and missing articles should be considered grammar change only (score 3).	- After dinner I **have** a long peaceful nap. (7/#17)(Score 3)
• By contrast, cases of extra marking or more marked morphology should be considered as meaning change. For example, a present tense repeated as past or as future should be scored as meaning change (score 2).	- The restaurant **was** supposed to have ve- good food.(12/#24)(Score 2) - After the dinner I **will have** a long... sp- peaceful nap. (7/#8)(Score 2)
• Similarly, singular/plural differences between stimulus and repeated string change the meaning, not only the grammar (score 2).	- The **street** in the city is wide (3/#8)(Score 2)
• Changes of person (he for she or she for he) change the meaning; but problems of agreement (she...her versus she...his) should be considered grammatical change, not meaning change.	- **She** just finished painting ...**his** room inside (15/#14) (Score 2)(apartment is missing)
• Ambiguous changes in grammar that COULD be interpreted as meaning changes from a NS perspective should be scored as 2. That is, as a general principle in case of doubt about whether meaning has changed or not, score 2.	- The street**s** on the city **is** wide (3/#23)(Score 2) (We can't know whether the number agreement is just a grammar problem or an interpretation problem, but string is ambiguous in meaning: (a) a generic plural statement or (b) a statement about one street (score 2).

SCORE 4

Criteria	**Examples**
• Exact repetition: String matches stimulus exactly. Both form and meaning are correct without exception or doubt.	

Appendix C

Test tasks

Test Set A

Task 1

Directions: You will now be asked to give your opinion about a familiar topic. Give yourself 15 seconds to prepare your response. Then record yourself speaking for 45 seconds.

> Choose a place you go to often that is important to you and explain why it is important. Please include specific details in your explanation.

Task 2

Directions: You will not read a short passage and then listen to a conversation on the same topic. You will then be asked a question about them. After you hear the question, you will have 30 seconds to prepare you response and 60 seconds to speak.

Give yourself 45 seconds to read the article.

Bus Service Elimination Planned

The university has decided to discontinue its free bus service for students. The reason given for this decision is that few students ride the buses and the buses are expensive to operate. Currently, the buses run from the center of campus past university buildings and through some of the neighborhoods surrounding the campus. The money saved by eliminating the bus service will be used to expand the over-crowded student parking lots.

Directions: Now listen to two students discussing the article.

Male student: I don't like the university's plan.

Female student: Really? I've ridden those buses, and sometimes there were only a few people on the bus.

Male student: I see your point. But I think the problem is the route's out of date. It only gets through the neighborhoods that've gotten too expensive for students to live in. It's ridiculous that they haven't already changed the route – you know, so it goes where most off-campus students live now. I bet if they did that, they'd get plenty of students riding those buses.

Female student: Well, at least they're adding more parking. It's gotten really tough to find a space.

Male student: That's the other part I don't like, actually. Cutting back the bus service and adding parking's just gonna encourage more students to drive on campus. And that'll just add to the noise around campus and create more traffic…and that'll increase the need for more parking spaces.

Female student: Yeah, I guess I can see your point. Maybe it would be better if more students used the buses instead of driving.
Male student: Right. And the university should make it easier to do that, not harder.

Directions: Give yourself 30 seconds to prepare your response to the following question. Then record yourself speaking for 60 seconds.

The man expresses his opinion of the university's plan to eliminate the bus service. State his opinion and explain the reasons he gives for holding that opinion.

Task 3

Directions: Now listen to part of a lecture in a economics class.

Professor: So let's talk about money. What is money? Well, typically people think of coins and paper "bills" as money, but that's using a somewhat narrow definition of the term. A broad definition is this: money is anything that people can use to make purchases with. Since many things can be used to make purchases, money can have many different forms. Certainly, coins and bills are one form of money. People exchange goods and services for coins or paper bills, and they use this money, these bills to obtain other goods and services. For example, you might give a taxi driver five dollars to purchase a ride in his taxi. And he in turn gives the five dollars to a farmer to buy some vegetables.

But, as I said, coins and bills aren't the only form of money under this broad definition. Some societies make use of a barter system. Basically, in a barter system people exchange goods and services directly for other goods and services. The taxi driver, for example, might give a ride to a farmer in exchange for some vegetables. Since the vegetables are used to pay for a service, by our broad definition the vegetables are used in barter as a form of money.

Now, as I mentioned, there's also a second, a narrower definition of money. In the United States only coins and bills are legal tender—meaning that by law, a seller must accept them as payment. The taxi driver must accept coins or bills as payment for a taxi ride. OK? But in the U.S., the taxi driver is not required to accept vegetables in exchange for a ride. So a narrower definition of money might be whatever is legal tender in a society, whatever has to be accepted as payment.

Directions: Give yourself 20 seconds to prepare your response to the following question. Then record yourself speaking for 60 seconds.

Using the points and examples from the lecture, explain the two definitions of money presented by the professor.

Test set B
Task 1

Directions: You will now be asked to give your opinion about a familiar topic. Give yourself 15 seconds to prepare your response. Then record yourself speaking for 45 seconds.

> What kind of reading material, such as novels, magazines, or poetry, do you most like to read in your free time? Explain why you find this kind of reading material interesting.

Task 2

Directions: Read a passage about a topic in psychology. You will have 45 seconds to read the passage. Begin reading now.

Actor-observer
People account for their own behavior differently from how they account for the behavior of others. When observing the behavior of others, we tend to attribute their actions to their character or their personality rather than to external factors. In contrast, we tend to explain our own behavior in terms of situational factors beyond our own control rather than attributing it to our own character. One explanation for this difference is that people are aware of the situational forces affecting them but not of situational forces affecting other people. Thus, when evaluating someone else's behavior, we focus on the person rather than the situation.

Directions: Now listen to part of a lecture in a psychology class.

Professor: So we encounter this in life all the time, but many of us are unaware that we do this. Even psychologists who study it, like me. For example, the other day I was at the store and I was getting in line to buy something. But just before I was actually in line, some guy comes out of nowhere and cuts right in front of me. Well, I was really annoyed and thought, "That was rude!" I assumed he was just a selfish, inconsiderate person when, in fact, I had no idea why he cut in line in front of me or whether he even realized he was doing it. Maybe he didn't think I was actually in line yet…But my immediate reaction was to assume he was a selfish or rude person. OK, so a few days after that, I was at the store again. Only this time I was in a real hurry—I was late for an important meeting—and I was frustrated that everything was taking so long. And what's worse, all the checkout lines were long, and it seemed like everyone was moving so slowly. But then I saw a slightly shorter line! But some woman with a lot of stuff to buy was walking toward it, so I basically ran to get there first, before her, and well, I did. Now, I didn't think of myself as a bad or rude person for doing this. I had an important meeting to get to—I was in a hurry, so, you know, I had done nothing wrong.

Directions: Give yourself 30 seconds to prepare your response to the following question. Then record yourself speaking for 60 seconds.

Explain how the two examples discussed by the professor illustrate differences in the ways people explain behavior.

Task 3

Direction: Now listen to a conversation between a student and her advisor.

Advisor: OK, Becky, so, you've chosen all your courses for next term?
Student: Well, not really, professor. Actually, I've got a problem.
Advisor: Oh?
Advisor: Yeah, well, I still need to take an American literature course; it's required for graduation. But I've been putting it off. But since my next term is my last…
Advisor: Yeah, you can't put it off any longer!
Student: Right. The thing is though, it's not offered next term.
Advisor: I see. Hmm. Ah, how about, ah, taking the course at another university?
Student: I thought about that. It's offered at City College, but, that's so far away. Commuting back and forth would take me a couple of hours, you know, a big chunk of time with all my other studies and everything.
Advisor: True, but it's been done. Or, ah, there are a couple of graduate courses in American literature. Why not take one of those?
Student: Yeah, but, wouldn't that be hard, though? I mean, it's a graduate course; that'd be pretty intense.
Advisor: Yeah, it'd probably mean more studying than you're used to, but I'm sure it's not beyond your abilities.

Directions: Give yourself 20 seconds to prepare your response to the following question. Then record yourself speaking for 60 seconds.

The speakers discuss two possible solutions to the woman's problem. Briefly summarize the problem. Then state which solution you recommend and explain why.

Test set C

Task 1

Directions: You will now be asked to give your opinion about a familiar topic. Give yourself 15 seconds to prepare your response. Then record yourself speaking for 45 seconds.

> Some students prefer to work on class assignments by themselves. Others believe it is better to work in a group. Which do you prefer? Explain why.

Task 2

Directions: The university's Dining Services Department has announced a change. Read an announcement about this change. You will have 45 seconds to read the announcement. Begin reading now.

Hot Breakfasts Eliminated
Beginning next month, Dining services will no longer serve hot breakfast foods at university dining halls. Instead, students will be offered a wide assortment of cold breakfast items in the morning. These cold breakfast foods, such as breads, fruit, and yogurt, are healthier than many of the hot breakfast items that we will stop serving, so health-conscious students should welcome this change. Students will benefit in another way as well, because limiting the breakfast selection to cold food items will save money and allow us to keep our meal plans affordable.

Direction: Now listen to two students discussing the announcement.

Female Student: Do you believe any of this? It's ridiculous.
Male Student: What do you mean? It is important to eat healthy foods.
Female Student: Sure it is, but they're saying yogurt's better for you than an omelet, or than hot cereal? I mean whether something's hot or cold, that shouldn't be the issue. Except maybe on a really cold morning, but in that case, which is going to be better for you—a bowl of cold cereal or a nice warm omelet? It's obvious; there's no question.
Male Student: I'm not going to argue with you there.
Female Student: And this whole thing about saving money.
Male Student: What about it?
Female Student: Well, they're actually going to make things worse for us, not better. 'Cause if they start cutting back and we can't get what we want right here, on campus, well, we're going to be going off campus and pay off-campus prices, and you know what? That will be expensive. Even if it's only two or three mornings a week, it can add up.

Directions: Give yourself 30 seconds to prepare your response to the following question. Then record yourself speaking for 60 seconds.

The woman expresses her opinion of the change that has been announced. State her opinion and explain her reasons for holding that opinion.

Task 3

Directions: Now listen to part of a lecture in a psychology class. The professor is discussing advertising strategies.

Professor: In advertising, various strategies are used to persuade people to buy product. In order to sell more products, advertisers will often try to make us believe that a product will meet our needs or desires perfectly, even if it's not true. The strategies they use can be subtle, uh, "friendly" forms of persuasion that are sometimes hard to recognize.

In a lot of ads, repetition is a key strategy. Research shows that repeated exposure to a message, even something meaningless or untrue, is enough to make people accept it or see it in a positive light. You've all seen the car commercials on TV, like, uh, the one that refers to its "roomy" cars, over and over again. You know which one I mean. This guy is driving around and he keeps stopping to pick up different people—he picks up 3 or 4 people. And each time, the narrator says, "Plenty of room for friends, plenty of room for family, plenty of room for everybody." The same message is repeated several times in the course of the commercial. Now, the car, uh, the car actually looks kind of small. It's not a very big car at all, but you get the sense that it's pretty spacious. You'd think that the viewer would reach the logical conclusion that the slogan, uh, misrepresents the product. Instead, what usually happens is that when the statement "plenty of room" is repeated often enough, people are actually convinced it's true.

Um, another strategy they use is to get a celebrity to advertise a product. It turns out that we're more likely to accept an advertising claim made by somebody famous—a person we admire and find appealing. We tend to think they're trustworthy. So, um, you might have a car commercial that features a well-known race car driver. Now, it may not be a very fast car—uh, it could even be an inexpensive vehicle with a low performance rating. But if a popular race car driver is shown driving it, and saying, "I like my cars fast!" then people will believe the car is impressive for its speed.

Directions: Give yourself 20 seconds to prepare your response to the following question. Then record yourself speaking for 60 seconds.

Using points and examples from the lecture, explain how persuasive strategies are used in advertising.

Appendix D

Post questionnaire

After Guided Planning Conditions

1. Please mark your confidence in performing each of the three test-tasks below.

Test	Confidence				
	I was not confident at all.	I was somewhat not confident.	I was confident on an average level.	I was fairly confident.	I was completely confident.
Independent task	☐	☐	☐		
Integrated task: Reading & Listening	☐	☐	☐		
Integrated task: Listening	☐	☐	☐		

2. Please mark your perceptions on the appropriateness of the planning time per test-task type.

Test	Appropriateness of planning time				
	It was not sufficient at all.	It was somewhat not sufficient.	It was just right.	It was fairly sufficient.	It was excessively sufficient.
Independent task	☐	☐	☐		
Integrated task: Reading & Listening	☐	☐	☐		
Integrated task: Listening	☐	☐	☐		

3. To what extent did you use the **particular planning activity** for which test-task type?

Test	Usefulness		
	I did not use in all cases.	I fairly used the planning activity.	I used the planning activity a lot.
Independent task	☐	☐	☐
Integrated task: Reading & Listening	☐	☐	☐
Integrated task: Listening	☐	☐	☐

4. To what extent was the **particular planning activity** effective/useful for performing which test-task type?

Test	Usefulness				
	It was not useful at all.	It was somewhat not useful.	It was useful on an average level.	It was fairly useful.	It was completely useful.
Independent task	☐	☐	☐	☐	☐
Integrated task: Reading & Listening	☐	☐	☐	☐	☐
Integrated task: Listening	☐	☐	☐	☐	☐

5. What are the pros and cons for using the **particular planning activity**?

_____

_____

_____

6. Have you practiced the **particular planning activity** you have just done before?
☐ Yes ☐ No

7. Have any of your teachers taught you how to plan before speaking? ☐ Yes ☐ No
If yes, was the planning activity you just did now taught by your teachers? ☐ Yes ☐ No

After Unguided Planning Conditions

1. Please mark your confidence in performing each of the three test-tasks below.

Test	Confidence				
	I was not confident at all.	I was somewhat not confident.	I was confident on an average level.	I was fairly confident.	I was completely confident.
Independent task	☐	☐	☐		
Integrated task: Reading & Listening	☐	☐	☐		
Integrated task: Listening	☐	☐	☐		

2. Please mark your perceptions on the appropriateness of the planning time per test-task type.

Test	Appropriateness of planning time				
	It was not sufficient at all.	It was somewhat not sufficient.	It was just right.	It was fairly sufficient.	It was excessively sufficient.
Independent task	☐	☐	☐		
Integrated task: Reading & Listening	☐	☐	☐		
Integrated task: Listening	☐	☐	☐		

3. To what extent did you use the **particular planning activity** for which test-task type?

Test	Usefulness		
	I did not use in all cases.	I fairly used the planning activity.	I used the planning activity a lot.
Independent task	☐	☐	☐
Integrated task: Reading & Listening	☐	☐	☐
Integrated task: Listening	☐	☐	☐

4. To what extent was the **particular planning activity** effective/useful for performing which test-task type?

Test	Usefulness				
	It was not useful at all.	It was somewhat not useful.	It was useful on an average level.	It was fairly useful.	It was completely useful.
Independent task	☐	☐	☐	☐	☐
Integrated task: Reading & Listening	☐	☐	☐	☐	☐
Integrated task: Listening	☐	☐	☐	☐	☐

5. Indicate (by ticking all relevant boxes) which of the following things you did during your planning time before you started speaking.

	With 15 seconds	With 20 seconds	With 30 seconds
I thought about grammar in my head.	☐	☐	☐
I practiced useful sentences or phrases in my head.	☐	☐	☐
I wrote down useful sentences or phrases on paper.	☐	☐	☐
I made a list of vocabulary in my head.	☐	☐	☐
I wrote down vocabulary in my head.	☐	☐	☐
I made a list of useful organizing and/or linking language in my head.	☐	☐	☐
I wrote down useful organizing and/or linking language on paper.	☐	☐	☐
I practiced pronunciation in my head.	☐	☐	☐
I tried to decide what topic I would talk about.	☐	☐	☐
I thought about how to organize my ideas.	☐	☐	☐
I thought about the content and ideas needed for the question.	☐	☐	☐

Table for question 5 (cont'd)

I wrote down ideas in my first language & then translated them.	☐	☐	☐
I thought about nothing.	☐	☐	☐
I did other things (please tell me what you did)	☐	☐	☐

Other (specify):

Appendix E

One-on-one interview questions

1. Which planning activities were most helpful in performing the task?
2. Did you find the planning times appropriate for each test-task type? Say why or why not.
3. Do you think you used the planning time as well as you could have? Say why/ why not.
4. Have you ever been given instruction/training on how to use planning time? If yes, how useful was it? If no, do you think it would help to have this kind of training?

Appendix F

Scoring rubric for speaking tasks

Task 1 (Independent task)

Score	General Description	Delivery	Language Use	Topic Development
2	The response addresses the task, but development of the topic is limited. It contains intelligible speech, although problems with delivery and/or overall coherence occur; meaning may be obscured in places. A response at this level is characterized by at least two of the following:	Speech is basically intelligible, though listener effort is needed because of unclear articulation, awkward intonation, or choppy rhythm/ pace; meaning may be obscured in places.	The response demonstrates limited range and control of grammar and vocabulary. These limitations often prevent full expression of ideas. For the most part, only basic sentence structures are used successfully and spoken with fluidity. Structures and vocabulary may express mainly simple (short) and/ or general propositions, with simple or unclear connections made among them (serial listing, conjunction, juxtaposition).	The response is connected to the task, though the number of ideas presented or the development of ideas is limited. Mostly basic ideas are expressed, with limited elaboration (details and support). At times relevant substance may be vaguely expressed or repetitious. Connections of ideas may be unclear.
1	The response is very limited in content and/or coherence or is only minimally connected to the task, or speech is largely unintelligible. A response at this level is characterized by at least two of the following:	Consistent pronunciation, stress, and intonation difficulties cause considerable listener effort; delivery is choppy, fragmented, or telegraphic; frequent pauses and hesitations.	Range and control of grammar and vocabulary severely limit (or prevent) expression of ideas and connections among ideas. Some low-level responses may rely heavily on practiced or formulaic expressions.	Limited relevant content expressed. The response generally lacks substance beyond expression of very basic ideas. Speaker may be unable to sustain speech to complete the task and may rely heavily on repetition of the prompt.
0	Speaker makes no attempt to respond OR response is unrelated to the topic.			

Task 2 and Task 3 (Integrated task)

Score	General Description	Delivery	Language Use	Topic Development
4	The response fulfills the demands of the task, with at most minor lapses in completeness. It is highly intelligible and exhibits sustained, coherent discourse. A response at this level is characterized by all of the following:	Speech is generally clear, fluid, and sustained. It may include minor lapses or minor difficulties with pronunciation or intonation. Pace may vary at times as speaker attempts to recall information. Overall intelligibility remains high.	The response demonstrates good control of basic and complex grammatical structures that allow for coherent, efficient (automatic) expression of relevant ideas. Contains generally effective word choice. Though some minor (or systematic) errors or imprecise use may be noticeable, they do not require listener effort (or obscure meaning).	The response presents a clear progression of ideas and conveys the relevant information required by the task. It includes appropriate detail, though it may have minor errors or minor omissions.
3	The response addresses the task appropriately, but may fall short of being fully developed. It is generally intelligible and coherent, with some fluidity of expression, though it exhibits some noticeable lapses in the expression of ideas. A response at this level is characterized by at least two of the following:	Speech is generally clear, with some fluidity of expression, but it exhibits minor difficulties with pronunciation, intonation, or pacing and may require some listener effort at times. Overall intelligibility remains good, however.	The response demonstrates fairly automatic and effective use of grammar and vocabulary and fairly coherent expression of relevant ideas. Response may exhibit some imprecise or inaccurate use of vocabulary or grammatical structures or be somewhat limited in the range of structures used. Such limitations do not seriously interfere with the communication of the message.	The response is sustained and conveys relevant information required by the task. However, it exhibits some incompleteness, inaccuracy, lack of specificity with respect to content, or choppiness in the progression of ideas.

Appendix G

Elder and Iwashita's (2005) rating scales on fluency, accuracy, and complexity

Fluency

5	Speaks without hesitation; speech is generally of a speed similar to a native speaker.
4	Speaks fairly fluently with only occasional hesitation, false starts and modification of attempted utterance. Speech is only slightly slower than that of a native speaker.
3	Speaks more slowly than a native speaker due to hesitations and word-finding delays.
2	A marked degree of hesitation due to word-finding delays or inability to phrase utterances easily.
1	Speech is quite disfluent due to frequent and lengthy hesitations or false starts.

Accuracy

5	Errors are barely noticeable.
4	Errors are not unusual, but rarely major.
3	Manages most common forms, with occasional errors; major errors present.
2	Limited linguistic control: major errors frequent.
1	Clear lack of linguistic control even of basic forms.

Complexity

5	Confidently attempts a variety of verb forms (e.g., passives, modals, tense and aspect), even if the use is not always correct. Regularly takes risks grammatically in the service of expressing complex meaning. Routinely attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is occasionally awkward or incorrect.
4	Attempts a variety of verb forms, even if the use is not always correct. Takes risks grammatically in the service of expressing complex meaning. Regularly attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is awkward or incorrect.
3	Mostly relies on simple verb forms, with some attempt to use a greater variety of forms. Some attempts to use coordination and subordination to convey ideas that cannot be expressed in a single clause.
2	Produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty.
1	Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect idea across clauses.

Appendix H

Basic descriptive statistics for the raw coding data

Test Set A for *fluency* indices

Fluency	Test Set A								
	UG (N = 32)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Filled Pauses (Num)	3.36 (1.92)	4.80 (3.14)	4.27 (2.66)	4.20 (3.22)	4.80 (4.37)	5.18 (4.36)	3.03 (2.84)	4.10 (3.10)	4.16 (3.53)
Unfilled Pauses (Num)	4.83 (2.41)	5.13 (2.03)	5.89 (2.42)	6.27 (2.45)	6.00 (3.07)	5.73 (2.04)	5.82 (2.14)	5.92 (2.44)	5.86 (2.38)
Reformulations (Num)	0.78 (0.78)	1.13 (0.78)	1.03 (0.87)	0.79 (0.87)	1.02 (0.84)	0.79 (0.75)	0.76 (0.71)	1.27 (0.83)	1.19 (0.73)
Repetitions (Num)	0.70 (0.86)	1.14 (1.32)	1.10 (1.02)	0.83 (0.97)	1.20 (1.17)	0.97 (0.94)	1.18 (0.88)	1.15 (1.00)	1.33 (1.13)
Replacements (Num)	0.50 (0.67)	0.77 (0.87)	1.02 (0.88)	0.56 (0.95)	1.03 (1.12)	0.61 (0.69)	0.68 (0.76)	1.05 (1.12)	1.00 (0.92)
Hesitations (Num)	0.50 (0.77)	0.55 (0.78)	0.65 (0.71)	0.26 (0.44)	0.59 (0.72)	0.35 (0.66)	0.27 (0.45)	0.58 (0.94)	0.48 (0.65)
False starts (Num)	0.33 (0.53)	0.27 (0.44)	0.31 (0.49)	0.26 (0.61)	0.39 (0.54)	0.21 (0.48)	0.27 (0.45)	0.36 (0.53)	0.25 (0.42)

Test Set B for *fluency* indices

Fluency	Test Set B								
	UG (N = 33)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Filled Pauses (Num)	3.49 (2.46)	4.80 (3.04)	4.27 (2.93)	4.06 (3.23)	5.44 (3.88)	4.92 (4.49)	3.82 (2.25)	4.17 (3.82)	3.30 (2.95)
Unfilled Pauses (Num)	6.25 (3.13)	6.09 (2.84)	6.39 (2.57)	7.24 (2.54)	7.73 (2.27)	7.48 (2.30)	7.09 (3.38)	7.18 (3.18)	7.18 (3.32)
Reformulations (Num)	0.70 (0.68)	0.97 (0.85)	1.12 (0.64)	0.73 (0.67)	1.12 (0.78)	1.02 (0.81)	1.11 (0.73)	1.30 (1.23)	1.17 (0.67)
Repetitions (Num)	0.91 (0.74)	1.15 (1.03)	1.15 (1.00)	0.88 (1.06)	1.23 (1.22)	1.09 (1.03)	0.80 (0.84)	0.58 (0.71)	0.64 (1.85)
Replacements (Num)	0.67 (0.65)	1.18 (1.41)	0.96 (0.92)	0.58 (0.66)	0.92 (1.10)	0.73 (0.79)	0.80 (0.84)	0.58 (0.71)	0.82 (1.07)
Hesitations (Num)	0.20 (0.39)	0.44 (0.66)	0.36 (0.60)	0.17 (0.41)	0.42 (0.57)	0.41 (0.57)	0.33 (0.54)	0.70 (1.01)	0.41 (0.63)
False starts (Num)	0.10 (0.26)	0.33 (0.46)	0.18 (0.41)	0.10 (0.29)	0.29 (0.45)	0.27 (0.52)	0.14 (0.34)	0.27 (0.45)	0.18 (0.37)

Test Set C for *fluency* indices

Fluency	Test Set C								
	UG (N = 32)			GW (N = 33)			GT (N = 32)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Filled Pauses (Num)	3.59 (3.23)	4.80 (3.00)	4.44 (3.04)	3.97 (3.43)	5.21 (5.15)	5.09 (4.71)	3.03 (2.60)	3.47 (3.86)	4.03 (4.10)
Unfilled Pauses (Num)	6.23 (2.24)	6.70 (2.86)	6.14 (2.95)	6.73 (2.08)	7.12 (3.04)	5.99 (2.80)	6.47 (2.76)	7.13 (2.78)	6.41 (2.85)
Reformulations (Num)	0.89 (0.78)	1.23 (0.81)	1.03 (0.82)	1.05 (0.54)	1.02 (0.78)	1.26 (1.06)	0.78 (0.61)	1.11 (0.90)	1.00 (0.72)
Repetitions (Num)	1.00 (1.02)	1.39 (0.92)	1.48 (1.17)	0.71 (0.84)	0.89 (0.79)	0.68 (0.79)	0.42 (0.71)	1.09 (0.89)	0.73 (0.72)
Replacements (Num)	0.80 (0.90)	1.13 (1.29)	0.98 (0.85)	0.58 (0.79)	0.67 (0.89)	0.80 (0.83)	0.47 (0.66)	0.64 (0.89)	0.75 (1.02)
Hesitations (Num)	0.36 (0.65)	0.30 (0.52)	0.44 (0.67)	0.42 (0.90)	0.42 (0.74)	0.39 (0.83)	0.25 (0.66)	0.52 (0.83)	0.45 (0.73)
False starts (Num)	0.22 (0.55)	0.19 (0.40)	0.20 (0.40)	0.12 (0.33)	0.08 (0.25)	0.14 (0.34)	0.03 (0.18)	0.13 (0.34)	0.34 (0.48)

Test Set A for *accuracy* indices

Accuracy	Test Set A								
	UG (N = 32)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Error-free clauses (Num)	2.39 (1.34)	2.28 (1.46)	2.90 (1.25)	2.05 (1.40)	2.05 (1.39)	2.61 (1.29)	2.05 (1.34)	2.15 (1.56)	2.03 (1.49)
Lexical errors (Num)	1.58 (0.76)	1.70 (0.94)	1.69 (0.73)	1.77 (1.07)	1.79 (1.20)	1.76 (1.00)	1.35 (1.06)	1.50 (1.00)	1.42 (0.88)

Test Set B for *accuracy* indices

Accuracy	Test Set B								
	UG (N = 33)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Error-free clauses (Num)	2.18 (1.26)	2.27 (1.46)	3.03 (1.47)	2.58 (1.80)	2.21 (1.39)	3.12 (2.06)	2.85 (1.40)	2.94 (1.30)	3.58 (1.92)
Lexical errors (Num)	1.24 (0.93)	1.58 (0.84)	1.70 (0.82)	1.30 (0.93)	1.47 (0.95)	1.47 (0.93)	1.03 (0.84)	1.38 (1.19)	1.56 (1.37)

Test Set C for *accuracy* indices

Accuracy	Test Set C								
	UG (N = 32)			GW (N = 33)			GT (N = 32)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Error-free clauses (Num)	3.13 (1.63)	3.05 (1.30)	3.14 (1.95)	2.41 (1.18)	2.97 (1.57)	3.49 (1.53)	3.16 (1.71)	3.23 (1.95)	3.28 (1.67)
Lexical errors (Num)	1.25 (0.95)	1.22 (1.07)	1.44 (1.03)	1.00 (0.97)	1.23 (1.23)	0.97 (1.34)	0.72 (0.89)	0.83 (0.99)	0.95 (1.23)

Test Set A for *complexity* indices

Complexity	Test Set A								
	UG (N = 32)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
AS-Units (Num)	2.75 (1.24)	3.47 (1.53)	3.31 (1.24)	2.99 (1.34)	3.85 (1.71)	3.67 (1.68)	2.85 (1.14)	3.35 (1.43)	3.11 (1.18)
Subordinate clauses (Num)	1.66 (1.00)	2.03 (1.07)	2.03 (1.05)	1.47 (0.98)	1.73 (1.16)	2.02 (1.42)	1.64 (0.86)	1.68 (1.14)	1.80 (1.03)

Test Set B for *complexity* indices

Complexity	Test Set B								
	UG (N = 33)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
AS-Units (Num)	2.91 (1.07)	3.61 (1.32)	4.08 (1.74)	3.09 (1.38)	3.46 (1.49)	4.23 (1.93)	2.77 (1.08)	3.49 (1.18)	4.11 (1.46)
Subordinate clauses (Num)	1.27 (0.84)	1.38 (0.82)	1.86 (1.25)	1.42 (1.00)	1.65 (1.12)	2.35 (1.15)	1.29 (1.22)	1.71 (1.06)	2.09 (1.22)

Test Set C for *complexity* indices

Accuracy	Test Set C								
	UG (N = 32)			GW (N = 33)			GT (N = 32)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
AS-Units (Num)	2.88 (0.98)	3.39 (1.17)	3.89 (1.24)	2.85 (0.97)	3.30 (1.47)	3.74 (1.47)	2.73 (1.07)	3.54 (1.10)	3.52 (1.06)
Subordinate clauses (Num)	0.81 (0.78)	1.47 (1.04)	2.20 (1.42)	1.20 (1.09)	1.76 (1.25)	1.97 (1.38)	1.17 (0.85)	1.56 (1.01)	1.58 (0.87)

Test Set A for *CAF ratings*

Rating category	Test Set A								
	UG (N = 32)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Accuracy	3.48 (1.01)	3.45 (0.96)	3.42 (1.02)	3.38 (1.08)	3.24 (0.90)	3.27 (0.91)	3.59 (0.80)	3.79 (0.89)	3.72 (1.02)
Complexity	3.34 (1.04)	3.38 (1.03)	3.31 (1.24)	3.27 (1.04)	3.59 (0.98)	3.38 (1.00)	3.73 (0.79)	3.79 (0.84)	3.75 (0.92)
Fluency	3.53 (1.13)	3.50 (1.11)	3.34 (1.16)	3.35 (1.24)	3.55 (1.07)	3.30 (1.13)	3.68 (0.94)	3.99 (0.84)	3.81 (0.87)

Test Set B for *CAF ratings*

Rating category	Test Set A								
	UG (N = 33)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Accuracy	3.45 (0.71)	3.44 (0.77)	3.49 (0.66)	3.44 (0.87)	3.38 (0.96)	3.47 (0.90)	3.41 (0.78)	3.59 (0.91)	3.73 (0.81)
Complexity	3.52 (0.76)	3.61 (0.76)	3.74 (0.77)	3.41 (0.97)	3.36 (1.02)	3.55 (0.91)	3.38 (0.76)	3.56 (0.82)	3.73 (0.74)
Fluency	3.42 (0.88)	3.65 (0.70)	3.64 (0.77)	3.29 (1.07)	3.29 (1.09)	3.49 (0.95)	3.44 (0.84)	3.53 (0.85)	3.77 (0.70)

Test Set C for *CAF ratings*

Rating category	Test Set C								
	UG (N = 33)			GW (N = 33)			GT (N = 33)		
	IP	IT-RL	IT-L	IP	IT-RL	IT-L	IP	IT-RL	IT-L
Accuracy	3.48 (1.01)	3.45 (0.96)	3.42 (1.02)	3.38 (1.08)	3.24 (0.90)	3.27 (0.91)	3.59 (0.80)	3.79 (0.89)	3.72 (1.02)
Complexity	3.34 (1.04)	3.38 (1.03)	3.31 (1.24)	3.27 (1.04)	3.59 (0.98)	3.38 (1.00)	3.73 (0.79)	3.79 (0.84)	3.75 (0.92)
Fluency	3.53 (1.13)	3.50 (1.11)	3.34 (1.16)	3.35 (1.24)	3.55 (1.07)	3.30 (1.13)	3.68 (0.94)	3.99 (0.84)	3.81 (0.87)

REFERENCES

REFERENCES

AERA, APA, NCME. (2014). *Standards for educational and psychological testing*. Washington, DC: American Education Research Association.

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

Bachman, L. F. (2004). *Statistical analyses for language assessment*. Cambridge, England: Cambridge University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice*. Oxford: Oxford University Press.

Batstone, R. (2005). Planning as discourse activity. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp. 277-296). Amsterdam: John Benjamins.

Barkaoui, K., Brooks, L., Swain, M., & Lapkin, S. (2012). Test-takers' strategic behaviors in independent and integrated speaking tasks. *Applied Linguistics, 34*, 304-324.

Bechger, T. M., Maris, G., & Hsiao, Y. P. (2010). Detecting halo effects in performance-based examinations. *Applied Psychological Measurement, 34*, 607–619.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: the case of systematicity. *Language Learning, 33*(1), 1-17.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

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

Bowles, M. A. (2011). Measuring implicit and explicit knowledge: What can heritage learners contribute? *Studies in Second Language Acquisition, 33*, 247-271.

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

Brown, G., Anderson, A., Shilcock, R., & Yule, G. (1984). *Teaching talk: Strategies for production and assessment.* Cambridge: Cambridge University Press.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, V. Kuiken & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity,*

accuracy and fluency in SLA (pp. 21-46). Amsterdam: John Benjamins.

Butler, F., Eignor, D., Jones, S., McNamara, T., & Suomi, B. (1999). *TOEFL 2000 speaking framework: A working paper.* (TOEFL Monograph Series Report No. 20). Princeton, NJ: Educational Testing Service.

Butterworth, B. (1975). Hesitation and semantic planning in speech. *Journal of Psycholinguistic Research, 4*, 75–87.

Bygate, M. (1996). Effects of task repetitions: Appraising the developing language of learners. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 136-146). Oxford: Heinemann.

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6*, 284–290.

Cheng, L., & DeLuca, C. (2011) Voices from test-takers: further evidence for language assessment validation and use. *Educational Assessment, 16*(2), 104-122.

Choi, I.-C. (2008). The impact of EFL testing on EFL education in Korea. *Language Testing, 25*(1), 39-62.

Cox, T. L., Bown, J., & Burdis, J. (2015). Exploring proficiency-based vs. performance-based items with elicited imitation assessment. *Foreign Language Annals, 48*(3), 350-371.

Crisp, V. (2012). An investigation of rater cognition in the assessment of projects. *Educational Measurement: Issues and Practice, 31*(3), 10-20.

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

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.

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227-1237.

Cublio, J., & Winke, P. (2013). Redefining the L2 listening construct within an integrated writing task: Considering the impacts of visual-cue interpretation and note-taking. *Language Assessment Quarterly, 10*(4), 371-397

Davis, A., Brown, A., Elder, C., Hill, K., Lumley, T., & McNamara, T. (1999). *Studies in language testing: Dictionary of language testing*. Cambridge: Cambridge University Press.

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

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics, 36*(2), 223-243.

Eckes, T. (2009). Many-facet Rasch measurement. *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment.*

Educational Testing Service (2012). *The official guide to the TOEFL®* test. Princeton, NJ: Educational Testing Service.

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

Elder, C., Iwashita, N., & McNamara, T. (2002). Estimating the difficulty of oral proficiency tasks: What does the test-taker have to offer? *Language Testing, 19*, 347-368.

Ellis, R. (1987). Interlanguage variability in narrative discourse: Style shifting in the use of the past tense. *Studies in Second Language Acquisition, 9*, 1–20.

Ellis, R. (2005). *Planning and task performance in a second language.* Amsterdam: John Benjamins.

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

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied Linguistics, 27*(3), 464–491.

Ferrari, S. (2012). A longitudinal study of complexity, accuracy and fluency variation in second language development. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 277–297). Philadelphia: John Benjamins.

Field, A., Miles, J., Field, Z. (2012). *Discovering statistics using R*. London: Sage Publications Limited.

Fisher, R. A. (1922). On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London, A222*, 309-368.

Fisicaro, S. A., & Lance, C. E. (1990). Implications of three causal models for the measurement of halo error. *Applied Psychological Measurement, 14*, 419-429.

Fleiss, J. L. (1981). *Statistical methods for rates and proportions*. New York: John Wiley.

Foster, P. (1996). Doing the task better: How planning time influences students' performance. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 126–135). London: Heinemann.

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition, 18*(3), 299-324.

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

Fox, J., & Cheng, L. (2007). Did we take the same test? Differing accounts of the Ontario Secondary School Literacy Test by first and second language test-takers. *Assessment in Education: Principles, Policy and Practice, 14*(1), 9–26.

Freed, B. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 243-265). Ann Arbor: The University of Michigan Press.

Friedman, A. (2012). How to collect and analyze qualitative data. In A. Mackey & S. M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 180-200). Chichester: Wiley-Blackwell.

Ghisletta, P., & Spini, D. (2004). An introduction to generalized estimating equations and an application to assess selectivity effects in a longitudinal study on very old individuals. *Journal of Educational and Behavioral Statistics, 29*, 421-437.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and (+/_ Here-and-Now): Effects on L2 oral production. In M. Garcia-Mayo (Ed.), *Investigating Tasks in Formal Language Learning*. (pp. 44–68). Clevedon: Multilingual Matters.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press.

Grainger, P., Purnell, K., & Kipf, R. (2008). Judging quality through substantive conversations between markers. *Assessment and Evaluation in Higher Education, 33*, 133–142.

Griffin, Z. M., & Bock, K. (2000). What the eyes say about speaking. *Psychological Science, 11*(4), 274–279.

Harvill, L. M. (1991). NCME instructional module: Standard error of measurement. *Educational Measurement: Issues and Practice, 10*(2), 33–41.

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

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.

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

Hunt, K. W. (1965). *Differences in Grammatical Structures Written at Three Grade Levels*. National Council of Teachers of English, Urbana, IL. Research Report No. 3.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of information processing approach to task design. *Language Learning, 51*, 401-436.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning, 64*(4), 809-854.

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

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

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

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

Kormos, J. (2011). Speech production and the Cognition Hypothesis. In P. Robinson (Ed.), *Second Language Task Complexity: Researching the Cognition Hypothesis of Language Learning and Performance* (pp. 39-60). Amsterdam: John Benjamins.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*, 48-60.

Kyle, K., Crossley, S. A., & McNamara, D. S. (2016). Construct validity in TOEFL iBT speaking tasks: Insights from natural language processing. *Language Testing*, *33*, 319-340.

Lance, C. E., LaPointe, J. A., & Stewart, A. M. (1994). A test of the context dependency of three causal models of halo rater error. *Journal of Applied Psychology, 79*, 332-340.

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75*, 440-448.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written

production. *Applied Linguistics, 16*, 307-322.

Lazaraton, A. (2002). *A qualitative approach to the validation of oral language tests.* Cambridge: Cambridge University Press.

Lee, S., & Winke, P. (2018). Young language learners' response processes when taking computerized tasks for speaking assessment. *Language Testing, 35*(2), 239-269.

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

Levelt, W. (1989). *Speaking: From intention to articulation.* Cambridge, MA: The MIT Press.

Linacre, J. M. (1989). Many-faceted Rasch measurement. Chicago, IL: MESA Press.

Linacre, J. M. (2000). Item discrimination and infit mean-squares. *Rasch Measurement: Transactions of the Rasch Measurement SIG, 14*(2), 743.

Linacre, J. M. (2017). *FACETS Rasch-model computer program* (Version 3.80.3) [Computer software]. Chicago, IL: Winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 484–509.

Long, M. (2015). Second language acquisition and task-based language teaching. Maiden: Wiley-Blackwell.

Lumely, T. (2002). Assessment criteria in a large-scale writing test: What do they really mean to the raters? *Language Testing, 19*, 246-276.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*, 85-104.

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods, 1*, 30–46.

McNamara, T. (1995). Modelling performance: opening Pandora's box. *Applied Linguistics, 16*, 159-179.

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

Mislevy, R. J., Steinberg, L. S. & Almond, R. G. (2003). On the structure of educational assessments. *Measurement: Interdisciplinary Research and Perspectives, 1*, 3–67.

Mochizuki, N., & Ortega, L. (2008). Balancing communication and grammar in beginning level foreign language classrooms: A study of guided planning and relativization, *Language Teaching Research, 12*, 11–37.

Moss, P. A., Girard, B. J., & Haniford, L. C. (2006). Validity in educational assessment. *Review of Research in Education, 30*, 109–162.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement, 5*(2), 189-227.

Nitta, R., & Nakatsuhara, F. (2014). A multifaceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing, 31*(2), 147-175.

Norris, J. M. (2009). Task-based teaching and testing. In M. J. Long & C. J. Doughty (Eds.), *The handbook of language teaching* (pp. 578-594). Chichester, UK: Wiley-Blackwell.

Norris, J. M. (2016). Current uses for task-based language assessment. *Annual Review of Applied Linguistics, 36*, 230-244.

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

Ockey, G. J., Koyama, D., & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test. *Language Assessment Quarterly, 10*(1), 1–17.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*, 109-148.

Ortega, L. (2000). *Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners*. Unpublished Doctoral dissertation, University of Hawai'i at Manoa.

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77-109). Amsterdam: John Benjamins.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October 3-6). *An investigation of elicited imitation tasks in crosslinguistic SLA research*. Paper presented at the Second Language Research Forum, Toronto.

Papageorgiou, S., Stevens, R., & Goodwin, S. (2012). The relative difficulty of dialogic and monologic input in a second-language listening comprehension test. *Language Assessment Quarterly, 9*(4), 375-397.

Park, H. I. (2015). *Language and cognition in monolinguals and bilinguals: A study of spontaneous and caused motion events in Korean and English*. Unpublished doctoral dissertation. Washington DC: Georgetown University, Department of Linguistics.

Plakans, L. (2010). Independent versus integrated writing tasks: A comparison of task representation. *TESOL Quarterly, 44*(1), 185–194.

Plakans, L., & Gebril, A. (2016). Exploring the relationship of organization and connection with scores in integrated writing assessment, *Assessing Writing, 31*, 98-112.

Polio, C. (1997). Measures of linguistic accuracy in second language writing research. *Language Learning, 47*, 101-143.

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

Richards, J., Platt, J., & Platt, H. (1996). *Dictionary of language teaching and Applied Linguistics.* London: Longman.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes, 14*, 423-441.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*, 99-140.

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

Robinson, P. (2003). Attention and memory. In C. J. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp.631-678). Oxford: Blackwell.

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

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* (pp. 3-37). Amsterdam: John Benjamins.

Rutherford, K. (2001). *An investigation into the effects of planning on oral production in a second language*. Unpublished master's thesis, University of Auckland, Auckland, New Zealand.

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

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.

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

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy,

fluency, and lexis. *Applied Linguistics*, *30*(4), 510-532.

Skehan, P. (2016). Tasks versus conditions: Two perspectives on task research and their implications for pedagogy. *Annual Review of Applied Linguistics, 36*, 34-49.

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*, 185-211.

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

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

Spolsky, B. (1985). The limits of authenticity in language testing. *Language Testing, 2*(1), 31–40.

Strauss, A. L., & Corbin, J. (1998). *Basics of qualitative research: Technique and procedures for producing grounded theory* (2nd ed.). London, England: SAGE.

Stiger, T. R., Kosinski, A. S., Barnhart, H. X., & Kleinbaum, D. G. (1998). ANOVA for repeated ordinal data with small sample size: A comparison of ANOVA, MANOVA, WLS and GEE methods by simulation. *Commun Stat B Simul Comput, 27*, 357–375.

Swain, M. (1984). Large-scale communicative testing: a case study. In Savignon, S. J. and Berns, M. (Eds.), *Initiatives in communicative language teaching* (pp. 185-201). Reading, MA: Addison-Wesley.

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *Canadian modern language review*, *50*(1), 158-164.

Tajima, M. (2003). *The Effects of Planning on Oral Performance of Japanese as a Foreign Language*. Unpublished doctoral dissertation. Purdue University, West Lafayette.

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.

Tracy-Ventura, N., McManus, K., Norris, J., & Ortega, L. (2013). "Repeat as much as you can": Elicited imitation as a measure of oral proficiency in L2 French. In P. Leclercq, H. Hilton, & A. Edmonds (Eds.), *Proficiency as- sessment issues in SLA research: Measures and practices*. Clevedon, UK: Multilingual Matters.

Trofimovich, P., & Baker, W. (2006). Learning second language suprasegmentals: Effect of L2 experience on prosody and fluency characteristics of L2 speech. *Studies in Second Language Acquisition, 28*, 1-30.

Van Gorp, K., & Deygers, B. (2013). Task-based language assessment. In A. Kunan (Ed.), *The companion to language assessment: Vol. 2. Approaches and development* (pp. 578 -593). Oxford, UK: Wiley-Blackwell.

Van Moere, A. (2012). A psycholinguistic approach to oral language assessment. *Language Testing, 29*(3), 325-344.

VanPatten, B. (1990). Attending to form and content in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12*, 287-301.

Weigle, S. C. (1998). Using FACETS to model rater training effects. *Language Testing, 15*(2), 263-287.

Weigle, S. (2004). Integrating reading and writing in a competency test for non-native speakers of English. *Assessing writing, 9*, 28–47.

Wendel, J. (1997). *Planning and second language narrative production*. Unpublished doctoral dissertation, Temple University, Japan.

West, D. E. (2012). Elicited imitation as a measure of morphemic accuracy: Evidence from L2 Spanish. *Language and Cognition, 4*(3), 203–222.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 21-44.

Wigglesworth, G. (2001). Influences on performance in task based oral assessments. In M. Bygate, P. Skehan, and M. Swain (Eds.), *Researching pedagogic tasks, second language learning, teaching and testing* (pp. 186-209). London: Longman.

Wigglesworth, G., & Elder, C. (2010). An investigation of the effectiveness and validity of planning time in speaking test tasks. *Language Assessment Quarterly, 7*(1), 1-24.

Wigglesworth, G., & Frost, K. (2017). Task and performance-based assessment. In E. Shohamy, I. Or, & S. May (Eds.), *Language testing and assessment* (pp. 121-133). Springer.

Winke, P., Gass, S., & Myford, C. (2012). Raters' L2 background as a potential source of bias in rating oral performance. *Language Testing, 30*(2), 231–252.

Wolfe, E. W. (1997). The relationship between essay reading style and scoring proficiency in a psychometric scoring system. *Assessing Writing, 4*(1), 83-106.

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22*, 463–508.

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