“I think that’s what he’s doing”: Effects of intentional reasoning on second language (L2) speech performance

Abstract

This study advances our understanding of the effects of task design on task complexity and second language (L2) performance. The research reported here focused on examining the impact of degree of intentional reasoning, operationalised at two levels of task content and task instructions, on language performance and perceptions of task difficulty. Using a mixed-methods approach, the study drew on quantitative and qualitative data collected from 20 Jordanian L2 learners performing video-based oral narratives and completing retrospective questionnaires. The results suggest that intentional reasoning has a noticeable effect in generating more syntactically complex and accurate language, and also influences perceptions of task difficulty. However, a higher intentional reasoning demand is associated with less lexical diversity and inconsistent patterns of fluency. An important finding of the study is that the link between the cognitive demands and the language used to convey intentional reasoning should be carefully considered when selecting analytic measures of complexity and accuracy. The implications of these findings for two most widely-used models of task complexity, i.e. Cognition Hypothesis (Robinson, 2007) and Limited Capacity model (Skehan, 2009) are discussed.

Key words: task complexity, intentional reasoning, second language, speech performance, task-based language teaching

1. Introduction

A leading line in research in task-based language teaching (TBLT) has been how task design influences second language (L2) performance and acquisition, offering a new perspective to understanding the role of task in producing and acquiring L2. For example, TBLT has provided ample research evidence that suggests manipulating task design, e.g. task structure, influences cognitive processes involved in language production and promotes accuracy and fluency of L2 performance (Foster & Skehan, 1996; Author 2 & collaborator, 2008; Wigglesworth, 2001).

TBLT has also been successful in persuading L2 pedagogy that tasks are important teaching devices that activate naturalistic L2 acquisitional processes and facilitate development of learner interlanguage (Skehan, 1998; Bygate et al., 2001; Bygate, 2015). For similar reasons, syllabus writers find TBLT research findings beneficial in designing communicative materials and developing effective syllabi for L2 teaching purposes. However, it is essential to understand how task design entails differences in task complexity and, particularly, task difficulty. Thus the proposal that task difficulty can inform L2 syllabi (Nunan, 1988; Ellis, 2000) and the pledge that TBLT research can help develop an index of task difficulty have been appealing to L2 practitioners for a long time.

In this regard, what L2 educators have hoped to see is an analytic framework that allows for a systematic evaluation and analysis of a) task complexity, i.e. the extent to which task demands deplete learners' cognitive resources and information processing capacity, and b) task difficulty, i.e. the extent to which learners find a task demanding and challenging to perform. Such a framework should then serve to help teachers and learners understand how task complexity and/or difficulty could help or hinder language development.

Therefore, investigating the effect of task complexity on learners' performance and acquisition has become an increasingly important research agenda for TBLT research (Ellis, 2000; Jackson & Suethanapornkul, 2013; Robinson, 2001, 2011; Skehan, 1998, 2014), as the findings of this research can shed light on different aspects of the L2 acquisition process and provide a much-needed link to current pedagogy. Although there is substantial interest in the study of variables that influence task complexity, not much attention has been dedicated to the study of intentional reasoning, or to developing an index of task difficulty (Jackson & Suethanapornkul, 2013; Ellis, 2012; Skehan, 2014).

The abstract nature of many of the variables contributing to task complexity and task difficulty, and the methodological complexities involved in examining and evaluating them are some of the challenges researchers in this area face. In this article we set out to examine one such variable, intentional reasoning (IR), which is reported to have a significant influence on task complexity (Jackson & Suethanapornkul, 2013; Robinson, 2007), but has not yet been researched systematically as a variable in its own right, nor for its relation to how a task is perceived as difficult by the task-taker. We aim to help fill the existing gap in TBLT research through the findings from our recent study by a) examining the contribution of IR to task complexity, b) introducing a more systematic approach to defining, evaluating and operationalising IR, and c) investigating the impact of IR on perceptions of task difficulty. We contextualise the study first by examining the crucial literature relating to task complexity, task difficulty, and IR.

2. Literature Review

2.1 Models of Task complexity

As noted earlier, in L2 research, tasks have commonly been seen as a research instrument that could potentially allow researchers to examine different aspects of L2 acquisition and production processes. In particular, tasks would allow researchers to examine increasing degrees of task complexity, and how such complexity could prompt or hinder spoken language during the task performance itself. Hence, a need evolved for a framework to analyse, evaluate and operationalise task design systematically. With this need in mind, researchers have long been engaged in defining task complexity as a key component of task design, and identifying the features, variables and conditions that contribute to it (Robinson, 2001, 2015; Skehan, 2001, 2014). Early conceptualisations, e.g. in Prabhu's (1987) study, saw that complexity was contingent on both cognitive and linguistic demands which a task imposed on learners during task performance, but without specifying in detail what these demands were, or how they related to each other. Skehan (2001) defined task complexity as the extent to which a task depletes the learner's attention. Robinson takes this attentional focus further, articulating complexity as the effect on "conceptualisation, attention, memory, and reasoning processes during task performance" (Robinson, 2015, p.95).

Building up on earlier research testing these constructs, Skehan and Robinson have developed two similar frameworks for evaluating, analysing and examining task complexity, although each draws on different underpinning assumptions. In addition to analysing task complexity, the two frameworks are interested in the effects task complexity has on different aspects of language performance, divided to categories of form (complexity and accuracy) versus meaning (fluency).

Robinson's (2001, 2007, 2015) framework, known as the Cognition Hypothesis, distinguishes between three factors: 1) task complexity (task-dependent cognitive demands), 2) task conditions (task-interactive demands), and 3) task difficulty (learner-dependent factors). The framework is motivated by a multiple-resource perspective on attentional capacities; this framework assumes that the human brain functions as a multiple-resource attentional system in which exhausting attention in one pool does not restrict the attentional resources available in other pools; however, variables relating to task complexity affect attentional resources differently, and may restrict performance accordingly.

Robinson divides variables along two dimensions. First are ‘resource-directing’ variables, i.e. those which help attentional focus through the amount of contextual support and reasoning demands. Second are ‘resource-dispersing’ variables, which are affected, for example, by task structure, planning time or amount of prior knowledge. Robinson (2007) argues that increasing task complexity along the resource-directing dimension, e.g. increasing the need for intentional reasoning (IR), directs learners’ attention to a range of linguistic and functional requirements, leading learners’ attention to improve both aspects of the linguistic form, i.e. complexity and accuracy of performance. By contrast, a task involving ‘resource-dispersing’ variables, e.g., less planning time, increases the demands on performance causing depletion of attentional resources; this results in less attention being available for processing, and therefore may negatively affect speech performance across the three aspects of performance, i.e. complexity (syntactic and/or lexical), accuracy and fluency (CAF).

Skehan’s Limited Capacity model (1998, 2009, 2014) distinguishes between three broad factors that regulate task complexity, but does not distinguish task difficulty as a component part of complexity, like Robinson’s; rather, Skehan sees difficulty in relation to how demanding the task is overall. Skehan’s factors of task complexity are: 1) code complexity (the language required, e.g. linguistic complexity), 2) cognitive complexity (the thinking required, e.g. amount, organisation and familiarity of information), and 3) communicative stress (the performance conditions for task performance). There is some degree of overlap in these models, e.g. code complexity relates to a small extent to Robinson’s earlier mention of conceptualisation requirements; communicative stress may be argued to relate to Robinson's notion of task difficulty.

The main difference in Skehan's model compared to Robinson's is that it assumes the human brain operates on a limited-capacity attentional system in which attentional resources are limited. As a result, when performing a demanding task, competition for allocating attention can be anticipated. This is to say, increasing demands of the task in any of the three dimensions of complexity, accuracy, fluency, would result in a trade-off between aspects of form, i.e. accuracy and complexity. Unlike Robinson's Cognition Hypothesis, Skehan's model (2014) does not distinguish between resource-directing and resource-depleting variables, and argues that the link between task characteristics and language performance should be explored "on a case-by-case basis" (p: 7).

Further to the debates arising from the different conceptualisations of complexity as noted in the two models above, it is clear that operationalising complexity is similarly hard to pin down. Pallotti (2009, 2014), arguing for the multifaceted nature of complexity in applied linguistics, contends that it is important to distinguish between complexity "directly arising from the number of linguistic elements and their interrelationships", and complexity as the "cognitive cost" arising from performing the task (Pallotti, 2014: 30). In addition, even if using existing definitions to operationalise complexity, it can be unclear whether complexity is viewed similarly from the perspective of a task-taker or an expert task-designer. Highlighting the need for validation studies that systematically examine the varying operationalisations of task complexity in TBLT research, Révész, Michel and Gilabert (2016) evaluated the manipulation of task complexity through use of dual-task methodology, self-ratings, and expert judgements. The findings of their study suggested that manipulation of task complexity results in different levels of cognitive complexity, and has an impact on both learner perceptions and expert judgements.

Different conceptualisations and operationalisations of task complexity can similarly be found in relation to task difficulty. Depending on which model one is working with, task difficulty can be characterized in two different ways. Robinson (2001, 2011) views task difficulty related to individual variables such as motivation and anxiety, which affect the learner's ability to cope with task demands. Skehan (2014: 6), on the other hand, considers difficulty as "inherent in the task, rather than learner-dependent", and argues that task difficulty is influenced by a range of different task-internal and task-external factors. These different perspectives and definitions for task difficulty create problems for researchers interested in how complexity and difficulty interact. We therefore specifically examine the construct of task difficulty within the context of task complexity in the current study, to try and tease out whether it can be identified as a separate factor, as in Robinson's model, or should indeed be associated with a mix of task-internal and task-external factors, as in Skehan's model.

Current operationalisations of the construct are also inconsistent – here, we assume that task difficulty, however it originates in either model, can be seen in perceptual terms as the level of challenge associated with performing a given task, as experienced by the learners during performance. Following previous research (Ortega, 2005; Author 2, 2009a, 2009b), we explored learners' perceptions of difficulty through retrospective analysis.

2.2 Intentional reasoning

We now turn to the role of intentional reasoning (IR), identified by Robinson as a specific resource-focusing variable. Compared to studies examining other cognitive and contextual variables contributing to task complexity, e.g. organisation of information and planning time, the existing research on intentional reasoning (IR) is less substantial in amount, less systematic in defining and operationalising IR, and less convincing in terms of consistency of its findings. The first challenge confronted by the current body of research is to provide a clear and coherent definition of IR that allows for a systematic evaluation of the construct. In general terms, IR refers to task requirements to explain other people's intentions and reasons, i.e. "understanding and explaining the motives, beliefs and thoughts which cause others to perform certain actions" (Robinson, 2007, p.194). Such requirements may be seen as demonstrating the ability to use more complex abstract language (see below). Within this definition, Robinson (2015) hypothesizes that tasks that require describing motion events (spatial reasoning), explaining reasons (causal reasoning), and reading others' minds (intentional reasoning) direct learners' attention to use more accurate and complex language to convey the reasoning demands of the tasks.

In terms of cognitive psychology, IR refers to the process of predicting and describing others' behaviours based on one's own assumptions about their beliefs, desires and knowledge (Astington & Baird, 2005). Leighton (2004) suggests that IR is a process of describing conclusions, drawn deductively or inductively, about others' thoughts and beliefs. These notions of the role of IR are predicated on a view of actions as executed intentionally, rather than randomly or unpremeditatedly (Bratman, 1987). IR is therefore seen as a critical element involved in a) observing others' actions and behaviours, and b) arriving at conclusions about others' thoughts, intentions and beliefs. IR as a cognitive process is assumed to be serial

(Leighton, 2004) and dependent on a chain of logical premises and hypotheses (Gilhooly, 2004). Thus the serial processes and semantic operations to create the logical chain involved in IR should lead to extra burdens on attention and working memory in constructing an appropriate preverbal message, i.e. at the Conceptualiser stage in Levelt's model (1989) of speech production. Psycholinguistically, depicting these thoughts, describing and justifying them involve use of specific language that denotes intentionality and reasoning. For example, describing IR is expected to elicit more complex structures (e.g., logical subordinators), and more complex lexis (e.g., cognitive status verbs). It is therefore logical to assume that tasks that involve IR – i.e. hypothesising, interpreting and drawing conclusions about others' behaviours - will necessarily be more demanding, and therefore more complex, according to either of our frameworks discussed above; higher IR should therefore affect linguistic performance measured across typical CAF variables.

Two TBLT studies have so far investigated the effects of IR on language performance (Ishikawa, 2008; Robinson, 2007), but both reveal variation in their operationalisations of IR and inconsistency in their findings. Robinson (2007) employed three interactive narratives with increased IR demands, e.g. simple, medium and complex IR demands. The simple task required reasoning about the intention of only one character, whereas the more complex tasks required reasoning about a number of characters whose intentions were dependent on others' thoughts and beliefs. Although performance in the more complex tasks generated more interaction, it failed to show higher levels of syntactic complexity, accuracy or fluency measured by general linguistic indices such as clauses per C-unit, percentage of error-free C-units, type-token ratio, or syllables per second. Ishikawa (2008) used hypothetical situations about the relationship between people in their workplace as the context for manipulating three self-created levels of IR demands: no IR, simple IR, and complex IR. While the no-IR task required only describing relationships between members of staff, the two IR conditions required reporting troubles in relations between two workers (simple IR) or four workers (complex IR). The results of Ishikawa (2008) revealed the IR tasks were associated with increased complexity (S-nodes per T-unit and Guiraud 2000) and accuracy (percentage of error-free T-units), but decreased fluency (speech rate, and dysfluency). The findings of this study strongly supported the predictions on the Cognition Hypothesis regarding the positive effect of IR demands on complexity (syntactic and lexical), and accuracy. However, we challenge the no-IR condition in Ishikawa's study, as we do not see IR as a dichotomous category, but rather a continuum in which oral narratives can be defined as more or less demanding (in line with the cognitive literature reviewed earlier).

In a systematic review of the literature on the Cognition Hypothesis, Jackson and Suethanapornkul (2013) have identified two shortcomings with research in reasoning demands: a) the paucity of research examining the different aspects of reasoning demands, and b) the lack of consistency in operationalisation of reasoning demands. Taking this further, we would argue that IR has been neither adequately defined nor systematically operationalized in TBLT before. The current study aims to address these limitations, with a particular focus on testing the predictions of Robinson's Cognition Hypothesis, given the specification of that model for the interplay between complexity, difficulty and oral performance.

3. Research Questions

The study reported here aimed firstly to investigate the effect of degree of IR demands on L2 learners' oral performance in terms of syntactic and lexical complexity, accuracy and fluency; secondly, their perceptions of task difficulty. Two levels of intentional reasoning, i.e. more vs.

less IR demands, were used to operationalise task complexity in this study. The study was designed to answer the following research questions:

1. Does degree of intentional reasoning demand in an oral narrative task affect L2 learners' oral performance in terms of syntactic complexity, lexical complexity, accuracy and fluency?
2. Are learners' perceptions of task difficulty affected by degree of IR demand?

Following the assumptions of the Cognition Hypothesis (Robinson, 2011), it was hypothesised that the more complex task (more IR demands) would generate more syntactic and lexical complexity and more accuracy at the expense of less fluency. It was also hypothesized that the learners would perceive the more complex task (more IR demands) as more difficult.

4. Methodology

The study employed a within-participant design in which the participants completed two oral narrative tasks with different degrees of complexity operationalised by two levels of IR demand, and a retrospective questionnaire designed to tap their recall of perceived task difficulty. In order to avoid any influence of rehearsal and order on performance, the design was counterbalanced between the participants. The two levels of IR served as independent variables. The five dependent variables were the four aspects of oral performance, i.e., syntactic and lexical complexity, accuracy and fluency, and participants' perceptions of task difficulty. Twenty students at a secondary school in Jordan participated in this study. All participants were male and spoke Arabic as a first language. They were 16 years old and had been learning English for ten years at school. They reported they had never lived in an English-speaking country before. The course at the private school where the learners studied, and during which data were collected, offered five hours of English instruction weekly. The school aligned learners to three level-groups (A, B, C) based on their proficiency in English with 'A' being the highest proficiency level equal to B2 level of CEFR. The school uses an examination portfolio of different tests and continuous assessment, across all areas of language skill, to determine the students' proficiency levels. However, these are not internationally standardised tests. The participants in this study were all taken from level 'A'; the mean average of their English proficiency level was 91.2% (SD = 3.79) based on the portfolio scores from the battery of school tests and continuous assessment conducted at the time of the data collection.

4.1 Tasks and Materials

Two video-based oral narrative tasks were used to elicit learners' oral performance. The choice of a video-based narrative was motivated by the assumption that watching a video under time constraints would allow for very limited online planning (Skehan & Shum, 2014; Wang, 2014,) and therefore would enable us to observe the impact of task design on spontaneous task performance more clearly. The two video clips were selected from the *Pat & Mat* show (Beneš, & Jiránek, 1976), a silent animated television series about two friends who constantly encounter trouble and are consequently challenged with complicated situations. The characters, Pat and Mat, typically use a variety of creative, unpredictable and optimistic strategies to overcome a chain of obstacles. The episodes offer a rich stimulus for a watch-and-tell condition as viewers can narrate the actions and events while watching the story unfold. The silent nature of the episodes makes them useful tasks in which the learners can reason about Pat & Mat's intentions, read their thoughts, and predict their reactions.
In order to evaluate the comparability of the two prompts, de Jong and Vercelloti's (2015) framework was adopted in which the three dimensions of structure, storyline complexity and

number of elements were carefully considered. The shortlisted prompts all had schematic sequential structure (Author 2, 2009a), similar storyline complexity (Author 2 & collaborator, 2008), and equal number of characters and props (de Jong and Vercelloti, 2015), but they differed in terms of the amount of IR they required. A third researcher helped us to choose the two best video clips from a shortlist.

To ensure the two tasks differed in the degree of IR required, IR was further operationalised at instruction-level through explicit task instructions. In the *less-IR* condition, the participants were asked to *tell and describe* the events as they happened; in the *more-IR* condition, participants were asked to tell and describe and to '*explain why the characters solve their problems or behave in certain ways'*. By explaining why, participants were assumed to need to reason about Pat and Mat's intentions, trying to read their thoughts. In order to unfold the characters' intentionality, the participants would need to draw conclusions about Pat and Mat's solutions to their problems, and to predict their future actions where possible, while narrating the story. The instructions were presented to the participants in their L1 before each video was shown. The participants watched 90 seconds of each video clip, but they often spoke longer. The first clip showed the two characters cooking lunch outdoor when it started to rain. Trying to solve the problem, they used different strategies to start a fire to help with their cooking. The second clip showed the two characters driving a car when they were inspired to try and make their car fly. Facing this challenge, they used different plans and techniques to turn their car into a flying vehicle. In Video 1, the less complex task (-IR henceforth), Pat and Mat's actions were self-expressive, i.e. descriptive, with little need to reason why they were doing what they did. Video 2, representing the more complex task (+IR henceforth), showed the two characters engaging in activities that were initially unclear in terms of why Pat and Mat were doing what they did, and therefore there was a need for explaining the characters' intentions as the story unfolded. This uncertainty promoted opportunities for making predictions and hypothetical thoughts about what the characters would do next.

The retrospective task-difficulty perception questionnaire was administered immediately after the participants performed the two tasks; the questionnaire asked them a range of questions, both multiple-choice and open-ended. The multiple-choice questions asked the participants to rate the difficulty level of each task by ranking them from 1 (very easy) to 4 (very difficult). The open-ended questions asked the participants to describe why they found one task more difficult or easier than the other. In addition to investigating learners' perceptions of task difficulty and matching them against the two models of task complexity, including a retrospective questionnaire would enable us to validate the researchers' choice of the video clips, in terms of IR and task difficulty. This would provide the kind of validity evidence on the manipulation of task complexity called for in current research (e.g. Révész, et al., 2016).

4.2 Procedures

After ethical consent was sought, the data were collected at the participants' school where each individual was seen by one of the researchers in a quiet room. A pre-task planning-time opportunity was not provided as we were keen to examine the effects of IR under unplanned conditions. The two videos were presented in a counter-balanced order: half of the participants started with the –IR task while the other half started with the +IR task. A digital voice recorder with a headphone was used to record the participant's oral performances. The participants were asked to narrate the story while watching the video clips. At the end of the 90-second clips, they were given 20 extra seconds[i] to finish speaking if they already had not done so. The participants recorded their perceptions of task difficulty by filling in the questionnaire immediately after completion of the two tasks.

4.3 Data coding and analysis

The data were transcribed using SoundScriber software (Breck, 1998). The AS-unit (Foster, Tonkyn, & Wigglesworth, 2000) was employed as a basic unit to segment the transcriptions. The data were then coded for a range of measures of syntactic complexity, lexical complexity, accuracy and fluency. In this study, we employed three measures to examine syntactic complexity: mean length of AS-unit, mean length of clause, and ratio of subordination (number of clauses per total number of AS-units). The choice of these measures is in line with Norris and Ortega (2009) who argued that it is necessary to include both length and subordination measures for exploring syntactic complexity in intermediate levels of proficiency. It also supports a more nuanced approach to complexity as argued in recent research on syntactic complexity, e.g. Inoue (2016) who has shown that the choice of syntactic complexity measures should be carefully considered in relation to the nature of the task and task-essentialness, i.e. what is essential to attend to, in terms of grammatical structures, to perform the task successfully. Lexical variety was measured using D (Malvern & Richards, 2002), which is a type-token based measure that corrects for variations in text length. Given the text-internal nature of D, it is a preferred measure for examining language performance when elicited by stimuli with different topics or content (de Jong & Vercelloti, 2015). Voc-D in Coh-Metrix software (Graesser, McNamara & Louwerse, 2003) was used to calculate the D values. Two measures of accuracy were used: percentage of error-free clause (Foster & Skehan, 1996) and errors per 100 words (Mehnert, 1998). Both these measures are reported as reliable and useful measures of accuracy (Ellis & Barkhuizen, 2005)[ii]. Following Kahng (2014), Author 2 and collaborator (2011), and Author 2 (2011), six measures of fluency were used: repair measures (i.e. repetition, reformulation and false start), number of filled pauses, number of mid-clause silent pauses, number of end-clause silent pauses, mean length of mid-clause silent pauses and mean length of end-clause silent pauses. A threshold of 0.25 second, which has proved to be a reliable length for analysing silent pauses in L2 research (de Jong & Bosker, 2013), was used to identify a silent pause. All measures of fluency were calculated per 60 seconds of performance. Measures of length of silent pauses were calculated by use of GoldWave software (2009). To confirm the reliability of the coding, a second researcher checked 10% of the coded transcriptions independently. Pearson's *r* co-efficient of over .91% were achieved for all the measures.

For all the data, SPSS 21.0 was used to run the descriptive and inferential analyses. In relation to the oral performance scores, a MANOVA was employed to identify whether there were statistically significant differences between performances on selected measures in the -IR and +IR tasks. MANOVA was preferred over running a number of separate ANOVAs for its power to control the risks of Type 1 error (Pallant, 2013). The normality of the distribution of the data was checked by a Kolmogorov-Smirnov test. The non-significant results indicated that the data were normally distributed. This allowed us to continue with paired-sample t-tests to locate significant differences in performance on specific variables between the two groups. Cohen *d* effect size was calculated when significant differences were observed. In order to interpret the findings in terms of effect sizes, we note a difference between what Cohen (1988) considers as a small (0.2), medium (0.5) and large (0.8) effect size, and what Plonsky and Oswald (2014) report as small (0.6), medium (1.00) and large (1.40) for within-participant study design in applied linguistics. We follow Plonsky and Oswald's recommendations in this study.

With respect to analysing the questionnaires, descriptive statistical analysis and t-tests were used for the quantitative parts of the data, whereas the qualitative sections of the data were subjected to a thematic analysis (Creswell, 2013). In running a thematic analysis, the qualitative data were first examined to identify common patterns in the statements of the participants about their perceptions of difficulty. Where possible, the common patterns were clustered together to form a theme. The quantitative and qualitative data, including the most recurrent themes about why the participants found a task difficult, are discussed in the Results section.

5. Results

As noted earlier, the current study examined the effects of level of IR demands on different aspects of L2 oral performance, and participants' perceptions of task difficulty depending on level of IR. Descriptive statistics for oral performance through measures of syntactic complexity, lexical complexity, accuracy, and fluency are presented in Table 1 below.

INSERT TABLE 1 HERE

Results obtained from the descriptive analysis indicated that the +IR task generated more syntactically complex and accurate performance, whereas the -IR task elicited improved performance in terms of lexical complexity. With regard to fluency measures, the mixed results did not show a regular pattern in the data favouring one or the other IR condition.

Given the small sample size of the study, running a MANOVA with many dependent measures is not recommended (Tabachnick & Fiddel, 2007). Therefore, in line with previous research (Skehan & Foster, 2005; Author 2 & collaborator, 2005), we selected four measures of mean length of AS unit, D, percentage of error-free clauses, and mean length of mid-clause silent pauses for running the MANOVA. Our rationale for this selection was based on research findings (Author 2 & Collaborator, 2005) which showed these measures consistently loaded on the complexity, accuracy and fluency constructs across different tasks. The results of the MANOVA for these four measures revealed that the effect of IR was statistically significant for syntactic complexity (Wilks' Lambda = .632; F = 11.04, *p* = .004; η^2 = .368), lexical complexity (Wilks' Lambda = .654; F = 10.04, *p* = .005; η^2 = .346), and accuracy (Wilks' Lambda = .632; F = 11.50, *p* = .003; η^2 = .377). However, the effect was not statistically significant with regard to fluency (Wilks' Lambda = .866; F = 2.94, *p* = .102; η^2 = .134).

The generally significant results of the MANOVA across an indicative spread of variables allowed us to conduct paired-sample t-tests on all our variables to answer Research Question 1, by comparing the means of different measures in the two IR conditions. The results are presented in Table 2 below.

INSERT TABLE 2 HERE

With respect to complexity, it was hypothesised that +IR would generate more syntactically and lexically complex oral performance than the -IR task. The results of the t-tests showed that two of the three measures of syntactic complexity reached statistically significant differences with large effect sizes in favour of the +IR task in terms of mean length of AS-unit (*t* = -3.323, *p* = .004, *d* = -.85), and ratio of subordination (*t* = -2.962, *p* = .008, *d* = -1.01). Although mean length of clause was higher in +IR task, the variation failed to reach statistical significance (*t* = -.449, *p* = .659). In terms of lexical complexity, a statistically significant difference with a large effect size was obtained for lexical diversity D (*t* = 3.170, *p* = .005, *d* = 1.83). However, the result was in the opposite direction to the predictions of the study, i.e. more lexical diversity was associated with the performance in the -IR task.

Regarding the impact of IR on accuracy, the results indicated that more accurate language performance was observed in the +IR task. Both measures of accuracy reached statistically significant differences with large effect sizes for percentage of error-free clauses (*t* = -3.392, *p* = .003, *d* = -.89) and number of errors per 100 words (*t* = 2.878, *p* = .010, *d* = .79).

As regards fluency, in line with the Cognition Hypothesis, a less fluent performance was predicted in the more complex task +IR. The results of the t-tests in fact showed mixed results across the six measures of repair and breakdown fluency. Descriptively, the mid-clause silent pauses and filled pauses were fewer in the +IR condition, while repair measures were fewer in the –IR condition. However, the variation between the two tasks, also reflected in the standard deviations indicated in Table 2, failed to show any statistically significant differences, implying that degree of IR demand had no reliable effect on L2 learners' fluency.

To sum up the oral performance analysis, in relation to Research Question 1, the results provided only partial evidence that IR would systematically affect oral performance in line with the Cognition Hypothesis; measures of accuracy and syntactic complexity confirmed those predictions, whereas measures of fluency and lexical diversity did not conform to the predictions of the Cognition Hypothesis.

Turning now to data to answer Research Question 2 on perceived task difficulty, a questionnaire was used to collect quantitative and qualitative responses to perceptions of difficulty. Across a scale of 4 possible answers from very easy to very difficult, the majority of participants indicated that the +IR task was perceived as more difficult (M = 2.75, SD = .71) than the -IR task (M = 2.10, SD = .55) (see Figure 1 below). The t-test results revealed a statistically significant difference with a large effect size in favour of the +IR task ($t = 3.32$, $p = .004$, $\eta^2 = 1.02$).

INSERT FIGURE 1 HERE

As discussed earlier, thematic analysis was used for the qualitative data to examine why the participants rated the +IR task as more difficult and which specific aspects of the tasks they found more demanding. The results suggested that a number of factors affected the participants' perceptions of task difficulty. Table 3 below shows the most common themes, their frequency and percentages as well as examples for each theme.

INSERT TABLE 3 HERE

The most frequently-mentioned themes related to the participants' awareness of higher cognitive demands associated with the +IR condition, i.e. the requirement not only to describe the events but to justify, reason, and think ahead about the characters' intentions and behaviours while speaking. Participants' comments about the cognitive demands are divided into two categories of task-induced and task-inherent demands. As can be seen in Table 3, the task-induced cognitive demand category referred to the requirement imposed on the participants through explicit task instructions, encouraging them to provide reasoning about the characters' intentions, predicting and justifying their actions. Task-inherent cognitive demands were slightly more frequently rated, seen in statements in which the participants referred to the demands inherent in the task, largely related to topic familiarity and predictability of the events. For this reason, they found the +IR task more difficult since the video contained unpredictable or less familiar events. By contrast, the participants described the actions in the -IR video as more common and predictable and thus easier to describe.

The third most frequent theme identified related to linguistic demands including the need for a lexical item or a specific structure required to narrate the story. Statements in this category often referred to the participants' linguistic needs in expressing the reasoning required in the +IR condition. The final theme included comments and statements about the pressure of *speaking in real time*, i.e. having had to watch and narrate the story at the same time as events were developing. Although this kind of pressure is a characteristic of video-based oral narratives and should be equally felt in both IR conditions, all the comments were made about the +IR task. We will discuss these results in further detail below.

6. Discussion

The current study was designed to explore the effects of two levels of IR on L2 learners' oral performance, and how including different levels of IR may affect learners' perceptions of task difficulty. Examining the effects of IR on L2 performance was argued to be particularly important for our current theoretical understanding of task design in relation to complexity and difficulty, in that the two existing models of task complexity, i.e. Skehan's Limited Attention Capacity and Robinson's Cognition Hypothesis, predict different effects of complexity on L2 performance but omit a clear definition of IR within task complexity or task difficulty. We also sought to clarify methodological issues relating IR, being the first study to our knowledge to carefully define and systematically operationalise IR in order to test the connection between IR, task complexity and task difficulty.

To recap the results shown above, we found that performance in the +IR condition was associated with a) more complexity in terms of subordination and length of AS unit, and b) more accuracy in terms of percentage of error-free clauses and number of errors in 100 words. These results clearly support the predictions of the Cognition Hypothesis (Robinson, 2007) about the impact of task complexity on syntactic complexity and accuracy.

However, the results of lexical complexity contradict the assumptions of Cognition Hypothesis as performance in the -IR task was more lexically varied than that in the +IR condition. Similarly, the non-significant results for fluency measures implied that fluency was not affected by the IR demands in the way the Cognition Hypothesis expected. While performance was more fluent in the +IR condition in terms of number and length of mid-clause pauses and length of end-clause pauses, it was more fluent in the -IR condition with respect to filled pauses, number of end-clause pauses and repair measures.

These results are not fully consistent with findings from the only two previous studies we are aware of which included IR. Our data only partially confirm Robinson's (2007) findings; and like Ishikawa (2008), we found positive effects of IR on syntactic complexity and accuracy, but the negative effect on lexical complexity found here contradicts Ishikawa's results. We assume this is partly due to the inconsistency in operationalising IR and task complexity across the three studies, and argue strongly that these constructs must be more clearly and systematically defined for future research.

The findings of the current study are also different from the results of Jackson and Suethanapornkul's (2013) meta-analysis that revealed "small positive effects for accuracy and small negative effects for fluency" (p. 330). However, our results support Jackson and

Suethanapornkul's (2013) findings in that the predictions of the Cognition Hypothesis for complexity were not confirmed.

In terms of evaluating Skehan's and Robinson's models of task complexity, we focused particularly on testing the predictions of Robinson's Cognition Hypothesis. The findings of our study support his model in terms of accuracy and syntactic complexity (i.e. capacity to process form), but the results from the fluency and lexical complexity measures do not. These mixed findings therefore leave us with no conclusive claims about whether different aspects of speech production across the CAF triad improve (i.e. form + fluency) when the task is more cognitively demanding - a Cognition Hypothesis assumption, or whether different aspects of speech production compete with one another in the form of a tradeoff relationship (form vs. fluency) - a Limited Attention Capacity hypothesis. Clearly, further careful and systematic investigation of task complexity in general and of IR in particular is needed to untangle the cognitive implications of IR for task performance. We note that the participants in this study had good levels of proficiency with high levels of fluency, and suggest that further research should investigate IR demands in relation to individual learner variables such as proficiency level and cognitive processing, e.g. working memory capacity.

Another important area of research to focus on is to examine a wider range of performance measures of analysis for accuracy, fluency, complexity and lexis to see if operationalisation of IR interacts differently across the measures. Our data suggest there may be specific ways of using language for reasoning, justifications and expressing hypothetical thoughts which would affect these measures, particularly in regard to syntactic complexity, which has been claimed to require more nuance than simple count of clauses or length of clause (Norris and Ortega (2009). We are aware of the limitation of our lexical analysis as we only looked at D, while Jarvis (2013), among others, argues that to capture the lexical diversity, different measures of size, dispersion, and sophistication should be explored.

In our data on syntactic complexity, the two measures of length of AS unit and ratio of subordination showed statistically meaningful increases in the +IR task, but the difference between the mean length of clause in the two conditions was negligible. Not surprisingly, higher demand in IR is associated with more subordination, i.e. use of language that expresses the characters' intentions and reasons, e.g. '*I think they are planning to make an airplane'*. More subordination concomitantly results in longer AS units overall, but not necessarily longer clauses. In fact, the hypothetical expressions required to perform the task encouraged many short clauses such as '*I suppose'* and '*I believe'*. One conclusion we draw here is that while a higher degree of IR would complexify language in terms of subordination, it would also induce frequent use of short clauses; therefore, length of clause should not always be taken an indication of complexity. We note that the choice of analytic measures researchers employ to examine L2 performance will have a direct impact on the results they obtain. This confirms Inoue's (2016) recommendation that the choice of measures of complexity should be carefully considered.

IR had an interesting effect on accuracy also connected to the use of short clauses. We found that accuracy was significantly higher in the +IR condition, in that higher +IR demands seemed to encourage both more error-free clauses and fewer errors in 100 words, which we associate with the use of short clauses as noted above, which were mostly of an idiomatic nature for hypothetical expressions, but this needs further investigation to claim as a reliable connection.

As noted above, the findings for lexical complexity, measured by D, revealed that the -IR task was lexically more diverse. This refutes the assumption of the Cognition Hypothesis (Robinson, 2011) which states that high lexical complexity, inferring stronger grasp of less frequent words, would be an outcome of increased TC. It is possible to argue that the formulaic nature of the language used for explaining the characters' reasoning and justification (e.g. *I suppose*, *I assume, they want to*, etc.) may have encouraged a degree of repetitiveness in the learners' use of lexis in the +IR task, which in turn resulted in a lower index of D.

Given the number of formulaic phrases noted in the data, it is also possible to argue that the operationalisation of the IR condition in this study, i.e. task-inherent and task-induced reasoning demands in the +IR condition, may have encouraged repetitive use of certain kinds of lexical items that were intended to display the speakers' intentional reasoning. To test out this assumption, we used Compleat Lexical Tutor (2016) to run a post-hoc lexical frequency analysis to identify the most frequent words used by the participants in the two conditions. The results are presented in Table 4.

INSERT TABLE 4 HERE

As indicated in Table 4, the results showed that mental state verbs, conjunctions, modal verbs and adverbs of uncertainty were used more frequently in the +IR condition. In the order of their frequency, mental state verbs commonly used were *think, assume, seem,* and *want;* the most frequent conjunctions used were *so, but,* and *because*; the most recurrent adverbs of uncertainty used included *maybe, apparently,* and *probably*; and modal verbs used frequently were *may, could,* and *will*. These results can help explain the low index of D in the +IR condition, and highlight the importance of using more measures to investigate lexical complexity. The results also tie in with previous research (Albert, 2011; Author 2 & collaborator, 2011; Author 2 & collaborator, 2008; Tidball & Treffers-Daller, 2008) that contends that lexical complexity is at least to some extent contingent on the content of each task.

Turning finally to the notion of task difficulty, in line with previous research on cognitive demand effects on perceptions of task difficulty (Ahmadian, 2012; Gilabert, 2007; Author 2, 2009b; Révész et al., 2016), the results of the questionnaires clearly indicated that the participants perceived the +IR task as more difficult and the cognitive demands associated with it as a key factor affecting their perceptions of task difficulty. Although we had not asked the participants to rate the *mental effort* (Pallotti, 2014, Révész et al., 2016) associated with task performance, the qualitative responses provided by the students showed that the cognitive demands of the +IR condition influenced their perceptions of task difficulty. The findings of the current study combined with previous studies (Ishikawa, 2011; Robinson, 2007) provide empirical evidence that IR is an additional important variable to be considered in task design to ensure task difficulty can be adequately taken into account. The positive relationship between task difficulty and the cognitive demand of a task (whether task-inherent or task-induced) will certainly contribute to the development of a more reliable framework than currently exists for evaluating task difficulty. Recent research in this area, e.g. Révész et al (2016) provides strong evidence that learner self-rating of task difficulty can consistently help complement other sources of evidence, e.g. expert judgements and researcher manipulations of task complexity, to provide a valid and reliable framework for analysing task complexity and difficulty.

One interesting finding of the current study was that the participants attributed difficulty to both task-inherent cognitive demand, i.e. the content of the videos, and task-induced cognitive demand, i.e. task instructions. Firstly, this finding confirms that the operationalisation of IR at the two levels of task content and task instruction influenced learners' perceptions of task difficulty. Secondly, we argue that while the content of the +IR video engaged participants at the level of thinking, realising and understanding, the instructions were needed to encourage them to go beyond the familiar "*tell and describe*" level, using higher order cognitive and logical processes to explain, justify and predict events, which participants noted as unusual and much more demanding. Although these two types of demand may be inter-related, they seem to impose two different kinds of demand on the speech production process. The finding of the current study suggests that task-inherent (content-level) IR requirement made the task more demanding at the pre-verbal stage of Conceptualisation, whereas the task-induced (instruction-level) requirement increased the demand at the verbal stage of Formulation. The high demands on the two stages of production affected the participants' perceptions of difficulty. This post-hoc observation clearly warrants further examination, perhaps to be tied more closely to the predictions of Skehan's conceptualisation of task complexity which separates conceptualisation from encoding demands.

7. Conclusions

The findings of the study suggest that IR is an important aspect of task design that can affect L2 performance and learners' perceptions of difficulty. Although the findings only partially support each of the existing models of task difficulty, they provide robust evidence about the impact of IR on task design and L2 performance. Studies of this nature can also help TBLT researchers develop an index of task difficulty. Given the small scale of the study, we suggest that the findings are interpreted with caution and care. Clearly, a study with a larger sample size and with a design that allows for examining the interaction between IR and other cognitive processes and individual differences is needed. Processes like reasoning, justifying and predicting are not only highly demanding cognitive processes, they also require very specific language to convey the abstract and complex concepts of reasoning and intentions. We consider the relationship between task complexity and the linguistic requirements associated with task performance as an interactive and dynamic process that needs a more careful operationalisation, especially when using analytic measures of complexity, accuracy and lexis. It is hoped that the way IR has been operationalised and explored in this study can serve as a point of departure to develop a framework to investigate IR more systematically within the TBLT context. More comparable studies are still needed before such a framework can be recognised.

References

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

Albert, Á. (2011). When individual differences come into play. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 239-265). Amsterdam: John Benjamins.

Astington, J., & Baird, J. (2005). *Why language matters for theory of mind*. Oxford: Oxford University Press.

Beneš, L., & Jiránek, V. (1976). Pat & Mat, Czech stop-motion animated series. http://en.patmat.cz/home-pat-and-mat/

Bratman, M. (1987). Intention, Plans, and Practical Reasoning. Cambridge: Harvard University Press.

Breck, E. (1998). SoundScriber. Michigan: University of Michigan. Retrieved from http://www-personal.umich.edu/~ebreck/code/sscriber/

Bygate, M. (2015). *Domains and Directions in the Development of TBLT* (Vol. 8). Amsterdam: John Benjamins.

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching pedagogic task*. Harlow: Longman.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Creswell, J. (2013). *Research design: Qualitative, quantitative, and mixed methods approaches*. London: Sage publications.

De Jong, N., & Bosker, H. (2013). *Choosing a threshold for silent pauses to measure second language fluency.* Paper presented at the The 6th Workshop on Disfluency in Spontaneous Speech (DiSS).

De Jong, N., & Vercellotti, M. (2015). Similar prompts may not be similar in the performance they elicit: Examining fluency, complexity, accuracy, and lexis in narratives from five picture prompts. *Language Teaching Research*, 1-18.

Ellis, R. (2000). Task-based research and language pedagogy. *Language Teaching Research, 4*(3), 193-220.

Ellis, R. (2012). *Language teaching research and language pedagogy*. Sussex: Wiley Blackwell.

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

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in second language acquisition, 18*(03), 299-323.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354-375.

Gilabert, R. (2007). The simultaneous manipulation of task complexity along planning time and [+/- Here-and-Now]: Effects on L2 oral production *Investigating tasks in formal language learning* (pp. 44-68).

Gilhooly, K. (2004). Working memory and reasoning. In R. Sternberg & J. Leighton (Eds.), *The nature of reasoning* (pp. 49-77). Cambridge: Cambridge University Press.

GoldWave, I. (2009). GoldWave Software (Version V5.70). Retrieved from http://www.goldwave.coml

Graesser, A., McNamara, D., & Louwerse, M. (2003). What do readers need to learn in order to process coherence relations in narrative and expository text? In A. Sweet & C. Snow (Eds.), *Rethinking reading comprehension* (pp. 82-98). New York: Guilford Publications.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal*, 1-19.

Ishikawa, T. (2008). The effect of task demands of intentional reasoning on L2 speech performance. *The Journal of Asia TEFL, 5*(1), 29-63.

Ishikawa, T. (2011). Examining the influence of intentional reasoning demands on learner perceptions of task difficulty and L2 monologic speech. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (Vol. 2, pp. 307-330). Amsterdam: John Benjamins.

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.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(s1), 87-106.

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

Leighton, J. (2004). Defining and describing reason. In R. Sternberg & J. Leighton (Eds.), *The nature of reasoning* (pp. 3-11). Cambridge: Cambridge University Press.

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

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language testing, 19*(1), 85-104.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in second language acquisition, 20*(01), 83-108.

Norris, J., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics, 30*(4), 555-577.

Nunan, D. (1988). *The learner-centred curriculum: A study in second language teaching*. New York: Cambridge University Press.

Ortega, L. (2005). What do learners plan? Learner-driven attention to form during pre-task planning. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 77-110). Amsterdam: John Benjamins.

Pallant, J. (2013). *SPSS survival manual*. London: McGraw-Hill Education.

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

Pallotti, G. (2014). A simple view of linguistic complexity. *Second Language Research, 31*(1), 117-134.

Plonsky, L., & Oswald, F. (2014). How big is “big”? Interpreting effect sizes in L2 research. *Language Learning*, 64(4), 878-912.

Prabhu, N. (1987). *Second language pedagogy* (Vol. 20). Oxford: University Press Oxford.

Révész, A., Michel, M., & Gilabert, R. (2016). Measuring cognitive task demands using dual task methodology, subjective self-ratings, and expert judgments: a validation study. *Studies in Second Language Acquisition*, 38(4), 703-737.

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

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *IRAL-International Review of Applied Linguistics in Language Teaching, 45*(3), 193-213.

Robinson, P. (2011). *Second language task complexity: researching the cognition hypothesis of language learning and performance* (Vol. 2). Amsterdam: John Benjamins.

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and Directions in the Development of TBLT* (Vol. 8, pp. 87-121). Amsterdam: John Benjamins.

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

Skehan, P. (2001). Tasks and language performance. In M. Bygate, M. Swain & P. Skehan (Eds.), *Researching pedagogic tasks: Second language learning, teaching, and testing* (pp. 167-186). New York: Routledge.

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

Skehan, P. (2014). The context for researching a processing perspective on task performance. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 1-26). Amsterdam: John Benjamins.

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* (Vol. 11, pp. 193-216). Amsterdam: John Benjamins.

Skehan, P., & Shum, S. (2014). Structure and processing condition in video-based narrative retelling. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 187-210). Amsterdam: John Benjamins.

Tabachnick, B., & Fidell, L. (2007). *Using multivariate statistics* (5th ed.). MA: Allyn & Bacon.

Tidball, F., & Treffers-Daller, J. (2008). Analysing lexical richness in French learner language: What frequency lists and teacher judgements can tell us about basic and advanced words. *Journal of French language studies, 18*(03), 299-313.

Wang, Z. (2014). On-line time pressure manipulations. In P. Skehan (Ed.), *Processing Perspectives on Task Performance* (Vol. 5, pp. 27-61). Amsterdam: John Benjamins.

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

Table 1. Descriptive statistics for all oral performance measures

Dimension	Measures	- IR Mean (SD)	+ IR Mean (SD)
Syntactic Complexity	Mean length of AS-unit	7.73 (1.09)	8.95 (1.69)
	Mean length of clause	5.83 (.79)	5.91 (.81)
	Ratio of subordination	1.33 (.16)	1.50 (.16)
Lexical Complexity	Lexical diversity VOCD	27.37 (7.36)	24.62 (5.91)
Accuracy	Percentage of error free clauses	48.32 (15.76)	59.96 (9.66)
	Number of errors per 100 words	9.36 (2.77)	7.53 (1.73)
Fluency	Dysfluencies per minute	6.65 (4.22)	7.58 (5.18)
	Number of mid-clause silent pauses	5.85 (2.56)	5.55 (3.28)
	Mean length of mid-clause silent pauses	.79 (.28)	.65 (.26)
	Number of end-clause silent pauses	10.50 (3.01)	11.15 (3.45)
	Mean length of end-clause silent pauses	1.33 (.61)	1.13 (.46)
	Number of filled pauses	12.27 (5.28)	13.5 (5.11)

Table 2. T-test outputs and effect size

Dimension	Measures	- IR	+ IR	t-test	Sig. (2-tailed)	Effect size
		Mean (SD)	Mean (SD)	T	P	Cohen's d
Syntactic Complexity	Mean length of AS-unit	7.73 (1.09)	8.95 (1.69)	-3.32	.004*	-.85
	Mean length of clause	5.83 (.79)	5.91 (.81)	-.44	.659	---
	Ratio of subordination	1.33 (.16)	1.50 (.16)	-2.96	.008*	-1.01
Lexical Complexity	Lexical diversity VOCD	27.37 (7.36)	24.62 (5.91)	3.17	.005*	1.83
Accuracy	Percentage of error free clauses	48.32 (15.76)	59.96 (9.66)	-3.39	.003*	-.89
	Number of errors per 100 words	9.36 (2.77)	7.53 (1.73)	2.87	.010*	.79
Fluency	Dysfluencies per minute	6.65 (4.22)	7.58 (5.18)	-1.29	.211	---
	Number of mid-clause silent pauses	5.85 (2.56)	5.55 (3.28)	.36	.721	---
	Mean length of mid-clause silent pauses	.79 (.28)	.65 (.26)	1.71	.102	---
	Number of end-clause silent pauses	10.50 (3.01)	11.15 (3.45)	-.88	.389	---
	Mean length of end-clause silent pauses	1.33 (.61)	1.13 (.46)	1.71	.103	---
	Number of filled pauses	12.27 (5.28)	13.5 (5.11)	-1.70	.104	---

df = 19, *p < 0.05

Table 3. Participants' perceptions of task difficulty

Themes	Excerpts from the data	Frequency	%
Cognitive demand (task-inherent, e.g. familiarity, predictability)	*It is difficult because that's not common in normal life. Task was easy because they have done easy things that I can describe.*	**23**	**43%**

Cognitive demand (task-induced)	*I found it difficult because I have to read their minds and what they are thinking about.*	**22**	**41%**
Linguistic demands	*The first task was easy because the task doesn't need any hard words and meanings.*	**5**	**9%**
Time pressure	*The story is live and I have to think so fast to describe the actions.*	**4**	**7%**

Total number of comments: 54

Table 4. Lexical frequency analysis in +IR and –IR conditions

Word categories	**Times used in +IR**	**Times used in -IR**
Mental state verbs	152	41
Conjunctions	124	43
Modals verbs	99	27
Adverbs of uncertainty	42	5
Total	**417**	**116**

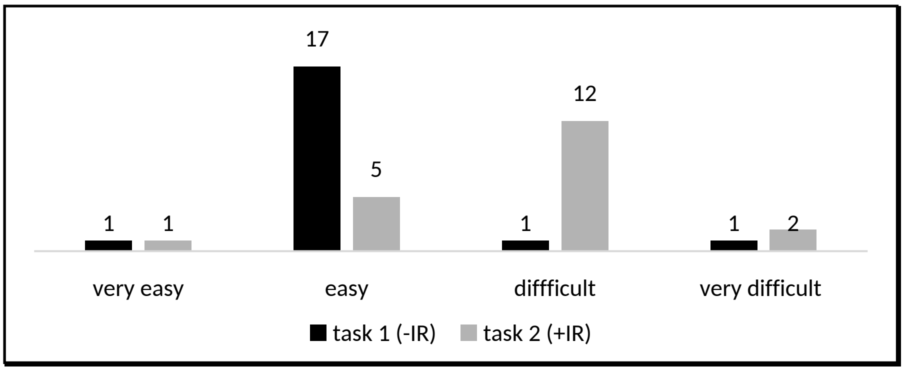

Figure 1. Participants' perception of task difficulty

[i] The inclusion of the extra 20 seconds was based on a post-pilot observation.

[ii] We did not include Foster and Wigglesworth's (2016) Weighted Clause Ratio measure, as this positively correlated with the percentage of error-free clauses in our study (Author 1 and Author 2, in preparation).