COMPLEXITY, ACCURACY, AND FLUENCY MEASURES IN ORAL PRE-TASK PLANNING: A SYNTHESIS

MITSUKO SUZUKI
University of Hawai'i at Mānoa

INTRODUCTION

Over the past few decades, researchers have investigated planning as a significant process in task-based language teaching (TBLT). Past studies have supported the effectiveness of planning in second language (L2) learners' oral production, especially in terms of fluency (e.g., Foster & Skehan, 1996; Gilabert, 2007; Ortega, 1999; Sasayama & Izumi, 2012; Yuan & Ellis, 2003). However, considering the influence of planning on accuracy (e.g., Foster & Skehan, 1999; Lee & Oh, 2007; Mehnert, 1998; Mochizuki & Ortega, 2008) and complexity (e.g., Bei, 2010; Kawauchi, 2005; Nitta, 2007; Wang & Song, 2015; Wigglesworth, 1997; Yuan, 2001), research has yielded mixed results. One of the reasons for this inconsistency in results may be the different units that studies have used to measure complexity, accuracy, and fluency (CAF). This variety makes comparisons among pre-task planning studies difficult (Ellis, 2009b). Although researchers in CAF have commented on this issue at large (e.g., Lambert & Kormos, 2014: Plonsky & Kim, 2016), they have not yet focused directly on pre-task planning. Therefore, the aim of this paper is to spark the discussions around the use of CAF measures by synthesizing existing pre-task planning studies and comparing the CAF measures employed in a set of selected studies.

A number of quantitative studies conducted between 1995 and 2016 were selected based on a set of inclusion criteria. In order to investigate the overall role of strategic planning in oral tasks, special focus was given to CAF measures and the operationalization of pre-task and main task, including (a) the instruction given prior to the planning, (b) type of pre-task planning activity, (c) length of planning time, and (d) type of main task. More than 200 studies were

Second Language Studies, 36(1), Fall 2017, pp. 1-52.

collected in the initial phase, of which 40 were selected for comparison. The overview of this research process and the findings will be presented after a brief review of existing pre-task planning studies. Finally, the paper will conclude with a discussion of how researchers can use CAF measures to develop a deeper understanding of pre-task planning.

LITERATURE REVIEW

As an alternative approach to focus on forms and focus on meaning (Ellis, 2005a; Skehan, 2003), TBLT has been widely practiced in various ESL and EFL settings. Although its definition varies among researchers, a task might simply be described as an "activity which requires learners to use language, with emphasis on meaning, to attain an objective" (Bygate, Skehan, & Swain, 2001, p. 11). Although researchers have identified task characteristics (e.g., number of elements), learner factors (e.g., differences in working memories), and modes (e.g., monologic or dialogic task) as variables that are relevant to L2 performance, the variables that help learners perform a task remain to be discovered. Among these variables, however, the role of pre-task planning has attracted significant research interest. Although there are a number of previous studies that have examined the influence of pre-task planning on L2 writing (e.g., Ellis & Yuan, 2004; Genc, 2012; Ong & Zhang, 2010), the present study will focus on the effectiveness of pre-task planning in oral performance, specifically, on CAF. While inclusion of writing tasks may bring a more profound understanding of pre-task planning effect, synthesizing both modes of language seemed problematic, as researchers have different opinions about the theoretical relationship between spoken and written language (Cleland & Pickering, 2006). A brief explanation of pre-task planning and its potential benefits in language performance as well as its methodological issues is presented below.

Pre-Task Planning and L2 Task Performance

Ellis (2005a) described planning as a "problem solving activity" (p. 3), since participants need to decide "what linguistic devices need to be selected in order to affect the audience in the desired way" (p. 3). In his review of task planning, Ellis (2005a) categorized task-based planning into pre-task planning and within-task planning. The latter is also known as on-line planning (Yuan & Ellis, 2003). The difference between these two types of planning is the time when it takes place: pre-task planning occurs before the task, while within-task occurs during the task. Pre-task planning can be further divided into rehearsal and strategic planning (Ellis, 2005a). In rehearsal, learners are given time to practice and complete their performance prior to the main task, and the main task consequently becomes a repetition of what the participants did in the planning phase. The focus of this study, however, is strategic planning, which provides learners prior time to consider the content and language to use but not to completely rehearse the task. These pre-task planning types are distinct from other pre-task activities (e.g., brainstorming, pre-teaching new phrase) in the sense that the actual task materials are ready in the planning stage (Ellis, 2005a).

Ellis (2005a) showed several other ways as well of subdividing pre-task planning. For instance, students can prepare for the main task individually or collaboratively with their peers. In addition, teachers can choose to conduct guided planning, where they give students explicit instructions on what and how to prepare, or unguided planning, which leaves students without any specific instructions. In the case of guided planning, teachers can bring attention to a particular linguistic form, meaning, or both. Note, however, that these past TBLT studies are not necessarily comparable, since their instructions vary from study to study (Ellis, 2005a). Furthermore, students' activities during this unguided planning time are unclear in most of these studies. In other words, the significance of what learners actually do in this given preparation time is not always apparent. Therefore, how each of these pre-task planning types or a combination of these pre-task planning types affects L2 learners' oral performance remains

under investigation.

Empirical Studies on Pre-task Planning Effect

Some researchers have conducted empirical studies on planning effects (e.g., Ellis, 2005b; Ellis & Yuan, 2004; Foster & Skehan, 1996; Gilabert, 2007; Sasayama & Izumi, 2012; Yuan & Ellis, 2003). One of the most typical measures used to gauge the effectiveness of planning has been CAF (Ellis, 2009a). In addition to these CAF measures, lexis has also been proposed as an importance index of learners' language proficiency (Housen, Kuiken, & Vedder, 2012; Skehan, 1998). Past studies have supported the effectiveness of planning on L2 oral production, especially in terms of fluency (e.g., Foster & Skehan, 1996; Gilabert, 2007; Ortega, 1999; Sasayama & Izumi, 2012; Yuan & Ellis, 2003).

In regard to the planning effect on lexical complexity and accuracy, however, research has yielded mixed results. Ellis (2009b) stated that pre-task planning might have a positive influence on syntactic complexity, but not as much on lexical complexity. Yuan and Ellis's (2003) study, in which pre-planning groups' lexical complexity did not statistically differ from that of the non-planning group, seems to support this argument. In contrast, participants in Kawauchi's (2005) study showed significant improvement in lexical density under the pre-task condition, regardless of their proficiency levels. Yuan and Ellis (2003) also examined the effectiveness of planning in accuracy and found that, although the planning group showed higher accuracy than the non-planning group, no significant difference was identified between the two groups (Yuan & Ellis, 2003). While their findings are in line with the findings of Crookes (1989) and other researchers, other studies have generated rather inconsistent results (e.g., Mehnert, 1998; Mochizuki & Ortega, 2008). These findings indicate that the impact of pre-task planning on complexity and accuracy is not as apparent as its impact on fluency.

One limitation in comparative analysis across planning studies is associated with the operationalization of tasks and how task variables interact with planning. Although the narrative

task has been widely adopted in pre-task planning studies, some studies have used other tasks (e.g., decision-making, instruction-giving) that require different participation styles (e.g., one-way or two-way task). Task characteristics, such as the number of choices in the decision-making task, were not identical across these studies either. Interpretation becomes even more complicated when differences in participant characteristics (e.g., L1 difference, L2 proficiency, familiarity with the partner) and task settings (i.e., lab, classroom, test) are considered. As some researchers (e.g., Levkina & Gilabert, 2012) have reported, careful consideration is necessary when comparing the results of the synergetic effects of pre-task planning and other task complexity factors.

It is also important to note the different operationalization of planning in terms of instruction and planning time. While some studies conducted unguided pre-task planning (e.g., Levkina & Gilabert, 2012), others supplied guidelines prior to the planning time. For instance, participants in Mochizuki and Ortega's (2008) study were instructed to focus on a particular grammatical structure, whereas Skehan and Foster (2005) advised learners to consider useful grammar structure without specifying any form. Some studies (e.g., Kawauchi, 2005) provided different modes of planning (e.g., note-taking, talk-aloud, and reading a model answer). As what participants are doing during this planning time has rarely been focused on (Ellis, 2009b), to what extent these guidelines were helpful for students remains unclear. Regarding preparation time, some studies provided five minutes (e.g., Mochizuki & Ortega, 2008), while others provided less (e.g., Iwashita, McNamara, & Elder, 2001) or more (e.g., Foster & Skehan, 1996; Gilabert, 2007; Ortega, 1999; Tavares, 2009; Yuan & Ellis, 2003). The results of past studies (e.g., Mehnert, 1998; Li, Chen, & Sun, 2015) indicate that the length of planning time creates varying degrees of complexity. In sum, it is unclear whether it is the pre-task planning activity, the planning time, or both that have an impact on task performance.

Methodological Issues in TBLT

Another reason for inconsistency in results may be the different CAF measures that researchers have used (Ellis, 2009b). Regarding accuracy, some studies have employed global measures (e.g., error-free clauses) to yield a more realistic picture of learners' performance (e.g., Foster & Skehan, 1996; Levkina & Gilabert, 2012), while others have focused instead on the use of a specific linguistic component (e.g., Crookes, 1989; Mochizuki & Ortega, 2008; Ortega, 1999) or have even used both measures (e.g., Gilabert, 2007). In a sense, implementation of both measures offers a more comprehensive picture of accuracy, since a focus solely on global measures will "have the disadvantage of being too broad to capture small changes…and [will] obscure errors in grammatical domains that may be important at a given level of development" (Ortega, 1999, p. 118). However, this variety makes comparison among studies difficult. As illustrated in Table 1, the measures used for fluency and complexity are not necessarily consistent among studies, not even those that use a similar type of task.

Table 1

Methodological Comparison of Planning Studies on Narrative Tasks

Study	Foster & Skehan (1996)	Ortega (1999)	Yuan & Ellis (2003)	Gilabert (2007)
Planning	10-minute pre-task planning	10-minute strategic Planning	10-minute pre-task planning; on-line planning	10-minute pre-task planning
Task type	personal task; narrative task; decision-making task	narrative task	narrative task	narrative task (four different level of complexity)
Lexical Complexity	N/A	TTR	MSTTR	Guiraud's index
Syntactic Complexity	clauses per C-unit; variety of verb forms	words per utterance	sentence-node per T-unit; variety of verb forms	sentence-node per T-unit
Accuracy	error-free clauses	native-like use of Spanish articles; noun-modifier agreement	error-free clauses	error-free T-unit; target-like use of English articles; self-repairs
Fluency	number of replacement; false start; repetition	syllables per second in pruned speech rate	syllables per minute	pruned and unpruned speech rate

Note. TTR = type token ratio; MSTTR = mean segmental type-token ration

This lack of operational consistency is not an issue unique to pre-task planning, and addressing this methodological inconsistency is a serious challenge in the field of TBLT (Lambert & Kormos, 2014: Plonsky & Kim, 2016). Although TBLT has benefitted from the implementation of the CAF concept, its variability in operationalization prevents researchers from drawing firm conclusions (Plonsky & Kim, 2016). Researchers have also claimed that the variability in measures is problematic to internal validity, especially because TBLT studies typically have small sample sizes and multiple variables (Plonsky & Kim 2016). While Plonsky and Kim (2016) suggest that CAF measures should be "theoretically motivated and used in a way that allows for comparability across studies" (p. 90), to the author's best knowledge, there has not been a synthesis that explores the measures used in pre-task planning studies. It is not clear, then, the extent to which past pre-task planning studies are comparable.

Research Questions

The aim of this study is to review and synthesize the methodological approaches that have been adopted in pre-task planning studies. To be specific, there are two objectives behind the present study. The first objective is to provide the field of applied linguistics with an overview of the research designs and procedures employed in these studies. Considering the necessity of identifying the potential task factors that can enhance or reduce planning effects, it seems important to capture an overall picture of the research designs used in the primary studies. The second objective is to report how researchers have operationalized CAF in this domain. To the best of the author's knowledge, past syntheses of pre-task planning effects (i.e., Ellis, 2009b; Skehan & Foster, 2012) have not fully examined the CAF measures used in these studies (for recent reviews of CAF in general, see Housen & Kuiken, 2009; Housen, Vedder, & Kuiken, 2012). The first objective, which corresponds to Research Question 1, focuses on the data collection, while Research Question 2, corresponding to the second aim, is more focused on the analytical tools used in the literature. The research questions are as follows:

1. To date, how has primary pre-task planning research operationalized pre-task planning in terms of the following: (a) the instruction given prior to the planning, (b) type of pre-task planning activity, (c) length of planning time, and (d) type of main task?
2. What CAF measures have been implemented in pre-task planning studies? To what extent has this variability in measurement influenced the CAF results?

METHOD

Data Collection: Identifying Primary Source

Following In'nami and Koizumi's (2010) suggestion, the current research was started by searching for articles in the Education Resources Information Center (ERIC) and Linguistics and Language Behavior Abstracts (LLBA) databases, which are the two most frequently used electronic databases in the field of applied linguistics. According to their study, these sources provide the most extensive coverage of journals in the field (In'nami & Koizumi, 2010). The following keywords were combined or truncated as search terms: pre-task planning, task, task-based language learning, fluency, accuracy, complexity, cognitive process. The journals in these databases (e.g., *Annual Review of Applied Linguistics*, *Applied Language Learning*, *TESOL Quarterly*) were also manually searched in order to ensure the coverage of the relevant articles. Then, using the reference sections of these articles, the relevant journals and articles that were not covered in this electronic database search were manually searched and identified. In addition, state-of-the-art articles, dissertations, edited books (e.g., Ellis, 2005b) and book chapters (e.g., Levkina & Gilabert, 2012; Sasayama & Izumi, 2012), as well as their reference sections were reviewed for further data collection. These included articles written in English, Spanish, and Japanese (i.e., the languages spoken by the author of this study).

Inclusion and Exclusion Criteria

Although research synthesis is intended to mitigate the variance in study quality, explicit criteria are necessary to achieve a logical analysis. As Field and Gillett (2010) pointed out, "[O]ne red sock (bad study) amongst the white clothes (good studies) can ruin the laundry" (p. 668). In addition, unlike general quantitative studies where random sampling is preferred, synthetic studies usually aim to collect primary sources that are comparable (Norris & Ortega, 2006). Thus, as Norris and Ortega (2006) noted, "[R]esearch synthesis always includes an explicit articulation of how the relevant literature was searched and how primary studies were selected for review" (p. 6).

Over 200 quantitative studies relevant to pre-task planning and CAF were identified. These studies were examined to determine whether they matched the present study's research questions. Through the process of examining the collected reports and referring to past TBLT-related synthesis (e.g., Jackson & Suethanapornkul, 2013), the following emerged as inclusion criteria:

1. The study was published between 1995 and October 2016. The mid-nineties were when CAF were proposed as the three principal dimensions of L2 production and proficiency (Housen, Kuiken, & Vedder, 2012; Skehan, 1996). Thus, this time frame seems to cover the relevant research on pre-task planning with reference to the CAF model.
2. The study was experimental or quasi-experimental, designed to explore the effect of pre-task planning on L2 learners' task performance (i.e., CAF).
3. The study quantitatively measured oral CAF as dependent (i.e., outcome) variables through the comparison of identical participants' performance across different conditions (i.e., with or without pre-task planning) or through group comparison (i.e., a control group without pre-task planning and experimental group with pre-task planning).
4. The participants in the study were L2 or foreign language learners who were 18 years of

age or older. Since the age effect on the cognitive process and task performance is unclear, the present study removed the age variable by limiting the participants to adult learners of additional language(s).

5. The task(s) in the study was or were monologic or dyadic oral communication task(s). Here, tasks are defined, as by Ellis (2003, 2009a), as comprising language practices with four underlying principles: focus on meaning, inclusion of gaps, reliance on one's resource, and use of language as means. Activities that were less meaning oriented and that were equivalent to simple translation (e.g., Zhigang & Xudong, 2008) were thus excluded from the present study. In addition, the current study did not include computer-mediated studies, since this factor could influence the performance compared to face-to-face interaction (Lin, 2014).
6. The study described the characteristics of pre-task planning (e.g., guided or unguided, length of planning time, type of planning) and main task(s) (e.g., task complexity) in enough detail for them to be coded.

At the same time, reports were excluded from the analysis due to the following reasons:

1. The paper was a review of past studies, not including new empirical data (e.g., Skehan, 2016).
2. The article analyzed data from reports that were already included in the synthesis (e.g., Foster, 2011, compared data from Foster, 2000, and Foster & Skehan, 1996, which have been included in the current synthesis).
3. Specific CAF measures were not targeted in the analysis. Studies based on CAF performance scores marked by raters (e.g., Castro, Cabrera, & Martínez, 2009; Xi, 2005) were not included in the analysis.
4. The study used descriptive research design (e.g., Foster, 1996).
5. Additional treatment or instructions were given during the main tasks. When the participants had received a specific instruction on performance, its influence on the

output was considered. For instance, Birjandi and Seifoori (2009) implemented a 15-week-long training program on metacognitive strategies prior to the pre-task planning. Whether these pressured outputs were comparable with outputs that did not have such directions was uncertain, and therefore these results were not included in the analysis.

Coding and Analyzing Data

A coding sheet was developed to profile the details of the primary studies. Table 2 summarizes the items in the sheet. The coding of substantive features included independent variables, while methodological features organized the statistical information in the primary studies. As shown in the table, high-inference variables (e.g., task type, guidance) had several levels within each item.

Table 2
Substantive and Methodological Features Coded

Methodological features	Substantive features
Research Design	Planning
With-in	Guidance
Between	Guided
Sample Size	Unguided
	Teacher-led
Learner Characteristics	
L1	During-planning activity
Proficiency level	Think-silence
Age (mean and range)	Think-aloud
Target language	Note-taking
Institution (high school, university)	
Learning context (ESL, EFL, LOTE)	Planning Time
Research Setting	Main task
Classroom	Task type
Lab	Narrative
Test	Instructions
	Decision-making
CAF Measures	Opinion-giving
	Personal Information
Descriptive Statistics	Others
	Interaction
Results	Monologic
	Interactive
	Task Time

Pre-task planning. In coding the pre-task planning, the extent to which the participants were given time, guidance, and tools, as well as the type of planning, were coded. Regarding the planning time, preparation times for not only planned but also unplanned settings were recorded. For the higher-inference variables, a brief description is provided in Table 3. When the participants were not given any instructions on how to use the planning time, the study was coded as unguided. In contrast, studies were coded as guided planning when the participants were "given specific advice about what and how to plan" and "directed to attend to linguistic form, to meaning or to form and meaning" (Ellis, 2005a, p. 5). The guided planning was further divided into undetailed and detailed. In the undetailed studies, participants were briefly instructed to think about what to say and how to deliver the speech. Detailed studies, on the other hand, had more guidelines to organize ideas. For instance, Skehan and Foster (2005) had a list of advice, such as "think what you already know about each of the questions for judgement," "think what grammar you need to do the task," and "think how to avoid difficulties and solve problems with grammar and vocabulary" (p. 216).

In order to make comparison possible, guided planning targeting a specific grammar point (i.e., detailed guidance) was excluded from the analysis. For instance, Sangarun (2005) and Mochizuki and Ortega (2008) had instructions on particular grammar points (i.e., relative clause and comparatives, respectively) in their planning guidelines. Their degree of guidance differed from those of other studies (e.g., Skehan & Foster, 2005) which advised learners to think about the language forms they could use in general.

Table 3
Codes related to Pre-task Planning

Feature	Code	Definition
Guidance	Unguided	Participants were simply told to plan the study
	Guided	Through oral or written instruction, participants' attention was directed to specific aspect of planning (i.e., language, structure, and/or content)
	Teacher-led	The teacher led the planning session
During-planning	Think-silence	Plan what and how to say silently with no tools
	Note-taking	Pencil and paper were provided to take notes

Main task. The tasks were categorized into six types, as summarized in Table 4. Narrative included one-way story-telling tasks in which participant(s) had to retell a story using a sequence of pictures (e.g., Sasayama & Izumi, 2012; Yuan & Ellis, 1993) or a video clip (e.g., Wang, 2014). Tasks coded as instructions were similar to what Pica, Kanagy, and Falodun (1993) labeled as information gap: the information flowed one way (i.e., one dominant interactant holds and offers the information), and all the participants were required to interact toward a singular goal. In contrast, decision-making tasks consisted of two-way information exchange that allowed participants to make a choice based on a shared set of options. The tasks categorized under decision-making were slightly different from those under opinion, in that the latter did not require the participants to conclude with a single decision. Tasks coded as opinion simply asked the participants to share their own view or preference on a certain topic. The fifth task type category, personal information tasks, was based on facts about the participant (i.e., less reasoning demand), unlike opinion tasks, which involve some reasoning demand.

Table 4

Task Type Codes and Description

Task	Description	Information flow
Narratives	Story-telling using sequence of pictures.	1 way
Instructions	Give instructions to one's (imagined) partner. Participants do not have shared access to the same information.	1 way
Decision-making	Make a choice from a range of provided items. All the participants have access to the same information.	1 way or 2 way
Opinion-giving	Deliver a view or preference on a certain topic, but a single solution is not required.	1 way or 2 way
Personal Info	Describe personal experience.	1 way
Others	Tasks that did not fit into any of the above categories.	

Methodological features. Features related to methodology included research design, learner characteristics, research setting, and statistical analyses. In order to record the research designs, the number of participants and their experimental condition (i.e., within-subject or between subject) were coded.

Participants' proficiency had sub-categories of impressionistic judgment, institutional status, in-house assessment, and standardized test, based on Thomas's (1994) classification. Impressionistic judgment refers to studies that evaluated participants' proficiency level through the researcher or teacher's subjective assessment, while institutional status used the course that learners were enrolled in as the baseline (e.g., studies that considered learners' level as beginner if they were enrolled in an elementary level class). If the judgment was based on a locally developed test, the study was classified as in-house assessment. Finally, studies were coded as standardized tests when the researcher reported the participants' score on an established proficiency test (e.g., IELTS, TOEFL).

Other features related to learner characteristics and research settings were coded as well: age, institution (e.g., high school or university), first language and target language, learning context (i.e., ESL, EFL, LOTE), and research context (i.e., lab, classroom, test). As for statistical analyses, choice of CAF measures (e.g., number of pauses, error-free clauses, clauses per AS-unit), procedure (e.g., t-test, ANOVA), and reports on the following were coded, if available: *t*-value, *F*-value, *p*-value, standard error (*SE*), standard error of measurement (*SEM*), effect size.

Coding of results. The results of the studies were coded as effective, not effective, and mixed. In some of the studies, performance under the planning condition was significantly enhanced in terms of CAF. These studies were coded as effective, whereas those that did not find such a statistical difference were categorized as not effective. Not all studies, however, had such a clear result. Some researchers reported mixed findings, where the effectiveness of pre-task planning, measured in statistical testing, depended on the type of pre-task planning and main task, as well as the amount of time allocated for planning. These results were categorized as mixed.

RESULTS

Overview of the Literature Search

In the initial literature search, over 200 studies were identified as being relevant to strategic planning. Through careful selection and evaluation, 40 empirical studies were then chosen to be included in the synthesis. These studies included 17 peer-reviewed journal articles, nine book chapters, 13 MA and PhD dissertations, and one conference proceeding. An overview of these studies is presented in Tables 5-7 (see Appendix A for the list of all 40 studies).

Table 5
Research Design and Context (*N* = 40)

Design and Context		*N*	%
Design of Study			
	with-in	12	30.0%
	Between	27	67.5%
	both	1	2.5%
Context of the study			
	Laboratory	27	67.5%
	Class	7	17.5%
	Test	5	12.5%
	Not reported	1	2.5%
Learning Context			
	ESL	11	27.5%
	EFL	21	52.5%
	LOTE	6	1.5%
	ESL and EFL	1	2.5%
	Not reported	1	2.5%

Note. ESL = English as Second Language; EFL = English as Foreign Language; LOTE = Language Other Than English

Table 5 indicates that 27 out of 40 strategic planning studies examined the effectiveness of planning through between-subject effect and often in laboratory settings (*n* = 27). Table 5 also shows that most of the studies targeted English language learners (*n* = 33), mostly in EFL

contexts (n = 21). Other studies targeted languages other than English (i.e., LOTE) including French (Gaillard, 2013), German (Mehnert, 1998), Spanish (e.g., Gutiérrez, 2013; Ortega, 1999), and Japanese (e.g., Nakakubo, 2011). The average number of participants was 45.9, ranging from 6 (Spetch, 2014) to 143 learners (Nakakubo, 2011). The target participants were undergraduate or adult learners who were 18 years of age or older.

For the studies to be comparable, their planning conditions must be operationalized. All the studies investigated reported on the amount of time allocated for planning. Among the 40 studies, three assessed the influence of different planning time on CAF (i.e., Mehnert, 1998; Wang & Song, 2015; Wigglesworth & Elder, 2010), and, therefore, there were two or more planning time conditions. Other studies compared the performance of the experimental (i.e., planning group) and control groups (i.e., non-planning group) after a certain planning time. As shown in Table 6, the planning time ranged from 1-10 minutes, but more than half of the studies allocated 10 minutes (n = 28) for planning.

Table 6 also shows the types of guidance that these studies provided to the participants. Note that some of the researchers used different instructions in a single study, since they aimed to compare the effects of guided, unguided, and teacher-led planning on performance. While 32 studies offered planning time without any guidance, eight studies included guided planning, including linguistic, content, and structural instructions. In regard to linguistic guidance, the degree of specificity ranged from those which simply asked the participants to come up with useful grammar and vocabulary by themselves to those which directed them to particular grammar points (see Appendix B for specific examples). For content and structural instruction, worksheets helped students brainstorm and organize their thoughts. Four studies did not report on the kind of instructions or guidance that were offered to the participants. Table 6 also reveals that most of the studies encouraged learners to take notes during the planning (n = 30). The exceptions were studies conducted by Foster and Skehan (1999) and Gaillard (2013), which involved not only individuals' self-planning but also teacher-led guidance (i.e., prior presentation by teachers on planning). Foster and Skehan (1999) also included student-led planning, in which students brainstormed on how to perform together in small groups. Two other exceptions, categorized in *Others*, included a study conducted by Kawauchi (2005),

which required students to plan their task in silence after reading a model answer, and a study conducted by Guará-Tavares (2008), in which the researcher asked questions to the students during the planning session in order to elicit their thoughts at the site. Note here, that not all participants had access to these notes or other prior works during the target task. In many cases (e.g., Gilabert, 2007; Nitta, 2007; Ortega, 1995; Saeedi, 2013; Tajima, 2003), participants were not allowed to use their notes during their performance.

Table 6

Operationalization of Planning

Planning Condition		*n*
Time (minutes)		
	1	5
	2	2
	3	4
	5	5
	8	1
	10	28
Guidance		
	Unguided	32
	Guided	8
	Teacher-led	2
	Not reported	4
During pre-task planning		
	Think-silence	1
	Note-taking	30
	Others (i.e., Student-led; Verbal protocol)	2
	Not reported	10

The analysis of the pre-task planning studies also revealed the types of task frequently implemented in the literature (see Table 7). Twenty-five studies implemented one particular activity, and other studies involved more than one task in their experiments. As pointed out in previous studies (e.g., Skehan, 2001), most of the main tasks conducted in the collected studies were narratives (i.e., story-telling). The majority of these narrative tasks were picture based,

with the exception of three studies that used wordless films (e.g., Wang, 2014). Although most of these were conducted as monologic tasks that required each individual to describe the story to the researcher, some of the tasks allowed negotiation of meaning to some extent. For instance, participants in Lee and Oh's (2007) study were assigned either a speaker or listener role to add more authenticity and meaningfulness to the task. The second most implemented task was decision-making (n = 8), followed by opinion-giving (n = 6), instruction (n = 4), and personal information (n = 2). As mentioned earlier in this paper, these tasks did not necessarily require exchange of information among participants. Finally, some of the tasks categorized under *Other* included picture comparison tasks (e.g., Rezaei & Tabatabaei, 2015), listening to and summarizing a telephone message (e.g., Wigglesworth, 1997), a pragmatic task (Mehnert, 1998), and topic talk (e.g., Bei, 2010).

Table 7

Task Types

Task	Description	*n*
Narratives	Story-telling using sequence of pictures.	25
Instructions	Give instructions to a (imagined) partner. Participants do not have shared access to the same information.	4
Decision-making	Make a choice from a range of provided items. All the participants have access to the same information.	8
Opinion-giving	Deliver a view or preference on a certain topic, but a single solution is not required.	6
Personal Info	Describe personal experience.	2
Others	Tasks that did not fit into any of the above categories.	9

Measures for Lexical Complexity

Of the 40 studies, less than half investigated the effect of planning on lexical complexity. Table 8 displays the different measures used in the studies and the results using these measures. Note that some studies used more than one measurement to assess the effectiveness of pre-task planning. The first column under Results indicates the number of studies that found pre-task planning effective for lexical complexity, whereas the second column (i.e., not effective) shows those that did not find such an effect. It appears that a slightly higher number of studies found pre-task planning to not be necessarily effective for lexical complexity. The numbers under the

Mixed column represent the number of studies that revealed a rather complex picture. For instance, in Mehnert's (1998) study, 10-minute pre-task planning was effective in increasing lexical density when the task was pragmatic role-playing (i.e., apologizing and giving excuse). However, Mehnert (1998) did not find any significant effect when the planning sessions were less than 10 minutes (i.e., 1- or 5-minute planning). Neither was there a significant difference between the planning and non-planning condition when the task was an instructional activity (i.e., explaining how to get to one's university from the airport). Hence, Mehnert's (1998) study was categorized under mixed.

Table 8
Lexical Complexity (18 studies)

Measure	*n* of studies	Results		
		effective (*n* = 8)	not effective (*n* = 11)	mixed (*n* =1)
TTR	5	1	4	0
Guiraud's index	4	4	0	0
Lexical Items/ Total words	2	2	0	0
Weighted lexical density	2	0	1	1
Number of word types	2	1	1	0
MSTTR	2	0	2	0
Lambda	1	0	1	0
MTLD	1	0	1	0
D-value	1	0	1	0

Note. TTR = Type-token ratio; MSTTR = Mean Segmental Type-token Ratio; MTLD = Measure of Textual Lexical Diversity

A closer look at the table reveals that researchers have focused on different aspects of lexical complexity. Fifteen studies examined the planning effect on lexical variety or the different types of words used in the speech: type-token ratio (TTR), Guiraud's index, the number of word types, mean segmental type-token ratio (MSTTR), Measure of Textual Lexical Diversity (MTLD), and D-value. Lexical density, or the proportion of content words in the speech, was measured by four studies, by calculating either the ratio of content words (i.e., lexical items) to total number of words or the weighted lexical density, which gives different weights to lexical items with lower and higher frequency. Only one study examined lexical sophistication (i.e., Lambda), which assesses the use of advanced vocabulary. It appears that

most of these pre-task planning studies have operationalized lexical complexity as the variety of word types available in the spoken production.

It is interesting to note that, although Guiraud's index and MSTTR are modified measures of TTR, these measures tend to reveal a different outcome. While TTR is dependent on text size or speech length, Guiraud's index divides types of vocabulary by square root of tokens, and MSTTR computes the average of TTR when the data is same-sized text samples.

According to van Hout and Vermeer's (2007) examination, Guiraud's index is "often a better transformation, at least from the perspective of concurrent validity" (p. 136) among these three measures. As shown in Table 8, most of the studies that used TTR and MSTTR did not find pre-task planning to be effective for lexical complexity. On the other hand, all four studies that employed Guiraud's index reported a significant increase in lexical complexity produced under planning condition.

Table 9 further analyzes the research design of 11 studies that used TTR, Guiraud's index, or MSTTR. Although there are differences in their research designs, these studies have common design features. Their participants were mainly adult English Language Learners (ELLs) at the intermediate level; eight studies were conducted in EFL contexts and the other three in ESL settings. The measures used in the latter three studies were not identical: one study used TTR, another focused on Guiraud's index, and the third implemented MSTTR. Furthermore, most of the studies offered 10-minute planning time under unguided settings. The main tasks, on the other hand, had slightly more variety, although most studies used narratives as their task.

The exceptions were two of the studies using Guiraud's index, which had the participants practice either instruction or decision-making tasks, and one MSTTR study, which had an opinion-giving activity. Hence, although several differences in research design features may have caused different task outcomes, the similarities among the 11 studies raise the possibility of lexical measures being a variable that affects the interpretation of the results.

Table 9

Comparison of Studies that Used TTR, Guiraud's Index, and MSTTR

Measures	*n* of studies	Research Context/ Design						
		participants	planning time (min)				guidance	main task
			0-4	5	8	10		
TTR	5	2 adult ELL (intermediate) 3 LOTE	-	-	1	4	4 unguided 1 guided	5 narratives
Guiraud's index	4	3 adult ELL (intermediate) 1 LOTE	-	1	-	3	4 unguided	2 narratives, 1 instruction, 1 decision-making
MSTTR	2	2 adult ELL (intermediate)	1	-	-	1	2 unguided	1 narrative 1 opinion-giving

Note. ELL=English Language Learners; LOTE = Language learners Other Than English

Measures for Syntactic Complexity

While few studies included syntactic complexity in their analysis, 36 studies focused on the effect of pre-task planning on syntactic complexity. Table 10 also shows that the studies considered here used a wide variety of measures to investigate this aspect of complexity. The majority of the studies were focusing on the number of clauses per unit or speech, but others mainly focused on the number of words, rather than on clauses. In other words, some of the studies were interested in counting how many segments the speakers were able to produce in their utterance, while others were interested in examining the length of those units. The most common unit was Analysis of Speech units (AS-unit), but T-units and c-units were also used in some studies. The clauses included not only independent clauses but also subordinate clauses (i.e., clauses containing a subordinating discourse marker) and dependent clauses (i.e., non-independent clauses other than subordinates). Less frequently observed was the use of verbs and Sentence-nodes (S-nodes). In total, 19 studies found pre-task planning to be effective for syntactic complexity, 28 studies found pre-task planning to not be effective, and seven studies had mixed outcomes.

Table 10

Syntactic Complexity Measures (36 studies)

Complexity Measures	*n* of studies	Results		
		effective (*n* = 19)	not effective (*n* =28)	mixed (*n* =7)
Clauses per AS-unit	12	7	5	0
Clauses per C-unit	8	3	4	1
Clauses per T-unit	5	2	2	1
Types of verb form	3	1	2	0
Words per T-unit	3	0	2	1
Words per utterance	3	2	1	0
S-nodes per T-Unit	3	0	3	0
Dependent clauses per T-unit	3	0	3	0
Number of passive voice	3	2	0	1
Dependent clauses per AS-unit	2	0	2	0
Subordinate clause per AS-unit	2	1	1	0
Number of subordinate clauses	2	1	0	1
Words per AS-unit	1	0	0	1
Words per clause	1	0	1	0
Words per C-unit	1	0	1	0
Discourse organization devices	1	0	0	1
Number of word families	1	0	1	0

Measures for Accuracy

All 40 studies examined oral accuracy in pre-task planning condition. In general, most of the studies found pre-task planning to not be significantly effective for improving oral task performance. As shown in Table 11, the 40 studies used global and/or specific measures to assess the development of oral accuracy. The data gathered for this synthesis suggest that past studies have a somewhat skewed emphasis on error-free clauses, or the number of clauses without any grammatical error. Over 70% of the studies defined accuracy as the number of error-free clauses. Few studies used units other than clauses, such as AS-unit and T-unit. On the other hand, researchers were also interested in counting how many errors the participants were making within a unit (e.g., number of errors per 100 words, clause) or their utterance in general. It is important to note that the operationalization of these measures as well was slightly different among the studies. For instance, in terms of error-free T-unit, some researchers regarded

sentences with up to one error as error-free T-units, whereas other studies adopted stricter criteria. Hence, the results of these studies might not always be comparable.

Table 11

Accuracy Measures (40 studies)

Accuracy Measures	*n* of studies	Results		
		effective (*n* = 12)	not effective (*n* = 45)	mixed (*n* =13)
Global Measures				
Error-free clause	30	7	18	5
Number of errors /100 words	6	1	4	1
Number of errors / Total words	2	1	1	0
Self-repairs/ Total error	2	0	2	0
Error-free T-units	2	1	1	0
Error-free AS-units	1	0	1	0
Errors per clause	1	0	1	0
Errors per T-unit	1	0	1	0
Errors per C-unit	1	0	1	0
Specific Measures				
Verb forms	10	1	6	3
Articles	6	0	5	1
Others (5 measures)	8	1	4	3

Specific measures, on the other hand, reflected the uniqueness of each study's target language. While the accurate use of verb forms and articles were commonly explored grammatical features, there were five other measures that were particularly unique to individual languages. Ortega (1999), for instance, examined Spanish language learners' article use and morphological agreement. Gutiérrez (2013) also investigated the planning effect on Spanish language learners' accuracy in gender and number concordances. In Tajima's (2003) study, the participants were Japanese language learners, and therefore the focus was on the use of Japanese participles.

Measures for Fluency

In total, 37 studies examined the effect of pre-task planning on fluency. Past pre-task planning studies examined oral fluency in terms of three different aspects proposed earlier by Tavakoli and Skehan (2005): speed fluency (i.e., speech rate), breakdown fluency (i.e., pausing and silence that interrupts the flow of speech), and repair fluency (i.e., repetition, replacement, reformation, and false starts that are used to repair the speech). Tables 12-14 summarize the measures and results for each of these fluency aspects.

Twenty-eight studies focused on the development of production rate. Among the three fluency aspects, production rate had the least variety of measures. As shown in Table 12, most of the studies adopted pruned speech rate per second (n = 5) or minute (n =18) and unpruned speech rate (n = 11). Interestingly, despite past reports on the effectiveness of planning on fluency, the studies synthesized here showed mixed outcomes for production rate. Studies that used Speech Rate A for their analysis tended to find an insignificant planning effect compared to those which used pruned speech rate. In addition, those studies that defined fluency as the mean length of run (i.e., mean number of syllables produced in utterances between pauses) were more likely to find positive planning effects on production rate. While most of the studies used syllables and/or words produced by the participants to measure their speech rate, two studies used mora, a non-syllabic phonological unit, for their analysis. These two studies targeted learners of the Japanese language, which is classified as a mora-based, rather than syllable-based, language.

Table 12
Production Rate (28 studies)

Fluency Measures	*n* of studies	Results		
		effective (n = 16)	not effective (n = 15)	mixed (n =9)
Speech Rate B (pruned speech rate)	16	6	6	4
Speech Rate A (unpruned speech rate)	9	1	6	2
Mean Length of Run	7	4	0	3
Pruned speech rate (per second)	5	4	1	0
Number of moras	2	1	1	0
Number of words per turn	1	0	1	0

Table 13 shows the results for repair fluency. For repair fluency, most of the researchers focused on the following four features: reformulation, repetition, false start, and replacement. Foster and Skehan (1999) defined repetition as repeated words, phrases, or clauses without any modification; false start as an unfinished utterance; reformulation as repeated phrases or clauses with modification; and replacement as words used to immediately replace an utterance. Note that these definitions are rather subjective, compared to the speech rate or pause length that can actually be measured using software such as ELAN. Past studies have counted or standardized these four self-repair indexes based on the speech time or words produced. Here, again, the effectiveness of pre-task planning seems unclear. For those studies that simply counted the self-repair indexes, researchers found significant planning impact. However, for those studies that used the standardized measures, there was no observable planning effect on participants' oral performance. Thus, in contrast to the widely accepted understanding in the literature (e.g., Skehan, 2016), planning seems less effective for fluency development in terms of self-repair.

Table 13

Repair Fluency (16 studies)

Fluency Measures	*n* of studies	Results		
		effective (*n* = 12)	not effective (*n* = 10)	mixed (*n* =11)
Reformulation	7	3	2	2
Repetition	6	3	1	2
False starts	6	3	1	2
Replacement	6	3	1	2
Self-repairs per 100 words	2	0	1	1
Self-repairs / Total time	2	0	2	0
Self-repairs / Total words	2	0	1	1
Clauses with self-repairs	1	0	0	1
Self-repairs per C-unit	1	0	1	0

Note. Self-repairs include reformulation, repetition, false starts, and replacement

Finally, Table 14 displays the results for breakdown fluency. In general, studies have defined breakdown fluency as either the amount of silence (i.e., amount of silence, pause length per total speech, mean length of pause, amount of silence per 100 words, length of filled pauses, length of pauses at the end of AS-unit), or their frequency (i.e., frequency of mid- and end-of-clause, filled pause, pauses per total words, silent or filled pauses per c-unit). A few more studies have focused on pause length (n = 20) than on frequency (n = 15). While earlier studies examined pauses in general, researchers such as Skehan and Foster (2005) adapted Davies's (2003) suggestion that pauses at the end of the clause boundary are more likely to distinguish native and non-native speakers' performances. Some researchers also distinguished filled pauses (e.g., use of fillers to fill in the silence) from silent pauses. It is also important to note that the operationalization of these pauses differed among the studies too. Some studies defined pause as a silence that is over 0.2 seconds long within a turn or utterance, while other studies set 0.4 seconds or 1 second as their criterion. Hence, the focus (i.e., frequency or duration), type (i.e., filled or unfilled; mid- or end-of- -clause), and criterion (i.e., duration of the silence) of the pause varied from study to study.

Table 14

Breakdown Fluency Measures (21 studies)

Fluency Measures	*n* of studies	Results		
		effective (*n* = 16)	not effective (*n* = 16)	mixed (*n* = 4)
Length of pause / Total length	7	2	5	0
Number of pauses	6	3	2	1
Amount of silence	6	3	1	2
Number of pauses (mid- / end-of-clause)	3	2	1	0
Number of filled pauses	3	1	1	1
Mean length of pause	2	2	0	0
Amount of silence/100 words	2	1	1	0
Number of pauses/ c-unit	2	0	2	0
Mean length of filled pause	2	0	2	0
Number of pauses/ Total word	1	1	0	0
Length of pauses (mid-/ end of AS-unit)	1	1	0	0
Filled pauses / c-unit	1	0	1	0

DISCUSSION

Research Procedures in Pre-task Planning Studies

The first aim of this study was to identify, by focusing on the following four aspects, how researchers have operationalized pre-task planning: (a) instruction given prior to the planning, (b) type of pre-task planning activity, (c) length of planning time, and (d) type of main task. The overview of these pre-task planning studies revealed that, at the present stage of research, a relatively small number of task and planning variables has been investigated. Most of the pre-task planning phases were unguided, requiring the participants to take notes silently on their thoughts. Following Mehnert's (1998) finding, that one minute of planning is effective in improving oral accuracy, planning time ranged from a minimum of 1 minute to a maximum of 10 minutes. The main task tended to be monologic, asking participants to simply explain a certain series of pictures in their target language to their partners.

This fairly narrow scope of variables has also been pointed out by Ellis (2009b). In his review of pre-task planning studies, Ellis (2009b) noted that "there is considerable scope…to identify other variables which might impact upon how planning time is used and, subsequently, the nature of the performance that results" (p. 222). While unguided self-planning makes it easier to control the variables, it limits our understanding and implementation of pre-task planning. If one of the aims of TBLT is to provide communicative and interactional experiences for learners (Ellis, 2009a), in actual classroom settings, peer or group planning activities are more purposeful and natural than self-planning. Individual planning also appears to be methodologically problematic. Although note-taking was a common planning activity, these notes do not fully reveal how the participants used their planning time (Ellis, 2009b). Some researchers did attempt to analyze participants' mental processes by combining verbal protocols with note-taking (e.g., Guará-Tavares, 2009), but such verbal reports may not accurately reflect learners' thinking processes (Gass & Mackey, 2000). Although the present study did not include computer-mediated studies, use of computer-supported technologies may help future researchers to record and track students' planning process. It is also important to note that the studies included in this paper were mostly conducted in laboratory settings, targeting ELLs.

Both theoretical and empirical evidence suggest that these contextual features may influence the results to some extent (Plonsky & Oswald, 2014).

An even greater source of concern, however, is the number of studies that lack precision (i.e., concerning the operationalization of guidance and planning). As the current synthesis demonstrated, what researchers regard as guided planning are not always identical. For instance, some studies offered a general guidance on the use of language, while others directed participants to particular language feature. Within the unguided planning studies, most of the participants were allowed to take notes, but some studies did not clearly report what was available for the learners during the planning time. While this is not a unique problem to pre-task planning studies (e.g., Norris & Ortega, 2006; Plonsky & Kim, 2016), this insufficiency makes it hard to identify task variables that support higher or lower levels of performance. Pre-task planning activities and main task factors interact in complex ways (Ellis, 2005b; Levkina & Gilabert, 2012). Most of the studies excluded from the current research, however, did not fully elaborate on how these factors were operationalized. Thus, despite the increasing number of reports, the state of pre-task planning research does not seem to be robust enough to make a reasonable interpretation.

CAF Measures in Pre-task Planning Studies

Another aim of this study was to explore the CAF measures used in the previous pre-task planning studies. In the 40 studies examined here, various CAF measures were employed. As pointed out by other researchers (e.g., Skehan & Foster, 2011), lexical complexity has received less attention in pre-task planning studies. Even among the 18 studies that did examine lexical complexity, the focus was heavily on lexical diversity, compared to richness of content (i.e., lexical density) and advanced vocabulary use (i.e., lexical sophistication). While further investigation is necessary to determine the extent to which these methodological varieties influence the research results, it would be useful to consider which of these measures help capture the dynamic aspect of learners' performance. According to Nitta and Nakatsuhara (2014), learners' performance is not consistent within a task. In their findings, participants

uttered more words in the beginning, but spoke less as the task continued (Nitta & Nakatsuhara, 2014). Interestingly, in the current research, those studies that used indexes with a higher degree of sensitivity to text length variations (i.e., Guiraud's index) had a different outcome from those with lower sensitivity (i.e., TTR and MSTTR). Hence, lexical indices that are more sensitive to speech length (e.g., Guiraud's index, D-value) might be useful in evaluating student's achievement, especially when the main task is longer.

Compared to lexical complexity, more studies analyzed syntactic complexity, accuracy, and fluency. In terms of syntactic complexity and accuracy, some studies had a narrower focus than others. For instance, most of the studies measured accuracy by counting the error-free clauses (i.e., general focus), while other studies counted the number of errors within error clauses (i.e., specific focus). More specific were those studies that examined the use of language-specific features (e.g., morphological agreement in Spanish). Other studies examined different dimensions of fluency: speech rate, repair, and breakdown. While the constructs underlying fluency appear to be clear (Pallotti, 2009), it is important to point out that there are diverse ways of defining these constructs. For example, pauses were defined as silences of more than one second in some cases (e.g., Skehan & Foster, 1997), whereas other researchers had a shorter criterion (e.g., Tavakoli & Skehan, 2005; Wigglesworth & Elder, 2010). These results therefore substantiate previous claims about the diversity of the CAF measures used in pre-task planning studies (e.g., Ellis, 2009b; Skehan & Foster, 2011).

In contrast to previous findings, however, the present study did not find clear evidence of planning effects on CAF. The number of studies that found planning effective for fluency did not significantly differ from the number of those that did not identify such a trend. Similarly, the majority of the studies reviewed in the current study did not find any significant planning effects on accuracy or syntactic complexity. One reason for this inconsistency can be the range of measures identified in the present study. In their synthesis, Skehan and Foster (2012) focused solely on a certain aspect of breakdown (i.e., the frequency of pauses and duration of silence)

and repair fluency (i.e., self-repairs per 100 words). Their focus is more specific than that of the present study. On the other hand, Ellis's (2009b) review of strategic planning included only 19 articles that were available at the time of the research. Hence, the effectiveness of pre-task planning in oral performance might vary depending on the CAF measures used in the study.

What this range of studies demonstrates is that there are numerous ways to assess learners' performance. Thus, as discussed by other TBLT researchers (e.g., Pallotti, 2009; Plonsky & Kim, 2016; Norris & Ortega, 2009), it is unfeasible to draw firm conclusions based on CAF measures. Perhaps then, for further investigation, future works should move beyond the framework of CAF measures. One possible approach is to assess how pre-task planning can help learners fulfill communicative adequacy (Revesz, Ekiert, & Torgersen, 2014). The development of CAF could be an indicator of learners' language proficiency, but the ultimate and genuine purpose of task-based learning is to get meaning across and accomplish the task (Ellis, 2009a). Rather than positioning CAF as the main tool to assess the effectiveness of pre-task planning, as Pallotti (2009) suggested, CAF could be used to help interpret the achievement of the task. For instance, researchers could examine whether pre-task planning helped learners accomplish a task and how the productive use of the target grammar points (i.e., accuracy) facilitated their activity. In addition to seeing CAF as a supplementary measure, as Norris and Ortega (2009) proposed, CAF measurements could be "considerably more organic, in the sense that they need to capture the fully integrated ecology of CAF development in specific learning contexts over time" (p. 556). The complex development of L2 production cannot be captured using a single CAF index. Data should be analyzed and interpreted in accordance to learners' developmental stage (for more details, see Norris & Ortega, 2009). In this respect, research into pre-task planning will have practical implications for pedagogical decision-making. As mentioned earlier in this paper, TBLT researchers (e.g., Lambert & Kormos, 2014; Pallotti, 2009; Plonsky & Kim, 2016) are currently calling for such a theoretically oriented use of CAF measures.

CONCLUSION

This study, a review of research on pre-task planning, documents the vast learner- and task-related factors as well as CAF measures examined in this area. The findings from the present study raise some methodological concerns, including the lack of adequate observations of learners' planning procedure itself. While previous studies in this area have no doubt advanced our understanding of the role of pre-task planning in L2 performance, a different approach from the trend (i.e., self-planning through note-taking) might reveal a clearer picture of what process exactly helps learners prepare for their task performance. Equally, reporting about the research design in enough detail will enable future replication and synthesis (Mackey, 2012).

The analysis of 40 studies also showed how CAF measures should be carefully chosen when researching the pre-task planning effect on oral performance. The results of the present study suggested the downside of using CAF measures to examine the effectiveness of pre-task planning. One of the limitations of the present study is the small number of studies available, especially those targeting non-English language learners. Each language has its own unique feature (e.g., morphological agreement in Spanish, mora in Japanese) that is different from the English language system. Although the present study collected data written in English, Japanese, and Spanish, adding sources in other languages may add more insights. In addition, while this paper reviewed the planning effect on oral CAF, adding its influence on written CAF may add more insights to the analysis. Nevertheless, the results of the present study provided a comprehensive overview of the methods used in pre-task planning.

REFERENCES

* = article included in the synthesis

*Bei, X. (2010). *Exploring task-internal and task-external Readiness: The effects of topic familiarity and strategic planning in topic-based task performance at different proficiency levels* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (ProQuest No. 3446044).

Birjandi, P., & Seifoori, Z. (2009). The effect of training and task-planning on the complexity of Iranian learners' oral speech. *The Journal of Applied Linguistics, 2*(1), 57-80.

Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogical tasks: Second language learning, teaching, and testing.* London: Longman.

Castro, A. P., Cabrera, P. M., & Martínez, B. P. (2009). El efecto de la planificación en la expresión oral de alumnos de primaria [The effect of planning on Primary children's oral language production]. *Porta Linguarum, 14,* 189-195.

Cleland, A. A., & Pickering, M. J. (2006). Do writing and speaking employ the same syntactic representations? *Journal of Memory and Language, 54*, 185-198. doi:10.1016/j.jml.2005.10.003

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

Davies, A. (2003). *The native speaker: Myth and reality* (2nd ed.). Clevedon, Avon: Multilingual Matters.

*D'Ely, R. C. S. F. (2006). *A focus on learners' metacognitive processes: The impact of strategic planning, repetition, strategic planning plus repetition, and strategic planning for repetition on L2 oral performance* (Unpublished doctoral dissertation). Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil.

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

Amsterdam: John Benjamins.

Ellis, R. (2005a). Planning and task-based performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 3-36). Amsterdam: John Benjamins.

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

Ellis, R. (2009a). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics, 19*(3), 221-246. https://doi.org/10.1111/j.1473-4192.2009.00231.x

Ellis, R. (2009b). 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. http://doi.org/10.1093/applin/amp042

Ellis, R., & Yuan, F. (2004). The effects of planning on fluency, complexity and accuracy in second language narrative writing. *Studies in Second Language Acquisition, 26,* 59-84. doi:10.1017/S0272263104261034

Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical & Statistical Psychology, 63*(3), 665-694. http://doi.org/10.1348/000711010x502733

*Foster, P. (2000). *Attending to message and medium: The effects of planning time on the taskbased language performance of native and non-native speakers.* (Unpublished doctoral dissertation). London: King's College.

Foster, P. (2001). Rules and routines: A consideration of their role in the task based language production of native and non-native speakers. In M. Bygate; P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 75-93). Harlow: Longman.

*Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(3), 299-323.

*Foster, P., & Skehan, P. (1999). The influence of source of planning and focus on planning on

task-based performance. *Language Teaching Research, 3*(3), 215-247.

*Fujita, T. (2006). *The influence of pre-task planning, on-line planning and their combination on fluency, complexity and accuracy in foreign language performance* (Unpublished master's dissertation). University of Essex, Essex, UK

*Gaillard, C. (2013). *The effects of pre-speaking planning on students' performance during speaking tasks* (Master's dissertation). Retrieved from All Theses and Dissertations (No. 4067)

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research.* Mahwah, NJ: Lawrence Erlbaum Associates.

*Genc, S. Z. (2012). Effects of strategic planning on the accuracy of oral and written tasks in the performance of Turkish EFL learners. In A. Shehadeh & C. A. Coombe (Eds.), *Task-based language teaching in foreign language contexts: research and implementation* (pp. 67-88). Amsterdam: John Benjamins.

*Geng, X., & Ferguson, G. (2013). Strategic planning in task-based language teaching: The effects of participatory structure and task type. *System, 41*(4), 982-993. http://doi.org/10.1016/j.system.2013.09.005

*Gilabert, R. (2007a). *Task complexity and L2 narrative oral production.* (Unpublished Doctoral Dissertation). University of Barcelona, Spain.

*Gilabert, R. (2007b). The simultaneous manipulation of task complexity along planning time and +/- Here-and-Now: effects on L2 oral production. In M.P. García-Mayo (Ed.), *Investigating tasks in formal language settings* (pp. 44-68). Clevedon, England: Multilingual Matters.

*Gilanioglu, I. (2002). *The effects of task features on lexis and grammar in L2 oral performance* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (ProQuest No. 1784694090)

*Guará-Tavares, M. G. (2008). *Pre-task planning, working memory, capacity, and L2 speech*

performance (Unpublished doctoral dissertation). Federal University of Santa Catarina, Florianópolis, Santa Catarina, Brazil.

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

*Gutiérrez, E. R. (2013). El papel de la pretarea en la producción oral de los aprendices de Español en la universidad Francesa [The role of the pre-task in the oral production of Spanish language learners at a French university]. *MarcoELE Revista de Didáctica Español Lengua Extranjera, 16*, 1-38.

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473. https://doi.org/10.1093/applin/amp048.

Housen, A., Vedder, I., & Kuiken, F. (Eds.). (2012). *Dimensions of L2 performance and proficiency investigating complexity, accuracy, and fluency in SLA*. Amsterdam: John Benjamins.

In'nami, Y., & Koizumi, R. (2010). Database selection guidelines for meta-analysis in applied linguistics. *TESOL Quarterly, 44*(1), 169-184. doi: 10.5054/tq.2010.215253.

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

Jackson, D. O., & Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning, 63*(2), 330-367. http://doi.org/10.1111/lang.12008

*Kanda, M. (2015). *Development of English oral proficiency among Japanese high school students* (Doctoral thesis). Retrieved from ProQuest Dissertations and Theses database. (ProQuest No.3745812)

*Kawauchi, C. (2005). *Pre-task planning in L2 oral performance: Quantitative and qualitative*

approaches. Tokyo: Kinseido.

Keck, C, Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between task-based interaction and acquisition: A meta-analysis. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 91-131). Philadelphia: John Benjamins.

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics, 35*(5), 607-614.

*Laudari, S. (2013). Planning and performance: The effect of form focused guided planning on learner performance. *Journal of NELTA, 18*(1-2), 92-104.

*Lee, H., & Oh, M. (2007). The effects of pre-task planning on L2 narrative tasks. In S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, & I. Vedder (Eds.), *Complexity, accuracy and fluency in second language use, learning and teaching* (pp. 127-138). Brussels, Belgium: KVAB.

*Lee, J. E. (2013). The effects of task planning time and instructions on the oral performance of young EFL learners. *Modern Studies in English Language & Literature, 11*, 293-319.

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

*Levkina, M., & Gilabert, R. (2012). The effects of cognitive task complexity on L2 oral production. In A. Housen, F. Kuikuen, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency investigating complexity, accuracy, and fluency in SLA* (pp. 171-197). Amsterdam: John Benjamins.

*Li, L., Cheng, J., & Sun, L. (2015). The effects of different lengths of pretask planning time on L2 learners' oral test performance. *TESOL Quarterly, 49*(1), 38-66.

Lin, H. (2014). Computer-mediated communication (CMC) in L2 oral proficiency development: A meta-analysis. *ReCALL, 27*(3), 261-287. https://doi.org/10.1017/S095834401400041X

Mackey, A. (2012). Why (or why not), when and how to replicate research. In G. K. Porte (Ed.), *Replication research in applied linguistics.* (pp. 34-69). Cambridge: Cambridge University Press.

*Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition, 20*(1), 83-108.

*Mehrang, F., & Rahimpour, M. (2010). The impact of task structure and planning conditions on oral performance of EFL learners. *Procedia Social and Behavioral Sciences, 2*(2), 3678-3686. https://doi.org/10.1016/j.sbspro.2010.03.572

*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*(1), 11-37. http://doi.org/10.1177/1362168807084492

*Nakakubo, T. (2011). *The effects of planning on second language oral performance in Japanese: Processes and production* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (ProQuest No. 879636294)

*Nielson, K. B. (2014). Can planning time compensate for individual differences in working memory capacity? *Language Teaching Research, 18*(3), 272-293. doi:10.1177/1362168813510377

*Nitta, R. (2007). *The focus-an-form effects of strategic and on-line planning: An analysis of Japanese oral performance and verbal reports* (Unpublished doctoral thesis). University of Warwick, UK.

*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., & Ortega, L. (2006). The value and practice of research synthesis for language learning and teaching. In J. Norris & L. Ortega (Eds.), *Synthesizing research on language learning and teaching* (pp. 3-50). Amsterdam: John Benjamins.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in

instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-578. https://doi.org/10.1093/applin/amp044

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing, 19,* 218-233.

*Ortega, L. (1995). *The effect of planning in oral narratives by adult learners of Spanish*. (Report No. 15) Honolulu, Hawai'i: University of Hawai'i, Second Language Teaching and Curriculum Center.

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

Oswald, F. L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics, 30*, 85-110. http://doi.org/10.1017/S0267190510000115

Oxford, R. L. (1990). *Language learning strategies: What every teacher should know*. New York: Newbury House/Harper Collins.

Pallotti, G. (2009). CAF: defining, refining and differentiating constructs. *Applied Linguistics, 30*(4), 590-601.

Park, S. (2010). The influence of pretask instructions and pretask planning on focus on form during Korean EFL task-based interaction. *Language Teaching Research, 14*(1), 9-26.

*Philip, J., Oliver, R., & Mackey, A. (2006). The impact of planning time on children's task-based interactions. *System, 34*, 574-565.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language instruction. In G. Crookes & S. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 9-34). Clevendon, England: Multilingual Matters.

Plonsky, L., & Kim, Y. (2016). Task-based learner production: A substantive and methodological review. *Annual Review of Applied Linguistics, 36,* 73-97. https://doi.org/10.1017/S0267190516000015

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*, 878-912.

*Rafie, F. R., Rahmany, R., & Sadeqi, B. (2015). The differential effects of three types of task planning on accuracy of L2 oral production. *Journal of Language Teaching and Research, 6*(6), 1297-1304. http://doi.org/http://dx.doi.org/10.17507/jltr.0606.17

*Rezaei, Z., & Tabatabaei, O. (2015). Individual differences in L2 speech performance: The role of working memory capacity and pre-planning tasks. *Mediterranean Journal of Social Sciences, 6*(2), 533-542. https://doi.org/10.5901/mjss.2015.v6n2s1p533

*Righettini, M. (2009). *The effects of planning time and language levels on task performance of English language learners* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses. (ProQuest No. 3387819).

Robinson, P. (1995). Attention, memory and the 'noticing' hypothesis. *Language Learning, 45,* 283-331.

Robinson, P. (2001). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and Second Language Instruction* (pp. 193-226). Cambridge: Cambridge University Press.

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

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*(3), 638-641. doi:10.1037/0033-2909.86.3.638

*Rutherford, K. (2001). *An investigation of the effects of planning on oral production in a second language* (Unpublished MA thesis). University of Auckland, Auckland, NZ.

*Saeedi, M. (2013). The influence of strategic planning and storyline complexity on EFL learners' narrative retellings. *The Journal of Language Teaching, 1,* 20-36.

*Saeedi, M. (2015). Unguided strategic planning, task structure, and L2 performance: Focusing on complexity, accuracy, and fluency. *Journal of Applied Linguistics and Language*

Research, 2(4), 263-274.

*Sangarun, J. (2001). *The effects of pre-task planning on foreign language performance* (Doctoral dissertation). Retrieved from ProQuest dissertations and Theses database. (Dissertation No. NQ63632)

*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. 111-42). Philadelphia: John Benjamins.

*Sasayama, S., & Izumi, S. (2012). Effects of task complexity and pre-task planning on Japanese EFL learners' oral production. In A. Shehadeh, & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 23-42). Amsterdam: John Benjamins.

*Shafaei, A., Salimi, A., & Talebi, Z. (2013). The impact of gender and strategic pre-task planning time on EFL Learners' oral performance in terms of accuracy. *Journal of Language Teaching and Research, 4*(4), 746-753. https://doi.org/10.4304/jltr.4.4.746-753

Skehan, P. (1996). Second language acquisition research and task-based instruction. In J. Willis, & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 17-30). Oxford: Heinemann.

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

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167-185). Harlow: Longman.

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

Skehan, P. (2016). Tasks versus conditions: Two perspectives on task research and their implications for pedagogy. *Annual Review of Applied Linguistics, 36,* 34-39.

*Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on

foreign language performance. *Language Teaching Research, 1*(3), 185-211.

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

Skehan, P. & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance: A synthesis of the Ealing research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 199-220). Amsterdam: John Benjamins.

*Spetch, A. L. (2014). *The impact of strategic planning instruction on learners' accurate oral performance of English as a foreign language* (Unpublished master's thesis). Federal University of Santa Catarina, Florianópolis, Brazil.

*Taguchi, N. (2007). Task difficulty in oral speech act production. *Applied Linguistics, 28*(1), 113-135. http://doi.org/10.1093/applin/aml051

*Tajima, M. (2003). *The effects of planning on oral performance of Japanese as a foreign language* (Doctoral dissertation). Retrieved from ProQuest dissertations and Theses database. (ProQuest No. 305319387)

*Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239-277). Amsterdam: Benjamins.

Thomas, M. (1994). Assessment of L2 proficiency in second language research. *Language Learning, 44,* 307-337.

van Hout, R. W. N. M., & Vermeer, A. R. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93-115). New York, NY: Cambridge University Press.

*Wang, Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on*

task performance (pp. 27-62). Amsterdam: John Benjamins.

*Wang, H., & Song, F. (2015). Effects of strategic planning time on L2 paired oral test performance. *Chinese Journal of Applied Linguistics, 38*(3), 263-276. https://doi.org/10.1515/cjal-2015-0017

*Wendel, J. N. (1997). *Planning and second language narrative production* (Doctoral dissertation). Retrieved from ProQuest dissertations and Theses database. (ProQuest No. 9813575)

*Wigglesworth, G. (1995). The effect of planning time on oral test discourse. In A. J. Kunann (Ed.), *Validation in language assessment: Selected papers from the 17th Language Testing Research Colloquium, Long Beach*. Mahwah, NJ: Lawrence Erlbaum Associates.

*Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing, 14*(1), 85-106. http://doi.org/10.1177/0265532297014001

*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. https://doi.org/10.1080/15434300903031779

Xi, X. (2005). Do visual chunks and planning impact performance on the graph description task in the SPEAK exam? *Language Testing, 22*(4), 463-508. http://doi.org/10.1191/0265532205lt305oa

*Yuan, F. (2001). *The effects of planning on language production in task-based language teaching.* Retrieved from ProQuest Dissertations and Theses. (ProQuest No. 9997306)

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

Zhigang, M., & Xudong, W. (2008). Interpreting performance under different task-planning conditions. *Babel, 54*(3), 201-233. http://doi.org/10.1075/babel.54.3.01ma

Appendix A

Planning, Main Task, and CAF Measurements of Studies Included in the Synthesis

Reference	Planning		Main Task	CAF Measures			
	Time (min)	Activity		Complexity - Lexical	Complexity - Syntactic	Accuracy	Fluency
Bei (2010)	10	Unguided note-taking	speech	Lambda lexical items/ total words	clauses per AS words per AS words per clause	(weighted) error-free clauses NER per 100 words	speech rate A (second) number of pause amount of silence / 100 words self-repair per 100 words number of unfilled pauses MLR
D'Ely (2006)	10	Guided note-taking (detailed): lexis, syntax, content, and organization - Dictionary allowed	narrative (video-based)	(weighted) lexical density	clauses per c-unit	error-free clauses errors per c-unit	speech rate A speech rate B mean length of filled pauses filled pause / c-unit amount of silence pauses per c-unit self-repairs per c-unit
Elder & Iwashita (2005)	3	Unguided note-taking	narrative (picture-based)	-	clauses per c-unit	error-free clauses	number of pauses per second self-repairs / time
Foster (2000)	10	(1) Unguided note-taking (2) Guided: Written guidance on the lexis, syntax, content and organization	(1) personal information (2) narrative (picture-based) (3) decision-making	-	clauses per AS passive voice	errors per AS	self-repairs (repetition, false-starts, reformulation, replacement) number of pauses amount of silence

Foster & Skehan (1996)	10	(1) Unguided note-taking (2) Guided: Written guidance on the lexis, syntax, content and organization	(1) personal information exchange (2) narrative (3) decision-making	-	clauses per c-unit	error-free clauses	self-repairs (repetition, false-starts, reformulation, replacement) number of pauses amount of silence
Foster & Skehan (1999)	10	(1) Unguided note-taking (2) Teacher-led: lecture on content or language (3) Group-based: Group brainstorming on content or language	decision-making	-	clauses per c-unit	error-free clauses	number of pauses amount of silence / 100 words self-repair per 100 words words per turn (number of words)
Fujita (2006)	10	Unguided note-taking - Dictionary allowed	narrative (picture-based)	type-token ratio	clauses per T-unit	error-free clauses	speech rate A speech rate B

Gaillard (2013)	5	(1) Unguided note-taking (2) Teacher-led: Brainstorm the content, vocabulary, organization, and connectors	personal information	-	words per utterance	error-free clauses NER / Total	speech rate A (second) duration in seconds pause length/ total time
Genc (2012)	10	Unguided note-taking	narrative (picture-based)	-	-	error-free clauses verb forms	-
Geng & Ferguson (2013)	10	Guided note-taking (detailed): language and content	(1) decision-making (2) opinion-giving	-	clauses per AS	NER per 100 words	speech rate A (second)
Gilabert (2007a)	10	Unguided note-taking	narrative (picture-based)	Guiraud's index lexical items/ total words ratio of lexical to function words	s-nodes per T-unit	error-free T-units self-repairs articles, ratio of repaired to unrepaired errors	speech rate A speech rate B
Gilabert (2007b)	10	Unguided note-taking	narrative (picture-based)	Guiraud's index	s-nodes per T-unit	self-repairs	speech rate B

Guará-Tavares (2009)	10	-	narrative (picture-based)	-	clauses per c-unit	NER / Total	speech rate A speech rate B
Guará-Tavares (2008)	10	Unguided note-taking - Verbal protocol	narrative (picture-based)	-	clauses per c-unit	NER per 100 words error-free clauses	speech rate A speech rate B pauses per c-unit pause length/ total time
Kawauchi (2005)	10	(1) Unguided note-taking (2) Rehearsal (3) Silent model reading	narrative (picture-based)	word types	clauses per T-unit words per T-unit	verb forms	(total amount of speech) repetition
Lee & Oh (2007)	10	-	narrative (picture-based)	-	dependent clauses per T-unit words per T-unit	error-free T-unit errors per T-unit	pause length/ total time number of pauses / total words
Levkina & Gilabert (2012)	5	Unguided note-taking	decision-making	Guiraud's index	clauses per AS	errors per AS	speech rate B
Mehnert (1998)	1, 5, 10	Unguided note-taking	(1) Personal information (2) Pragmatic role-play	weighted lexical density	s-nodes per T-unit dependent clauses per T-unit length of c-unit	error-free clauses NER per 100 words word order and word choice	speech rate A speech rate B MLR number of pauses amount of silence
Nakakubo (2011)	10	Unguided note-taking	narrative (picture-based)	type-token ratio	clauses per T-unit words per T-unit	error-free clauses NER per 100 words number of non-repeated errors per 100 words	(meaningful) moras per minute pause length/ total time

Nitta (2007)	10	Unguided note-taking	narrative (picture-based)	-	clauses per AS discourse organization	error-free clauses verb forms articles	speech rate B amount of silence MLR pauses (mid- / end- of clause) number of filled pause self-repairs (repetition, false-starts, reformulation, replacement)
Nitta & Nakatsuhara (2014)	3	Unguided note-taking	decision-making	MTLD	clauses per AS	NER per 100 words	self-repairs / total words pause length / total time words per second
Ortega (1999)	10	Unguided note-taking	narrative (picture-based)	type-token ratio	words per utterance	morphology agreement; Spanish articles	speech rate B
Ortega (1995)	8	Unguided note-taking	narrative (picture-based)	type-token ratio	words per utterance	morphological agreement; Spanish definite articles	self-repairs / total words reformulation
Rafie, Rahmany, & Sadeqi (2015)	10	Unguided note-taking	personal information	-	-	error-free clauses	-
Rezaei & Tabatabaei (2015)	10	-	picture description and opinion	-	dependent clauses per AS	error-free clauses	number of pauses
Saeedi (2013)	10	Unguided note-taking	narrative (picture-based)	-	clauses per AS	error-free clauses	speech rate B

Saeedi (2015)	5	Unguided note-taking	narrative (picture-based)	-	clauses per AS	error-free clauses	speech rate B
Skehan & Foster (1997)	10	-	(1) personal information exchange (2) narrative (3) decision-making	-	clauses per c-unit	error-free clauses	pauses (mid- / end- of clause) MLR (phonation time)
Skehan & Foster (2005)	10	Unguided note-taking	decision-making	-	clauses per AS	error-free clauses (weighted)	pauses (mid- / end- of clause) number of filled pause MLR self-repairs (repetition, false-starts, reformulation, replacement)
Spetch (2014)	10	Unguided note-taking	narrative (picture-based)	-	-	error-free clauses errors per clause	-
Tajima (2003)	10	Unguided note-taking	instructions	Guiraud's index	clauses per AS	error-free clauses target-like use of particles	pause length/ total time number of moras / total time
Tavakoli & Skehan (2005)	5	Unguided note-taking	narrative (picture-based)	-	clauses per AS	error-free clauses	MLR mean length of pause number of pauses amount of silence speech rate A (second) self-repairs (repetition, false-starts, reformulation, replacement) phonation time
Wang	3	Unguided	narrative	D-value	clauses per AS	error-free clauses	pause length (mid-/ end of

(2014)		note-taking	(video-based)		subordinate clauses per AS		AS-unit) reformulation speech rate A (second)
Wang & Song (2015)	1, 2, 3	Unguided note-taking	opinion-giving	MSTTR	clauses per c-unit verb forms	error-free clauses verb forms	speech rate A speech rate B
Wendel (1997)	10	Guided note-taking (undetailed): language and content	narrative (video-based)	word types	number of subordinate clauses passive voice number of word families	verb forms	speech rate B mean length of pause
Wiggleswor th (1995)	1	Unguided note-taking	(1) listening summarization (2) telephone message leave	-	dependent clauses per T-unit	verb forms morphemes indefinite articles	clauses with self-repairs self-repair
Wiggleswor th (1997)	1	Unguided note-taking	(1) picture description (2) listening summarization (3) telephone message leave (4) personal information	-	number of subordinate clauses	verb forms morphemes indefinite articles	type-token ratio self-repair
Wiggleswor th & Elder (2010)	1, 2	Unguided note-taking	personal information	-	dependent clauses per AS subordinate clauses per AS	error-free clauses error-free AS	pause length / total time mean length of filled pauses self-repairs / time
Yuan (2001)	10	Guided note-taking (undetailed): language and content	narrative (picture-based)	type-token ratio	verb forms clauses per T-unit	error-free clauses verb forms	speech rate A speech rate B

Yuan & Ellis (2003)	10	Unguided note-taking	narrative (picture-based)	MSTTR	clauses per T-unit verb forms	error-free clauses verb forms	speech rate A speech rate B

Appendix B

Examples of Guidance for Planning

1. Unguided Planning

> You have 10 minutes to plan what to say and how to say it. Please make written notes in German, but do not write out everything in detail. You then have to talk without your notes.
>
> (Mehnert, 1998, p. 108)

2. Guided Planning (less specific)

> You can make notes during the ten minutes. but you won't be allowed to use these notes while doing the task. These are things you do to help you prepare:
> - think what problems your listener could have and how you might help her
> - think about how your listener can understand the order of the things she to do.
> - think of ways to make sure your friend won't get lost
> - think what grammar you need to do the task
> - think what vocabulary you need to do the task
> - think how to avoid difficulties and problems with grammar and vocabulary
>
> (Foster & Skehan, 1996)

3. Guided Planning (more specific)

(ア) Think of all words you want to use in your message, and note only one word for one meaning

(イ) Think of transition words or phrases, such as *first, second, next ... finally* that will connect your instructions so that it is easy for your friend to follow them.

(ウ) Think of grammatical structures that play an important role in the task and write the main parts of the grammatical structures.

Grammatical structures that are common and needed for instruction speech are:
- imperative form of verb
- should, must, can + verb 1
- prepositions
- Present Simple tense

(Sangarun, 2001, p. 309)