Effects of task type and language proficiency on dialogic performance and task engagement

Effects of task type and language proficiency on dialogic performance and task engagement

Abstract

This study examined the effects of task type and English proficiency on L2 learners' task performance and engagement. By collecting data from 15 learner dyads at three levels of proficiency (elementary, intermediate and advanced) performing three dialogic tasks (personal information, narrative and decision-making tasks), we examined their performance in terms of complexity, accuracy, lexis and fluency (CALF), and degree of task engagement in three dimensions of cognitive, social and behavioural engagement. The results suggested that task type had an impact on all aspects of linguistic performance as well as social and behavioural engagement. While the personal information task elicited the most fluent and accurate language, it was the least engaging task in terms of social engagement. The narrative task elicited the most syntactically complex language, but it was the lowest in terms of behavioural engagement. Language proficiency influenced accuracy and fluency of performance as well as cognitive engagement with the task. The results showed that advanced learners were the most fluent, accurate and cognitively engaged group of learners across the tasks. In general, the results suggest that task type not only encourages specific dimensions of performance in CALF measures, but it could also affect learners' cognitive, social and behavioural engagement.

Keywords: CALF, Task, Task-Based Research, Task Engagement, Task Performance

1. Introduction

Using *tasks* in second language (L2) teaching has been popular for decades as these pedagogic materials are assumed to help "engage naturalistic acquisitional mechanisms, cause the underlying interlanguage system to be stretched, and drive development forward" (Skehan, 1998, p.95). Tasks are also believed to promote L2 learning through learning by doing (Willis & Willis, 2007), and provide learners with "opportunities for authentic use of language in the classroom" (Faez & Tavakoli, 2019, p.2). Motivated by this, considerable effort in task-based research has been put into investigating the effects of task design on different aspects of L2 performance and learning. Many such studies have typically looked into the complexity, accuracy, lexis and fluency (CALF) of learner performance in order to gain a better understanding of L2 processing, production and development (Ahmadian, 2012; Foster & Skehan, 1996; Dörnyei & Kormos, 2000; Revesz et al., 2019; Tavakoli & Foster, 2008; to name a few). One line of inquiry in this field has focused on examining the effects of task design on L2 performance and acquisition (Kormos, 2014; Michel et al., 2019). While much of this research thus far has used monologic tasks to investigate the interplay between task types and oral performance, studies on how the design of dialogic tasks affects learners' oral production are scarce. For example, one such study which used two dialogic tasks (a narrative task and a decision-making task) was conducted by Foster and Skehan (1996). However, there is still a need for studies which examine the interplay between dialogic tasks and learners' oral production, particularly as they develop their proficiency. Examining dialogic tasks is important as dialogue is a typical and recurrent kind of human linguistic interaction and

a frequent mode of speech in everyday communication (Tavakoli, 2016; Michel et al., 2007). From an L2 learning perspective, examining dialogic task types is crucial since L2 research has provided strong evidence that dialogues provide a rich opportunity for learning through interaction and negotiation of meaning (Long, 1981; Mackey, 1999).

Another important aspect of task design rarely examined is *task engagement*. Task engagement, or "the degree to which learners are intensely involved in a learning task" (Aubrey, 2017, p.661), is hypothesized to play a crucial role in L2 learning (Butler, 2017; Phung, 2017). Research in mainstream education has provided ample evidence about the relationship between task engagement and academic success (Fredricks et al., 2004; Newmann, 1991; Reschly & Christenson, 2012), suggesting that pedagogic activities that promote engagement have a positive impact on learning and educational achievement. Given the strong evidence presented by educational studies, it is surprising that little research has been conducted in L2 contexts to examine task engagement and its effects on language learning. Specifically, little is known about: 1) whether task type and individual learner variables such as proficiency level have an impact on learner engagement, 2) whether such variables interact with each other to enhance or diminish engagement, and 3) what potential such variables have on performance and learning. This is a gap the current study sets out to help fill by examining the effects of dialogic task types and proficiency levels on task performance indicated by levels of CALF and dimensions of task engagement.

2. Task engagement

Although interest in engagement seems rather new in L2 research, *engagement* as a concept has been central to pedagogy for a long time as it promotes interaction, classroom participation and learner autonomy (Fredricks et al., 2004). Philp and Duchesne (2016) defined engagement as "a state of heightened attention and involvement" in the process of learning (Philp & Duchesne, 2016, p.51). To them, the significance of engagement is explained in the light of its connection to attention and conscious mental involvement, which will in turn make learners cognitively prepared for the process of learning. Other researchers contend that engagement reflects learners' interest and participation, and generates rich opportunities for effective learning processes and outcomes (Butler, 2017; Newmann, 1991; Philp & Duchesne, 2016; Phung, 2017), and yet others argue that the primacy of engagement lies in reflecting learner motivation which leads to effective learning (Fredricks et al., 2004). Regardless of what it actually represents, most researchers concede that engagement is a complex and multidimensional construct (Fredricks et al., 2004) with at least four dimensions: cognitive, behavioural, emotional and social (see Philp & Duchesne [2016] for a full discussion).

Cognitive engagement refers to the extent to which learners are engaged in a task through processes of sustained attention and mental effort (Fredricks et al., 2004; Helme & Clarke, 2001). When cognitively engaged in a task, learners appear to exchange ideas and ask questions more frequently. Fredricks et al. (2004) argue that cognitively engaged learners use metacognitive strategies to plan, self-regulate and assess their performance and its impact on their cognition. Other indicators of cognitive engagement are the use of discourse markers and connectors as well as questioning and exchanging ideas (see Helme & Clarke, 2001; Philp & Duchesne, 2016). These indicators of task engagement have been corroborated in a study conducted by Kang and Wang (2014). The participants who were actively engaged were those who used more discourse markers and a variety of interactional features (i.e., back-channelling, prompting and new topic initiations). For them, the use of these interactional features shows "an improved degree of engagement

in a conversation" (Kang & Wang, 2014, p.48).

Behavioural engagement refers to the physical aspects of engagement (Fredricks et al., 2004; Philp & Duchesne, 2016) such as time on task or number of turns taken in a dialogic task. Behavioural engagement is often measured quantitatively in terms of word and turn counts (see Bygate & Samuda, 2009; Dörnyei & Kormos, 2000), or qualitatively through classroom observations for participation, effort and conduct (Fredricks & McColskey, 2012). Emotional engagement refers to learners' emotions during task performance (Fredricks et al., 2004), and comprises both positive emotions, such as enthusiasm, interest, and/or enjoyment, and negative emotions including frustration, anxiety and boredom (Skinner et al., 2009). The literature on emotional engagement fails to show a unanimous consensus about what constitutes this construct or whether it can be objectively measured. Some researchers (e.g., Yazzie-Mintz, 2009) defined emotional engagement in relation to learners' feeling of connectedness or their sense of belonging to the educational context; others (e.g., Skinner et al., 2009) argue that emotional engagement represents learners' enthusiasm, anxiety and motivation. The lack of certainty about what constitutes emotional engagement and the unknown "source of the emotional reactions" (Fredricks et al., 2004, p.63) makes measuring emotional engagement difficult. Finally, *social engagement* demonstrates the extent to which learners are socially involved with one another during task completion (e.g., whether they listen to one another or help each other with ideas, experiences and skills required to complete the task). Examples of social engagement include collaborative moves during a conversation, scaffolding one another's language use, and negotiating meaning.

In sum, L2 researchers agree that task engagement is central to L2 learning particularly in classroom contexts, and as such, examining task engagement can shed light on the complex nature of L2 acquisition. Some researchers (Baralt et al., 2016) argue that focusing on only one dimension of task engagement, although revealing and helpful, would provide only a partial picture of the complex phenomenon, and therefore examinations of task engagement in its full capacity is recommended. For this reason, we aim to examine engagement in different dimensions: behavioural, cognitive and social.

3. Task type and task engagement

Motivated by the claim that tasks not only advance L2 learning, but also shape and facilitate learning processes and outcomes (Bygate, 1999; Ellis & Barkhuizen, 2005; Foster & Skehan, 1996; Walsh, 2002), a growing body of research has investigated the effects of task type on L2 performance and development (e.g., Bygate, 1999; Ellis, 2005; Foster, 1996; Foster & Tavakoli, 2009; Gilabert, 2007; Ortega, 1999; Skehan, 2009; Yuan & Ellis, 2003; Wigglesworth, 1997; to name a few). In this field, several task type taxonomies have been proposed over the past decades. Prabhu (1987), for example, classified tasks in terms of the kind of gap they presented and proposed three categories: *information gap, reasoning gap* and *opinion gap*. Pica et al. (1993) classified tasks in terms of task goals being *convergent* or *divergent*. *Convergent* tasks, in this sense, require learners to arrive at a consensus in order to complete the task, whereas *divergent* tasks allow learners to have different views on and an open discussion about a topic without having to reach an agreement. Based on this classification, therefore, a decision-making task in which one single decision has to be made is a *convergent* task, while a decision-making task that allows for different decisions to emerge is considered a *divergent* task. Another frequently cited classification is Ellis's (2009) categories of *focused* and *unfocused* tasks. According to Ellis, *focused* tasks refer to those designed to provide "opportunities for communicating using some specific linguistic feature" (p.223),

whereas *unfocused* tasks offer "opportunities for using language in general" (p.223) without concern for using a specific form. In this framework, the main aim is to distinguish tasks based on their linguistic requirements. While these taxonomies have been very useful in providing researchers and teachers with a framework to define and describe tasks, we consider them limited in at least two different regards. First, they all focus on only one aspect of the task design (e.g., the kind of gap or the linguistic requirement) rather than providing an overarching and comprehensive framework for analysing and evaluating task design. We consider *the kind of gap, task goal*, and *linguistic requirements* as representative of some of the characteristics necessary to consider when evaluating task type. Secondly, the proposed frameworks presume task qualities as dichotomies (e.g., tasks are either convergent or divergent), while we propose these qualities should be regarded as continua (e.g., a task is more convergent or less divergent than another). It is necessary to note that in this paper we do not aim to define task type taxonomies or to develop one; rather, we are interested in task types that are frequently used by teachers and researchers.

Task-based research has also paid considerable attention to investigating task complexity and its effects on task performance (Robinson, 2003; Skehan, 1998, 2009). This body of research has so far presented two models of task complexity: Skehan's (2009) Limited Attentional Capacity and Robinson's (2003) Cognition Hypothesis. The first model proposes that task complexity should be realised in the light of the fact that human's attentional resources are limited in L2 acquisition, and as such learners will not be able to attend to both form and meaning when performing a demanding L2 task. On the other hand, the second model suggests there are multiple pools of attention available to L2 learners and therefore a complex task pushes the learners to attend to both form and meaning resulting in language output that is of high linguistic complexity and accuracy (Robinson, 2003). Typical to this body of research is manipulating task design variables in order to see whether performance varies as a result of such operationalisations, and whether manipulating task complexity would facilitate L2 production and development. While this body of literature has been informative and beneficial for task design, we will not focus on these models in this study as we do not aim to manipulate task design features in the light of the variables proposed by either model. As discussed below, our study is focused on replicating the findings of Foster and Skehan (1996) with regard to task performance and engagement across different proficiency levels, and, therefore, issues related to task complexity models will not be discussed in this paper.

In one of the earliest studies investigating task type, Foster and Skehan (1996) examined the effects of three task types on performance indicated by levels of fluency, complexity, and accuracy of a group of pre-intermediate English language learners. The task framework they used consisted of three different tasks (personal information exchange, narrative, and decision-making) which were implemented under three conditions (with no planning, with planning but without detail and detailed planning). Their results suggested that the personal information exchange task elicited accurate and fluent language, but was associated with low syntactic complexity; the narrative task elicited high levels of complexity but low levels of accuracy; and the decision-making task elicited useful levels of accuracy and complexity. Since the publication of Foster and Skehan (1996), several studies have used these task types to examine different aspects of language performance and development (Michel et al., 2007; Qian, 2014). Kuiken and Vedder (2007) conducted a study to investigate the effects of characteristics of different tasks, operationalised in terms of task complexity, on learners' performance. Their results showed that task type affected learners' accuracy and lexical choices. Gilabert (2007) also investigated the effects of different task types (narrative, instruction-giving, and decision-

making tasks) on self-repairs as a measure of accuracy. His results showed an overall effect of task type across the three tasks. He also reported a dynamic interaction between the number of self-repairs and task type. While discussing the findings of all studies investigating task type on performance is beyond the scope of the current paper, the few example studies discussed here generally show the effects of task type on L2 performance.

Although *dialogue*, given its interactive nature, reflects language use more authentically and naturally than *monologue* (Guillot 1999; Van Lier 2004), task-based research has extensively examined monologic task performance (de Jong & Perfetti, 2011; Skehan & Foster, 1996; Tavakoli, 2011). A preference for examining monologic task performance in this area is often explained in terms of the degree of control associated with the predictability of the outcome of task performance, and clarity and ease of measuring learners' output in a monologic task (Tavakoli, 2016). Galaczi and Taylor (2018) summarise the challenges involved in assessing dialogic task performance in relation to:

- complexities embedded in examining and measuring dialogic performance,
- interaction not being "linear, predictable or tidy",
- dialogue being shaped by "personal, cognitive and contextual factors" that might be difficult to measure.

(p.219)

As discussed earlier, task engagement is an under-researched area in task-based research. One of the earliest studies exploring task engagement was done by Lambert et al. (2017). Comparing tasks with learner- and teacher-generated content, this research examined the effects of these two types of tasks on learner engagement in terms of behavioural, cognitive, and social components. Their results indicated that learner-generated tasks had a beneficial impact on all aspects of engagement. Phung (2017) examined task engagement in relation to learners' task preferences. Measuring engagement across behavioural, social and cognitive dimensions, Phung (2017) reported significant effects of learner preference on engagement, that is to say, learners were cognitively, socially and behaviourally engaged in the tasks they preferred. Dao (2020) investigated engagement in divergent versus convergent tasks. The study examined 16 dyads on their cognitive (idea units and language-related-episodes, LREs), emotional (explicit task enjoyment), and social (responsiveness) engagement while performing the tasks. The analysis involved a mixed-methods approach using scores for idea units and LREs and content analysis of a post-task questionnaire on learners self-reported accounts of engagement. The results suggested that learners showed greater cognitive and social engagement when involved in convergent tasks compared to divergent tasks. A study which investigated the relationship between task types (monologue and dialogic tasks), proficiency levels (B1, B2, C1 and C2) and engagement by means of interactive features (discourse management, grammatical resources, lexical resources and pronunciation) was conducted by Kang and Wang (2014). In the dialogic tasks, they found that higher-proficiency speakers generally had more discourse markers and turns between the two interlocutors. Their results suggested that the higher-level participants were more active in using interactive and cooperative features with their interlocutors. This group also showed more task engagement in the dialogic tasks. As can be seen, there are very few studies investigating task engagement, highlighting the need for further research into the relationship between task type and task engagement.

4. Proficiency level and language performance

From an L2 development perspective, exploring the ability to engage competently in a speaking task is linked to proficiency level (Galaczi, 2008; Galaczi & Taylor, 2019). Language proficiency, or "the linguistic knowledge and skills that underlie L2 learners' successful comprehension and production of the target language" (Gaillard & Tremblay, 2016, p.420), is considered an important variable that influences L2 performance and learning. This was corroborated by Kim et al. (2016) in a study in which 130 L2 learners were asked to perform two story-retelling tasks orally and in written form. By measuring levels of complexity, accuracy and fluency in learners' written and oral performance, the results showed a strong correlation between learners' proficiency and CALF levels. Several other studies including Kahng (2014) and Revesz et al. (2019) have demonstrated that fluency is significantly associated with proficiency. Such findings provide us with strong evidence of the relationship between oral performance and proficiency.

Given its impact on different aspects of language learning, it is intriguing to find out whether proficiency level has also an impact on learner engagement and performance. Despite the importance of proficiency in task performance and oral development, there remains a paucity of evidence on the relationship between task type, proficiency level, and task engagement. In this study, our interest in proficiency level is motivated by two factors. Firstly, we are interested in examining the effects of task type on language performance in relation to learners' proficiency levels. In using tasks, it has been assumed that as learners develop their proficiency, their language performance becomes more competent in CALF. The current study allows us to examine whether a linear interaction can be expected between performance and proficiency level in terms of CALF measures. Investigating effects of task type on oral performance across three different levels of proficiency is a gap in the literature the current study aims to help fill. Second, in order to develop a better understanding of task engagement, we are keen to examine engagement across different levels of proficiency, that is, elementary, intermediate and advanced levels, to see if development of proficiency is associated with more engagement. This is another novel aspect of our research as previous studies have not provided a perspective of engagement across different task types and proficiency levels. Thus, measures of CALF and social, behavioural and cognitive engagement are considered as dependent variables, whereas task type and proficiency level are the independent variables of the study.

5. Research aims and questions

Following Foster and Skehan (1996), we are interested in experimenting with three task types: personal information, narrative and decision-making tasks. Our rationale for using these tasks is based on both research and pedagogic principles. From a research perspective, the detailed results provided by Foster and Skehan (1996) about performance in these three task types allowed us to replicate their study in our new design (i.e., across different proficiency levels and in learners' natural setting, their L2 classroom). By examining classroom data from 15 learner dyads at three levels of proficiency (elementary, intermediate and advanced) performing the three tasks in a dialogic fashion, we aim to provide an insight into: a) how the different task types affect task engagement (in cognitive, behavioural and social dimensions) and task performance (indicated by levels of CALF), b) whether such effects, if any, are consistent across three levels of proficiency, and c) whether there is a relationship between task performance and task engagement. From a pedagogic point of view, given that many teachers use these task types frequently in language classrooms of different proficiency levels, it seems necessary

to examine in what ways these tasks affect learner performance and engagement. Three research questions guide our study:

RQ1: To what extent is learners' task performance, measured in CALF, affected by task type and proficiency level?
RQ2: To what extent is learners' task engagement affected by task type and proficiency level?
RQ3: Is there a relationship between task performance and task engagement?

What our study adds to the literature in this area is: 1) how these three task types influence learners' CALF and task engagement, 2) whether task type effects on performance reported by Foster and Skehan (1996) can be replicated across different levels of proficiency, and 3) whether the task effects, if any, interact with proficiency levels. To the best of our knowledge, this is the first study examining L2 learners' dialogic performance and engagement across three different task types and three proficiency levels.

6. Methodology

6.1 Context

The study took place at the language centre of a Mexican university located in central Mexico. The study was conducted in three on-going EFL classes at elementary, intermediate and advanced levels. The classes met for five hours per week focusing on the language skills (i.e., speaking, writing, reading and listening), grammar and vocabulary. The data were collected in three consecutive weeks during speaking practice, each week focusing on one of the three tasks. The sequence of the tasks was counterbalanced across the groups and dyads to avoid any practice effect. Table 1 below shows the counterbalance design of the study.

Table 1. Counterbalanced design of task performance across proficiency levels

	Week 1		**Week 2**		**Week 3**	
	Group 1 (3 dyads)	Group 2 (2 dyads)	Group 1 (3 dyads)	Group 2 (2 dyads)	Group 1 (3 dyads)	Group 2 (2 dyads)
Elementary	Task A	Task B	Task B	Task C	Task C	Task A
Intermediate	Task B	Task C	Task C	Task A	Task A	Task B
Advance	Task C	Task A	Task A	Task B	Task B	Task C

Note. A is the personal information, B is the narrative and C is the decision-making task.

6.2 Participants

We advertised our research project in the three EFL classes. All 45 learners enrolled in these classes expressed interest in participating in the study. Besides learning English, the 45 learners were studying undergraduate and postgraduate programmes from different disciplines in this university. They had all studied English for a minimum of six years in junior and high school prior to commencing their university studies. The learners were highly motivated about learning English because level of proficiency is an institutional requirement to graduate which varies across different disciplines in this university. Although the learners were placed in their classes through a university placement test, to ensure differences between the groups and to check the homogeneity of each group, we

further assessed their language proficiency level through the grammar section of the FCE test (Cambridge, 2018). In effect, our assessment of the participants' proficiency relied on both a university-wide placement test and our FCE grammar test. The learners' proficiency levels on the FCE test were categorised following the Cambridge guidelines (see Appendix 1). Based on the results of the FCE test, those learners who obtained scores between 0-7 were classified as elementary, those with a score of 8-11 as intermediate, and those achieving 12-18 as advanced proficiency level. As a result, the participants of the study are those who were placed in these proficiency level groups by the university and whose FCE grammar scores were in the range suggested above. While we acknowledge that the grammar section of the FCE test is limited in scope especially for providing an insight into the learners' speaking ability, we used these results as a post-hoc validation of the university placement exam that includes both written and spoken tests.

To ensure the class practices were not disrupted, we asked all 45 learners in these three classes to perform the tasks, but our database reported in this study is based on 30 participants because, after the data collection was complete, we removed the data that failed to meet our criteria for proficiency, attendance and clarity. First, based on the results of the FCE grammar test, data from five participants who were not in the expected range of language proficiency scores for their levels were excluded from the study. Then, we removed data from five participants (and inevitably their partners in a dyad) who either missed one of the sessions or whose audio recordings were not of good quality (e.g., too much background noise). As a result, the data we are reporting here come from 30 participants, 15 dyads, five dyads from each proficiency level (13 males and 17 females; aged 18 to 26). Given the one-week interval between task performances, we expected very little practice effect. Yet, the tasks were performed by the learner dyads in a counterbalanced sequence to avoid any practice effect of a specific task. Written ethical consent was obtained from all the participants before the study started.

6.3 Tasks and data collection procedures

In this study, we are particularly interested in dialogic tasks as they are central to L2 learning and development. For its purpose, we consider a monologue as production of sequences by one speaker, whereas a dialogue is regarded as "prototypically a joint enterprise involving more than one person" (Cameron, 2001, p.87). According to research in this area (Edwards, 2008), in addition to the core component of turn-taking (i.e., speakers taking turns to communicate), interruption, backchannel and simultaneous talk are other important features of a dialogue. The contribution of each speaker to a dialogue, however, may vary according to the purpose, context, individual differences and communicative function of the speaking activity. Given the purpose of the current study, we were interested in dialogic tasks in which both speakers had equal opportunities for communicating their message. Use of dialogic tasks which offer the participants equal opportunities for speaking was therefore deemed necessary as it allowed us to examine social and behavioural aspects of engagement. To achieve this aim, three dialogic tasks were used to collect data. The participants were told that they had to take both listener and speaker roles and take turns to speak, but they had some freedom about structuring their dialogues (e.g., who would start first or how long each turn is). Details of each task are presented below.

The first task was a personal information task in which the learners asked each other ten personal questions in the form of an interview using small cards with cues for questions for each side of the dyad (e.g., how do you like to spend your holiday? What's

your favourite day of the week?). The instructions indicated that they had five minutes to perform the task and that both members of a dyad had to read from the same cue cards asking questions and providing a response to each question. This was done to ensure they actively took part in the dialogue in both roles (asking and answering). The students had the freedom to choose how to approach the task (e.g., how many questions they asked in one turn), but in the data most dyads decided to alternate turns to ask questions and formulate responses. While some participants, predominantly from the elementary group, read out the questions from the cards in a verbatim manner, most participants re-produced the questions using their own words (e.g., changing words or the structure of the questions). Given that all the participants had identical prompts and the same opportunities for using them, we analysed each participant's output including the few verbatim questions we identified in the dataset[1]. Following Foster and Skehan (1996), we hypothesized that this task would not be cognitively demanding as it involved familiar personal information and a task type that they were already familiar with.

The second task was a narrative task which involved a series of six images following a fixed sequential story of two boys getting caught in the rain and looking for shelter at an abandoned and haunted house. While the story has a sequential structure with a clear beginning, middle and end, it has a surprising end when the two characters of the story find a dead man in one of the rooms. Based on the images, the learners were asked to work with each other to construct and narrate the story. The instructions informed the dyads that they had the freedom to decide how to approach the task in terms of who speaks first and what contribution each person makes. Previous research evidence (e.g., Robinson, 2001) suggests that narrative tasks are cognitively demanding as learners are required to interpret the sequential images, use specific vocabulary and structures required in the pictures and formulate utterances that describe the images and events that may not be familiar to the participants.

The third task was a decision-making task based on a set of six photos. The photos showed common activities that teenagers choose to do in their free time (e.g., playing computer games, going to the gym, etc.). The task instructions asked them to discuss "what were the young people in the pictures doing?". Based on the pictures that they had described, the dyads had to negotiate and choose "which activity is the most typical of the young people today". Similar to the instructions in other tasks, they were told they had five minutes to work together to complete the task, but they could structure it as they wanted. While the information in the task was familiar to the participants, we consider this task as cognitively demanding as it required discussion, negotiation and persuasion which add to the demands of performing a task in a second language. During the experiment, we did not find any interaction in which one member of the dyad was dominating the conversation, keeping quiet for an extended period (more than three seconds) or using time for extensive planning during the dialogues. However, we are aware that some planning inevitably took place between the turns.

It should also be mentioned that the three tasks, although very popular in many L2 classrooms, may not be considered as real-life dialogues since, for example, the speakers have to read the questions from a card to elicit information or work with a set of

[1] Because prompts have been reported in the literature to influence language performance (e.g., Leaper & Riazi, 2014; Shi et al., 2020), we acknowledge that the participants' output might have been partly influenced by the card prompts. However, the instances in which the participants read out the cards were few, and thus we believe these did not have a significant effect on the overall scores. However, as there were only a few instances in which the participants read out the cards, we believe these did not have a significant effect on the overall scores.

picture prompts to narrate a story. These tasks, however, are ecologically valid (Eckert, 2008) in the research context, and therefore, they will help improve the external validity of the findings of the study. The importance of ecological validity is highlighted by several researchers including Rogers and Cheung (2020), who argue that "a study can claim to have ecological validity if the experiment is similar to the context to which it aims to generalize" (p.3). The interest in external validity justifies the use of these dialogic tasks in the current study. Since we did not want to disrupt the usual teaching and learning process in these classes, we briefed the teachers about the purpose of the study and asked them to administer the three tasks in their classes while one of the researchers was present during the sessions for help and support with data collection (e.g., providing clarifications, helping with the voice recorders, and time keeping). The participants were paired in each class, and the dyads stayed together for all three task performances during the data collection. Digital audio recorders were used to record the performances, and at the end of the five minutes, a 'stop' sign was shown to the dyads.

6.4 Measures of analysis

The data analysed for the current study comprised 45 dialogues, 15 on each task, each dialogue lasting for five minutes or more. While most dyads stopped at the 5-minute stop sign shown by the researcher, some continued with their conversations until they felt they had completed the task. In our data analysis, we analysed the first five minutes of the dyads' performances. After the data were transcribed and word processed, in the form of a dialogue identifying each partner's contribution, they were segmented to AS-units (Foster et al., 2000) and clauses for each of the participants (30 participants in total) within the 15 dyads. The data were then subjected to a careful coding process for a number of task engagement and CALF measures. As can be seen below, the CALF measures and cognitive engagement represent each participant's performance in the dialogues almost independently of what the other participant had said or done. For measures of behavioural and social engagement, however, we considered the partners' reactions and contributions in order to determine the turns and instances of scaffolding. The choice of the measures is discussed below.

CALF measures. Following task-based research literature (Housen et al., 2012), we chose measures that are reported to be reliable indices representing the corresponding constructs of syntactic complexity, accuracy, lexical complexity and fluency. To measure syntactic complexity, ratio of subordination was selected because it is reported as one of the most reliable measures of complexity (Norris & Ortega, 2009; Skehan, 2009). For accuracy, the global measure of percentage of error-free clauses was used as it has been reported by several studies to be a reliable indicator of proficiency, particularly in relation to the development of proficiency (Tavakoli, 2018; Skehan, 2009). To ensure comparability, measures of accuracy and complexity are reported in percentages and ratios for each performance per task. To represent fluency, following de Jong et al. (2015) and de Jong (2018), frequency of silent pauses and total number of repair measures calculated per minute (i.e., repetition, hesitation and reformulation) were used to reflect breakdown and repair fluency. Pauses, defined as a period of silence longer than 0.25 a second, were measured by Goldwave (2018) software. Lexical diversity was represented by D measure (VocD) calculated by Text Inspector (2018). All the data were coded separately by each researcher and correlation coefficients of $r > .90$ were observed for different measures.

Measures of task engagement. Given our interest in the effects of task type on task engagement in this study, we examined the data in three dimensions for each participant: cognitive, behavioural and social engagement. For cognitive engagement, following Baralt et al. (2016), we used frequency of logical discourse markers as an indication of engagement in the task for causal and reasoning purposes. The number of logical discourse markers (e.g., because, so, therefore, etc.) was calculated by using Text Inspector's (2018) Metadiscourse facility for each participant in the dyads. For behavioural engagement, we used frequency of turns per participant per task to show the level of behavioural engagement (Dörnyei & Kormos, 2000). Frequency of turns was calculated manually on the transcripts considering the partner's behaviour and response. An agreement of 100% was found between first and second coder for this measure. To represent social engagement, we calculated frequency of instances of scaffolding per participant per task performance (Baralt et al., 2016). Scaffolding in L2 classroom is defined as "providing contextual supports for meaning through the use of simplified language, teacher modelling, visuals and graphics, cooperative learning and hands-on learning" (Ovando et al., 2003, p.345). For our analysis, scaffolding referred to learners' attempts to provide help to their partners (in terms of ideas, language or skills) in order to complete the task (e.g., helping each other pronounce a word, explaining a concept, completing a sentence, etc.). To code the data for scaffolding, we separately listened to the audio files and read the transcripts to identify instances of scaffolding. In the case of any disagreements, the instance of scaffolding was discussed until an agreement was achieved. Figures representing engagement measures refer to frequency of occurrence of each measure per performance for each individual. We believe using the raw data can better represent the degree of participants' engagement per task (e.g., the learner took six turns performing one task compared to three turns in another).

7. Analyses & results

To answer Research Question 1 (to what extent is learners' task performance, measured in CALF, affected by task type and proficiency level?), a repeated-measures multivariate analysis of variance (MANOVA) was first run to examine the effects of task type and proficiency level on the CALF measures. The MANOVA allowed us to explore the between-participant effect (effects of language proficiency), the within-participant effect (effects of task type), and the interaction between the two, if any. When significant results were obtained, effect sizes were calculated to show the power of the results. To interpret effect sizes for mixed ANOVAs, we followed Plonsky and Oswald's (2014) recommendations that for between group comparisons d values of .40 should be considered small, .70 medium, and 1.00 large. For within group comparisons, we will consider a *d* value of .60 as small, 1.00 as medium, and 1.40 as large (Plonsky & Oswald, 2014). Finally, given the repeated measures design of the study, we consider a Bonferroni corrected alpha level of 0.017 (0.05/3= 0.017). To examine the relationship between task performance and task engagement, we used Pearson product-moment correlations. The descriptive statistics for CALF measures across different task types and proficiency levels is provided on Table 2. A more detailed set of descriptive statistics for these measures is provided in Appendix 2.

Table 2. Descriptive statistics for CALF measures across task types and proficiency levels

Measures	Personal information task Mean (sd)	Narrative task Mean (sd)	Decision-making task Mean (sd)	Elementary Mean (sd)	Intermediate Mean (sd)	Advanced Mean (sd)
Frequency of pauses	11.60 (8.12)	20.70 (10.98)	19.30 (9.90)	17.20 (9.36)	22.43 (11.90)	11.96 (7.06)
Total number of repairs	6.60 (5.40)	8.10 (4.81)	7.86 (6.19)	8.30 (6.20)	7.33 (4.88)	6.93 (5.34)
Ratio of subordination	1.48 (.24)	1.70 (.36)	1.45 (.33)	1.50 (.34)	1.58 (.32)	1.53 (.34)
Percentage of error-free clauses	67.52 (12.10)	57.71 (14.53)	64.84 (13.18)	59.32 (14.91)	62.01 (11.67)	68.74 13.31)
Lexical diversity D	55.91 (10.56)	45.64 (9.47)	58.10 (16.49)	50.14 (13.68)	54.15 (15.14)	55.36 (11.52)

Note. N = 30.

As can be seen on Table 2, frequency of pauses varied widely across different task types and proficiency levels. The personal information task had the lowest number of pauses and the narrative task the highest, suggesting the narrative task elicited the least fluent performance. The same pattern was found for total number of repairs implying the participants made the fewest repairs in the personal information task and the most in the narrative task. In terms of complexity, the narrative task elicited the highest amount of subordination (1.70), while the other two tasks seemed similar in terms of the complexity of language produced (a mean of 1.48 for personal information and 1.45 for decision making task). The narrative task elicited the least accurate performance as well as the least lexical diversity suggesting the language used to perform this task was not very accurate or diverse. With respect to lexical diversity, it is possible to hypothesise that the limited range of lexis produced by the participants was influenced by the picture prompts as they did not encourage use of varied vocabulary items. Performances in the other two tasks were similar in terms of accuracy and lexical diversity.

As for the effects of proficiency, frequency of pauses does not show a clear progression pattern across levels of proficiency since learners at the intermediate level produced the highest number of pauses. For total repair, a linear pattern can be seen as the elementary learners produced the most and advanced learners produced the fewest number of repairs. The same pattern of progression can be seen across levels of proficiency for percentage of error free clauses and lexical diversity. The ratio of subordination did not change much as proficiency developed. These results suggest a clear effect of both task type and proficiency level on different CALF measures. In order to examine whether these differences are statistically meaningful, further inferential statistical analyses were conducted.

7.1 Repeated-measures MANOVA: Overall effects of the independent variables

Checking multivariate normality through a linear regression, the results of Mahalanobisi distances showed that our largest Mahal distance figure was 19.76 which is lower than the critical value of 22.46 suggested for a 6-dependent variable test, or 24.32 or a 7-dependent variable test (Pallant, 2014). This result implies that there are no multivariate outliers in the dependent variables. Levene's Test of Equality of Error Variances showed that the assumption of equality of variance has not been violated.

The multivariate analysis demonstrated a significant main effect for task type (Wilks' Lambda= .570; F= 4.99, p= .001; $\eta 2$= .245), a significant effect for proficiency level on performance (Wilks' Lambda= .389; F= 3.77, p= .001; $\eta 2$= .202), and an interaction effect (Wilks' Lambda= .774; F= 1.03, p= .427; $\eta 2$= .062). The results of the MANOVA indicating significant task and proficiency effects allowed us to continue the analysis with two-way mixed ANOVAs to examine the effects of the two independent variables on dependent variables of the study. In what follows, we present a summary of the results of the two-way mixed ANOVAs on different measures of performance.

7.2 Two-way mixed ANOVAs: Effects of task type and proficiency level

In running the two-way mixed ANOVAs, a Bonferroni post-hoc comparison and a corrected alpha level of 0.01 (0.05 divided by 5) was employed to evaluate the significant differences across task types and proficiency levels. Table 3 below shows the results of the mixed ANOVAs for CALF measures of performance.

Table 3. Results of mixed ANOVAs for effects of Task Type and Proficiency on CALF

Measures	Effects	F	P	Effect size
Frequency of silent pauses	TT	9.51	.001	.190
	LP	10.50	.001	.211
	INT	1.74	.148	.079
Total repair	TT	.621	.540	.015
	LP	.471	.626	.011
	INT	.433	.785	.021
Ratio of subordination	TT	5.38	.006	.117
	LP	.465	.630	.011
	INT	.612	.665	.029
% error-free clauses	TT	4.55	.01	.101
	LP	4.017	.01	.100
	INT	.413	.799	.020
Lexical diversity D	TT	8.67	.001	.176
	LP	1.46	.238	.035
	INT	1.44	.226	.067

Note. N= 30; Corrected alpha level of p < .017; TT= Task type;
LP = Level of proficiency; INT = Interaction effect

7.2.1 Effects of task type on CALF

The results suggested that task type affected performance in a number of ways. A significant effect of task type was found for frequency of silent pauses (F = 9.51, p =

.001; $\eta 2$= .190) with performance in the personal information task as the most fluent (fewest frequency of pauses). Although the highest number of pauses were seen in the narrative task, the differences between narrative and decision-making task were not statistically significant. Another significant difference emerging from the effects of task type was for ratio of subordination (F = 5.38, p = .006; $\eta 2$= .117) with learners producing the highest amount of subordination in the narrative task which was statistically different from the other two tasks. The effects of task type were also significant for accuracy and lexical diversity. For accuracy (F = 4.55, p = .01; $\eta 2$= .101), a significant difference was identified between the personal information task and other tasks, with most accurate performance elicited in the personal information task (i.e., when exchanging personal information). For lexical diversity (F = 8.67, p = .001; $\eta 2$= .176), the highest level of D was observed in the decision-making task in which the learners used a range of different words to complete the task. It is worth noting that all the effect sizes are small, according to Plonsky and Oswald's (2014) guidelines indicated above.

7.2.2 Effects of proficiency on CALF

As for the results of proficiency level, only two significant differences were observed: frequency of pauses and percentage of error-free clauses. For silent pauses (F = 10.50, p = .001; $\eta 2$ = .211), a significant difference was observed between the advanced level and the other two levels of proficiency, with the advanced level learners producing the fewest number of silent pauses. For percentage of error-free clauses (F = 4.17, p = .01; $\eta 2$ = .100), advanced learners produced statistically more accurate clauses than the other two levels. It is worth noting that effect sizes for these comparisons, ranging from .101 to .190, are considered small according to Plonsky and Oswald's (2014) criteria. No interaction effect was observed between task type and language proficiency.

To summarize the results, the analyses for Research Question 1 suggest that task type had an impact on frequency of pauses (fluency), ratio of subordination (syntactic complexity), percentage of error-free clauses (accuracy) and D (lexical complexity). The effects of proficiency level were observed for frequency of pauses and percentage of error-free clauses. These results will be discussed after the analysis of task engagement is presented.

7.3 Effects of task type and proficiency level on task engagement

To answer Research Question 2 (to what extent is learners' task engagement affected by task type and proficiency level?), two-way ANOVAs were used to examine the effects of task type and proficiency level on task engagement. The descriptive statistics for measures of task engagement is provided on Table 4 below.

Table 4. Task engagement across task type (means and standard deviations)

Measures of engagement	Personal information task Mean (sd)	Narrative task Mean (sd)	Decision-making task Mean (sd)	Elementary Mean (sd)	Intermediate Mean (sd)	Advanced Mean (sd)
Cognitive: logical DMs	3.83 (1.44)	3.33 (.88)	3.26 (1.20)	2.86 (1.27)	3.46 (1.00)	4.1 (1.02)

Behavioural: number of turns	11.53 (4.32)	8.10 (6.07)	11.46 (6.34)	9.93 (6.37)	9.23 (5.00)	11.93 (5.82)
Social: scaffolding	.20 (.48)	1.36 (1.80)	1.73 (2.39)	.76 (1.33)	1.00 (1.41)	1.53 (2.52)

Note. N = 30.

The descriptive statistics for different measures of engagement across task types (Table 4) suggests that the learners produced more logical discourse markers in the personal information task. While the personal information and decision-making tasks provided the learners with similar opportunities for turn taking, the learners took fewer turns in the narrative task. As for frequency of scaffolding, the personal information task elicited the fewest and the decision-making task the most instances of scaffolding. Examples of scaffolding across the tasks in a dyad from the advanced group (L25 and L26) are provided below.

Personal information task: no scaffolding

Narrative task:
L26: because these two peoples come <to a-> I don't know how to say that (pause)
L25: *stormy night? <stormy->*
L26: *stormy*

Decision-making task:
L26: young people is umm take care of an old man
L25: <take care-> *taking care*
L26: *taking care* of an old man and the old man is eating

The descriptive statistics for task engagement across different proficiency levels (Table 4) suggest that the use of logical discourse markers increased across proficiency level, as did the frequency of scaffolding. As for number of turns, while advanced learners took most turns, intermediate and elementary learners took fewer turns in their performance. It is necessary to note that for some of these measures the frequency of the engagement indices is very small (e.g., three discourse markers or one scaffolding on average). The two-way ANOVAs (Table 5) comparing task engagement across task types and proficiency levels showed three significant results.

Table 5. ANOVAs for effects of Task Type and Proficiency on task engagement

Measures	Effects	F	P	Effect size
Cognitive engagement: Logical discourse markers	TT	2.34	.102	.055
	LP	9.30	.001	.187
	INT	.526	.717	.025
Behavioural engagement: Number of turns	TT	3.77	.01	.085
	LP	1.92	.153	.045
	INT	1.46	.221	.067
	TT	6.48	.002	.138

Social engagement: Instances of scaffolding	LP	1.56	.216	.037
	INT	1.54	.198	.071

Note. N= 30; Corrected alpha level of p < .017; TT= Task type; LP = Level of proficiency; INT = Interaction effect

As for the results of task type, the ANOVAs showed two significant differences in the participants' engagement. For behavioural engagement across task types, the narrative task elicited the fewest number of turns while both the personal information and decision-making tasks elicited many more turns on average. Also, a significant difference was observed for scaffolding across task types, with the personal information task being significantly different from the other two tasks by providing the least opportunities for scaffolding. For the effects of proficiency level, a significant difference was observed for use of logical discourse markers. The post-hoc comparisons suggested that advanced learners were different from the other two groups in that they used logical discourse markers more frequently. To summarise the results, the advanced learners used the highest number of logical discourse markers, were engaged in scaffolding more frequently, and took more turns than the other two proficiency levels although these two comparisons did not reach a statistically significant level. The personal information task did not offer **rich** opportunities for scaffolding; the narrative task did not promote turn taking; but the decision-making task presented the participants with more opportunities for scaffolding and turn-taking.

7.4 Relationship between task performance and task engagement

Research question 3 looked at the relationship between aspects of task performance and task engagement (i.e., is there a relationship between task performance and task engagement?). To address this question, Pearson product-moment correlations were run between measures of task performance (i.e., the CALF measures used in the study) and measures of engagement (Cognitive, behavioural and social engagement). Following Plonsky and Oswald (2014), we interpret correlations (*r*) under .4 as weak, between .4 and .7 as moderate and above .7 as strong. Preliminary analyses were conducted to ensure there were no violations of the assumptions of normality, linearity, and homoscedasticity.

Table 6. Results of correlation analysis between task performance and task engagement

		Logical Discourse markers	Number of Turns	Scaffolding
Mean number of pauses	Pearson Correlation	-.055	-.340**	.047
	Sig. (2-tailed)	.609	.001	.660
Total repair	Pearson Correlation	.069	-.158	.069
	Sig. (2-tailed)	.520	.137	.519
Ratio of subordination	Pearson Correlation	.227*	-.726**	-.241*
	Sig. (2-tailed)	.032	.000	.022
% of error-free clauses	Pearson Correlation	.223*	.375**	.045
	Sig. (2-tailed)	.035	.000	.671
VOCD	Pearson Correlation	.371**	.244*	.065
	Sig. (2-tailed)	.000	.021	.544

N = 90
** Correlation is significant at the 0.01 level (2-tailed).
* Correlation is significant at the 0.05 level (2-tailed).

As can be seen in Table 6, the results demonstrate a range of significant correlations between engagement and CALF measures. Cognitive engagement (i.e., number of logical markers) correlated significantly with syntactic complexity (r = .227, p < .03), accuracy (r = .223, p < .04), and lexical diversity (r = .371, p < .001). These correlations were all positive, although weak (all under r = .4), suggesting a higher level of cognitive engagement was associated with more accurate and complex language. Behavioural engagement, operationalised in terms of number of turns, correlated with all measures of fluency (frequency of pauses r = -.340, p < .001), accuracy (percentage of error-free clauses r = .375, p < .001), syntactic complexity (ratio of subordination r = -.726, p < .001) and lexical diversity (D, r = .244, p < .02). These correlations are positive for accuracy and lexical diversity but negative for fluency and syntactic complexity, implying those who took more turns produced less fluent and syntactically complex language, but their performance was more accurate and lexically diverse. Finally, social engagement (i.e., frequency of scaffolding) negatively correlated with ratio of subordination (r = -.241, p < .02), suggesting those who offered less scaffolding produced more syntactically complex language. It is worth mentioning that total number of repairs did not correlate with any of the engagement measures. To summarise the results, the correlational analysis suggested that there was a relationship between different aspects of performance and task engagement by the participants. Specifically, the evidence suggests that behavioural engagement was correlated with several aspects of performance.

8. Discussion

The current study set out to investigate the effects of task type and proficiency level on task performance and engagement of 15 L2 learner dyads of English at a Mexican university. Motivated by previous research (Foster & Skehan, 1996), three task types were used to examine learners' dialogic performance and task engagement. We examined effects of task type across three proficiency levels: elementary, intermediate and advanced. Measures of CALF and social, behavioural and cognitive engagement were considered as dependent variables in the study. The findings of the study are discussed in relation to its three research questions.

8.1 Task performance

RQ1 examined the effects of task type and language proficiency on learner performance in terms of CALF measures. The results of the analyses suggested that task type had an impact on the fluency, syntactic complexity, accuracy and lexical diversity of learner performance. Specifically, the results showed that the performance in the personal information task was the most accurate and fluent; the performance in the narrative task was the most syntactically diverse but the least fluent and accurate; and the performance in the decision-making task was the most lexically diverse. These results are in line with Foster and Skehan (1996), in that the personal information task promoted fluent and accurate language; the narrative task elicited the most syntactically complex and least fluent performance; and the decision-making task encouraged reasonable levels of accuracy and complexity.

As for the personal information task, the results showed that the participants produced language of high fluency and accuracy. The high levels of learners' fluency here may suggest a high degree of automaticity in their performance in this task. The automatic use of language, represented by the fewer number of pauses in the personal information task may imply that the learners were familiar with the language structures and content of the task, had had enough opportunities to practice the language functions or a combination of both. The high level of accuracy in this task can also be interpreted in relation to the familiarity with the personal information they were providing. It is also possible to argue that the use of flash cards in the personal information task in our study allowed the participants to formulate the questions with changes of words or structure and to borrow linguistic units from them. This might have supported learner performance linguistically and eased off the pressure of producing questions and formulating the responses during the interactions. Future research will need to look into such effects in a more systematic manner.

Based on these findings, it can also be argued that the narrative task seemed more demanding as student performance was less fluent and accurate than that elicited by the other two tasks. In those two other tasks, the participants were working with information that was familiar to them, whereas in the narrative task they worked with a story that was new to them and included some unknown information and an unanticipated punchline. We believe these factors may have made the task more challenging to perform. In addition, the controlled nature of the narrative task in terms of what information should be provided and what linguistic items and structures were to be used to narrate the story accurately might have affected their performance. Our finding for accuracy of performance across the three tasks is in line with Gilabert (2007), who found that the narrative task elicited the least accurate performance.

The results indicated that the decision-making task, compared to the narrative, elicited language of slightly higher fluency and accuracy. While the narrative task elicited the least fluent language, the two tasks were not statistically different for pause and repair measures. This is intriguing as we had expected the familiar information in the decision-making task (e.g., activities young people choose to do in their free time) to help promote fluency and/or accuracy. The unanticipated finding warrants further research to examine whether the choice of information provided in the task, or the requirement of negotiation and persuasion has affected learner fluency in this task. In our study, we also measured lexical complexity and the results showed that the decision-making task elicited the most lexically diverse language. We argue that the different requirements of our decision-making task (e.g., six different activities and a need for describing, discussing, negotiating and persuading) have encouraged a diverse use of lexical items. Overall, it is interesting to see the same results as those reported by Foster and Skehan (1996) with learners at three different levels of proficiency, implying that task type influences performance systematically across proficiency levels.

As for the effects of proficiency level on performance, the results suggested that proficiency level had an impact on fluency and accuracy. For both, the advanced learners were different from those at the lower levels. Despite the variance in scores for syntactic complexity and lexical diversity across the three levels, the differences did not reach a statistically significant level. This suggests that what in effect distinguishes participants at these proficiency levels may not be the syntactic or lexical complexity of their language; rather, fluency and accuracy of their performance appears to be the key to differentiating the learners. These results also suggest that complexity, whether syntactic or lexical, is a characteristic of language to be encouraged by task design. Our results showed that total number of repairs learners made during task performance was not

influenced by either task type or proficiency level. This corroborates with the findings of Tavakoli, Nakatsuhara & Hunter (2020), who reported no systematic effects of proficiency or task type on repair measures. Finally, no significant interactions were found between task type and proficiency level when we ran the two-way ANOVAs, suggesting the effects of task type were consistent across the three proficiency levels.

Taken together, the above results have some implications for task performance in EFL classrooms. Before discussing the implications, however, it is necessary to note that the results are based on a small-scale study and therefore they should be interpreted cautiously. We believe that the evidence of the effects of task type on linguistic performance reported here, and supported by other studies (Gilabert, 2007; Tavakoli, 2009, 2011, 2018), could be used to design dialogic tasks to promote the development of language learning and interlanguage. As shown in our results, task type tended to consistently affect learners' oral performance in a number of dimensions that can be used for teaching and learning purposes. To promote learner's development of syntactic and lexical diversity, for example, task design could be employed to help provide learners with rich opportunities for such development.

8.2 Task engagement

RQ2 asked whether task engagement was affected by task type and proficiency level. The results of the ANOVAs suggested that task type had an impact on social engagement measured through frequency of scaffolding in task per participant, with the personal information task eliciting the lowest amount of scaffolding. It is possible to argue that in this task, familiarity with both the information and task type makes the task less demanding, compared to the other two tasks, and therefore the learners may not feel the need to provide scaffolding, that is, to help one another with ideas, language forms or skills. As for behavioural engagement, the narrative task elicited the fewest number of turns suggesting the learners had a low level of engagement in this task. This may imply that constructing a narrative did not require the learners to interact dynamically with one another through taking turns. Given the evidence from previous research about the cognitive demands of narratives (Foster & Skehan, 1996; Robinson, 2001), it is also possible to claim that the high demands of the task and the controlled nature of the output did not provide room for discussion and negotiation and hence fewer opportunities for turn taking. Cognitive engagement, operationalised as the use of logical discourse markers, was not affected by task type. The higher levels of engagement in the decision-making task, compared to the other two tasks, may have been promoted by the freedom they had in the decision-making task to choose what information to talk about or the structures to use to express their intended meaning. Dao (2020) also found support for the claim that the decision-making task promoted social engagement, perhaps because the learners were expected to discuss the topic and reach a decision. Based on these findings, we argue that the decision-making task provides a rich opportunity for behavioural and social engagement. Previous research (Butler, 2017; Lambert, Philp & Nakamura, 2017; Phung, 2017) has also suggested that giving control over the choice of the content and considering students' preferences are important aspects of task design that can promote engagement. The findings of our study clearly suggest that engagement is closely related to task design in that the different task types provided different opportunities for engagement. The decision-making task, which involved comparison, negotiation, discussion and decision making, seemed to offer a rich opportunity for engagement by encouraging working collaboratively and supporting one another during task performance.

Looking at the effects of proficiency on task engagement, there was a significant difference between the three levels of proficiency in terms of their cognitive engagement, where advanced learners used more logical discourse markers than the other two levels. This higher use of discourse markers may in fact reflect their ability to shape their speech which would also help promote the listeners' comprehension. This finding is corroborated by previous research in terms of L2 speaker linguistic behaviour across proficiency levels. Kang and Wang (2014), for example, reported that higher proficiency speakers were more socially engaged in supporting one another (e.g., back channelling and prompting sentences and phrases). Kang and Wang's study also demonstrated that higher proficiency speakers used more discourse markers and took more turns, indicating that the speakers were more engaged in the tasks. Our analysis also indicated that the advanced learners produced not only more scaffolding but also more back channelling and support to their partners. The support included correcting errors, providing alternative lexical choices and stepping in to complete an incomplete sentence or clause. Previous research in this area has considered these moves as "an improved degree of engagement" (Kang & Wang, 2014, p. 48). Finally, it is necessary to note that our data did not demonstrate high levels of cognitive or social engagement as the figures representing means of these variables were rather small (i.e., ranging from 0.2 to 1.73 for social and 2.86 to 4.1 for cognitive engagement per person per task). For this reason, the findings should be interpreted cautiously.

8.3 Relationship between task performance and task engagement

The results of our analysis above suggested that behavioural and cognitive engagement were positively related to the level of accuracy and lexical complexity of the participants' performance. This is to say, the participants who took more turns and produced more logical discourse markers were more accurate and used more diverse lexical items. This suggests that having a better command of lexical variety and linguistic accuracy may encourage both cognitive and behavioural engagement in a task. It may be the case, therefore, that feeling more confident about one's linguistic abilities in terms of accuracy and lexical variety may encourage learners to use more logical discourse markers and turns and thus increase their levels of cognitive and behavioural engagement. The negative relationship between the number of turns and measures of speed fluency and subordination, on the other hand, suggests that learners who take more turns are likely to be slower and produce language of lower syntactic complexity. The largest correlation observed in our analysis (r = -.726) between behavioural engagement and syntactic complexity implies that taking more turns is strongly linked to producing language of low subordination. It can thus be suggested that in the process of producing several turns required by the characteristics of the task, students' oral abilities may be directed towards constructing more accurate and lexically varied utterances at the expense of other dimensions. Similarly, the negative relationship between social engagement and syntactic complexity implies that those students who provide more scaffolding are likely to produce less syntactically complex language. A possible explanation for this might be that when scaffolding is formulated individually from one learner and provided to a peer, it may not necessitate complex language because, as stated by Skehan (2001), more complex language is produced by learners when scaffolding is provided in collaborative negotiations.

The above findings draw our attention to the importance of evaluating task performance and engagement together as it can better explain a learner's language and communication skills. In SLA research, it is believed that social and behavioural engagement help generate rich opportunities for effective learning processes and

outcomes; the findings of the current study help us understand that engagement might also impose restrictions on task performance.

9 Conclusions

The present study took a novel approach to examining effects of task type and proficiency on both task performance and engagement as it looked at learners' linguistic behaviour across three different levels of proficiency and three task types in a classroom context. We investigated both engagement and performance as we assume these are both important aspects of the L2 acquisition process with potentially significant implications for task design.

The findings of our study suggested that different task types provide varying opportunities for linguistic performance and task engagement. Some of these opportunities may lead to more effective L2 acquisition in terms of linguistic forms and task engagement. Performing a personal information task based on familiar information, for example, would enhance fluency and accuracy of performance but not complexity. This task would also provide rich opportunities for cognitive and behavioural engagement. Performing a narrative task would encourage syntactic complexity but perhaps at the cost of accuracy and fluency. Compared to the other two tasks, the narrative task would not provide rich opportunities for engagement. The decision-making task would promote the use of diverse lexical items and considerable levels of task engagement, but it may not involve use of syntactically complex language. Performing this task and fulfilling the range of its requirement (e.g., description, negotiation and persuasion), however, may have had a negative impact on accuracy and fluency. Our results also implied that higher proficiency learners were more accurate and fluent, but perhaps more importantly they were more engaged in task performance. In terms of task design, we are aware that in each task type, the topic might have influenced our findings. As such, it is necessary for future research to control for the interaction between task type and task topic.

Examining task engagement and its relationship to task performance should be considered an important contribution this paper has made to the field of TBLT research and pedagogy. The findings of the study have helped shed light on not only how learners engage in performance of different task types but also how learning opportunities vary as a result of engaging in a task. For example, the results have suggested that learners with higher accuracy and lexical diversity are more likely to be engaged in tasks cognitively and behaviourally. The findings have also shown that although behavioural engagement (e.g., number of turns a participant takes) in a dialogic task is known to provide learners with potentially rich opportunities for interaction and acquisition, learners who take more turns would be restricted in terms of producing syntactically complex and fluent language.

We believe that the above findings have some implications for language teaching and assessment. They suggest language teachers should consider engagement and performance hand-in-hand in their teaching and testing practices. The relationship between engagement and performance should be considered in choosing tasks for pedagogic purposes. For example, if the focus of a task is to enhance learners speed fluency or syntactic complexity, taking too many turns would slow learners down and result in with low syntactic complexity. This kind of evidence would help teachers make decisions concerning lesson planning and the development of lesson materials. The findings are also important in language assessment contexts as lack of engagement with a task can negatively influence raters' judgements and score allocations, raising issues of

test reliability and validity. Thus, assessing learning requires a number of aspects such as the relationship between task performance and engagement. The current study focused on a correlation analysis, and as such, it cannot make a claim about the cause-and-effect relationship between the two. The findings of the correlations, however, clearly indicate the need for future research to examine the potential for a causal relationship.

Finally, we are aware that our data is limited as it comes from a small sample size and is not supported by classroom observations or retrospective data. We recommend that future research should address these limitations. In particular, further research is needed to help us understand "how to engage all learners" (Philp & Duchesne, 2016, p.52) during task performance, and how to design tasks that provide better opportunities for learning that emerges from task engagement. Ultimately, such research could aid teachers in using task design to promote learners' L2 production and engagement.

References

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

Baralt, M., Gurzynski-Weiss, L., & Kim, Y. (2016). Engagement with language: How examining learners' affective and social engagement explains successful learner-generated attention to form. In M. Sato & S. Ballinger (Eds.), *Peer interaction and second language learning. Pedagogical potential and research agenda* (pp. 209–240). Amsterdam: John Benjamins.

Butler, Y. G. (2017). Motivational elements of digital instructional games: A study of young L2 learners' game designs. *Language Teaching Research, 21*(6), 735-750.

Bygate, M. (1999). Task as context for the framing, re-framing and un-framing of language. *System*, *27*(1), 33-48.

Bygate, M., & Samuda, V. (2009). Creating pressure in task pedagogy: The joint roles of field, purpose, and engagement within the interaction approach. In Mackey, A., Polio, C. (Eds.), *Multiple perspectives on interaction: Second language research in honor of Susan M. Gass* (pp. 90–116). New York: Routledge.

Cambridge (2018). *First Certificate of English.* Retrieved from: https://www.cambridgeenglish.org/exams-and-tests/first/results/

Dao, P. (2020). Effects of task goal orientation on learner engagement in task performance. *International Review of Applied Linguistics in Language Teaching*,*58*(3): 323-349.

de Jong, N.H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, *15*(3), 237-254. https://doi.org/10.1080/15434303.2018.1477780

de Jong, N.H., Groenhout, R., Schoonen, R., & Hulstijn, J.H (2015). Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics. 36*(2), 223–243. https://doi.org/10.1017/S0142716413000210.

Dörnyei, Z., & Kormos, J. (2000). The role of individual and social variables in oral task performance. *Language Teaching Research, 4*(3), 275–300.

Eckerth, J. (2009). Negotiated interaction in the L2 classroom. *Language Teaching, 42*(1), 109-130.

Ellis, R. (2005). Measuring Implicit and Explicit Knowledge of a Second Language: A Psychometric Study. *Studies in Second Language Acquisition, 27*(2), 141-172.

Faez, F. & Tavakoli, P. (2019). *Task-based Language Teaching*. Alexandria, VG: TESOL Press.

Foster, P. (1996). Doing the task better: How planning time influences students'

performance. In J. Willis & D. Willis (Eds.), *Challenge and change in language teaching* (pp. 126-135). Oxford, UK: Heinemann.

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.

Fredricks, J. A., & McColskey, W. (2012). The measurement of student engagement: A comparative analysis of various methods and student self-report instruments. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student engagement* (pp. 763–782). Boston, MA: Springer US.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of Educational Research, 74*(1), 59–109. https://doi.org/10.3102/00346543074001059

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning, 66*(2), 419-447.

Galaczi, E. (2008). Peer-peer interaction in a speaking test: The case of the *First Certificate in English* Examination. *Language Assessment Quarterly, 5(*2), 89-119. https://doi.org/10.1080/15434300801934702

Galaczi, E. & Taylor, L. (2018). Interactional competence: Conceptualizations, operationalizations and outstanding questions. *Language Assessment Quarterly, 15*(3), 219–236.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. International Review of Applied Linguistics in Language Teaching, *45*(3), 215-240. https://doi.org/10.1515/iral.2007.010

Helme, S., & Clarke, D. (2001). Identifying cognitive engagement in the mathematics classroom. *Mathematics Education Research Journal, 13,* 133–153.

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

Kang, K., & Wang, L. (2014). Impact of different task types on candidates' speaking performances and interactive features that distinguish between CEFR levels. *Cambridge Research Notes, 57,* 40-49.

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

Kim, Y., Nam, J., & Lee, S.-Y. (2016). Correlation of proficiency with complexity, accuracy, and fluency in spoken and written production: Evidence from L2 Korean. *Journal of the National Council of Less Commonly Taught Languages, 19,* 147-181. Retrieved from http://www.ncolctl.org/files/jncolctl-vol-19/Correlation-of-Proficiency-with-Complexity.pdf

Kormos, J. (2014). *Speech production and Second Language Acquisition.* Computer science.

Kuiken, F., & Vedder, I. (2007). Task complexity and measures of linguistic performance in L2 writing. *International Review of Applied Linguistics in Language Teaching, 45*(3), 261-284. https://doi.org/10.1515/iral.2007.012

Lambert, C., Philp, J., & Nakamura, S. (2017). Learner-generated content and engagement in second language task performance. *Language Teaching Research, 6,* 665–680.

Leaper D.A., & Riazi M. (2014). The influence of prompt on group oral tests. *Language Testing*, 31(2), 177-204.

Mackey, A. (1999). Input, interaction, and second language development. *Studies in second language acquisition, 21*(4), 557-587.

Michel, M.C., Kuiken, F. & Vedder, I. (2007). Effects of task complexity and task condition on Dutch L2. *International Review of Applied Linguistics, 45*(3), 241-259.

Michel, M.C., Revesz, A., Shi, D., & Li, Y. (2019). The effects of task demands on linguistic complexity and accuracy across task types and L1/L2 speakers. In Wen and Ahmadian (eds.), *Researching L2 task performance and pedagogy* (pp. 133-151). London: John Benjamins.

Newmann, F. (1991). Student engagement in academic work: Expanding the perspective on secondary school effectiveness. In J. R. Bliss & W. A. Firestone (Eds.), *Rethinking effective schools: Research and practice* (pp. 58–76). Englewood Cliffs, NJ: Prentice-Hall.

Ortega, L. (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition, 21*(1), 109-148.

Ovando, C., Collier, V., & Combs, M. (2003). *Bilingual and ESL classrooms: Teaching Multicultural Contexts* (3rd ed.). Boston: McGraw-Hill.

Pallant J. (2014). *SPSS survival manual, a step by step guide to data analysis using SPSS for windows.* McGraw Hill.

Philp, J. & Duchesne, S. (2016). Exploring engagement in tasks in the language classroom. *Annual Review of Applied Linguistics*, *36*(1), 50-72.

Phung, L. (2017). Task preference, affective response, and engagement in L2 use in a US university context. *Language Teaching Research, 21*(6): 751-766.

Plonsky, L., & Oswald, F.L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-91.

Reschly, A. L., & Christenson, S.L. (2012). Jingle, jangle, and conceptual haziness: Evolution and future directions of the engagement construct. In S. L. Christenson, A. L. Reschly, & C. Wylie (Eds.), *Handbook of research on student*

engagement (pp. 3–19). Springer Science + Business Media.

Revesz, A. J., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviours: A mixed methods study. *Studies in Second Language Acquisition, 41*(3), pp. 605-631.

Robinson, P. (2003). The Cognition Hypothesis of adult, task-based language learning. *Second Language Studies, 21*(2), 45-107.

Rogers, J., & Cheung, A. (2020). Does it matter when you review?: Input spacing, ecological validity, and the learning of L2 vocabulary. *Studies in Second Language Acquisition*, 1-19.

Shi, B., Huang L., & Lu, X. (2020). Effect of prompt type on test-takers' writing performance and writing strategy use in the continuation task. *Language Testing, 37*(3), 361-388.

Skehan, P. (2009). Modeling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(3), 510-532.

Skehan, P. (2001). Comprehension and production strategies in language teaching. In C. N. Candlin & N. Mercer (Eds.), English language teaching in its social context: A reader (pp. 75–89). London: New York: Routledge. https://doi.org/10.2307/3588440.

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

Skinner, E. A., Kindermann, T. A., & Furrer, C. (2009). A motivational perspective on engagement and disaffection: Conceptualization and assessment of children's behavioral and emotional participation in academic activities in the classroom. *Educational and Psychological Measurement, 69*(3), 493–525.

Tavakoli, P. & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning*. 58(2): 439-473.

Tavakoli, P. (2009). Assessing L2 task performance: Understanding the effects of task design. *System, 37*(3), 482-495.

Tavakoli, P. (2011). Pausing patterns: Differences between L2 learners and native speakers, *ELT Journal. 65*(1): 71-79.

Tavakoli, P. (2016). Speech fluency in monologic and dialogic task performance. *International Review of Applied Linguistics* (IRAL): Special Issue on fluency. 54(2): 131-150.

Tavakoli, P. (2018). L2 Development in an intensive Study Abroad EAP context. *System, 72*(1), 62-74.

Tavakoli, P. Nakatsuhara, F. & Hunter, A-M. (2020) Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal. 104*(1): 169-191.

Wigglesworth, G. (1997). An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, *14*(1), 85-106.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford: Oxford University Press.

Yazzie-Mintz, E. (2009). *Engaging the voices of students: A report on the 2007 & 2008 high school survey of student engagement.* Bloomington, IN: Center for Evaluation and Education Policy.

Yuan, F., & Ellis, R. (2003). The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 oral production. *Applied Linguistics*, *24*(2), 1-27.

Appendix 1. Cambridge score ranges for assigning the learners' proficiency levels

Grammar score ranges	Cambridge scale scores	Proficiency level
0-7	120	elementary
8-11	140	intermediate
12-18	160	advance

Appendix 2. Detailed descriptive statistics for all the measures

		Frequency of pauses		Total repair		Ratio of subordination		Percentage of error-free clauses		Voc-D		Logical Discourse markers		Number of Turns		Scaffolding	
Task type	Proficiency level	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Personal information	Elementary	14.80	11.60	8.20	8.10	1.36	.30	60.55	14.62	54.76	12.68	3.00	1.69	13.60	4.4	.40	.69
	Intermediate	12.90	4.20	6.10	3.54	1.50	.14	68.80	8.39	54.52	11.74	3.80	.91	10.10	4.62	.10	.31
	Advanced	7.10	4.95	5.50	3.37	1.57	.23	73.19	9.78	58.46	7.04	4.70	1.15	10.90	3.41	.10	.31
	Total	11.60	8.12	6.60	5.40	1.48	.24	67.52	12.10	55.91	10.56	3.83	1.44	11.53	4.32	.20	.48
Narrative	Elementary	17.50	6.91	8.00	5.14	1.71	.36	54.46	18.06	46.50	9.61	3.00	.81	6.70	6.36	1.00	1.63
	Intermediate	29.90	11.82	9.20	5.82	1.76	.28	56.27	9.61	46.13	12.14	3.30	1.05	6.50	2.01	1.60	1.89
	Advanced	14.70	7.45	7.10	3.44	1.63	.46	62.39	14.96	44.30	6.81	3.70	.67	11.10	7.69	1.50	2.01
	Total	20.70	10.98	8.10	4.81	1.70	.36	57.71	14.53	45.64	9.47	3.33	.88	8.10	6.07	1.36	1.80
Decision making	Elementary	19.30	9.42	8.70	5.63	1.44	.29	62.94	11.57	49.17	17.61	2.60	1.26	9.50	6.65	.90	1.59
	Intermediate	24.50	11.46	6.70	4.94	1.50	.44	60.94	13.68	61.79	17.80	3.30	1.05	11.10	6.50	1.30	1.15
	Advanced	14.10	6.41	8.20	8.05	1.40	.27	70.64	13.44	63.33	10.71	3.90	.99	13.80	5.71	3.00	3.43
	Total	19.30	9.99	7.86	6.19	1.45	.33	64.84	13.18	58.10	16.49	3.26	1.20	11.46	6.34	1.73	2.39
Total	Elementary	17.20	9.36	8.30	6.20	1.50	.34	59.32	14.91	50.14	13.68	2.86	1.27	9.93	6.37	.76	1.35
	Intermediate	22.43	11.90	7.33	4.88	1.58	.32	62.01	11.67	54.15	15.14	3.46	1.00	9.23	5.00	1.00	1.41
	Advanced	11.96	7.06	6.93	5.34	1.53	.34	68.74	13.31	55.36	11.52	4.10	1.02	11.93	5.82	1.53	2.52
	Total	17.20	10.46	7.52	5.47	1.54	.33	63.35	13.80	53.22	13.56	3.47	1.21	10.36	5.81	1.10	1.85