The effects of increasing reasoning demands on accuracy and complexity in L2 oral production

YAR-SING STACEY SHIAU & REBECCA ADAMS

University of Auckland

ABSTRACT

Robinson's (2001a) Cognition Hypothesis proposes that increasing the cognitive complexity of a second language task in specific ways should lead to more accurate and complex second language (L2) production. One of these ways is by increasing reasoning demands. To date, there are relatively few studies focusing on the role of reasoning demands in L2 task production. The present study is a replication and extension of Gilabert's (2007a) study which investigated the effects of manipulating task complexity through increasing reasoning demands of a task on accuracy in L2 oral production. This study goes beyond the original study in that it also examines the effects on the linguistic complexity of L2 production. Accuracy was measured by the number of errors per Analysis-of-Speech-unit, the ratio of total errors to total words, the number of self-repairs per AS-unit and the ratio of total self-repairs to total words. Complexity was measured by the number of clauses per AS-unit and the number of words per clause for syntactic complexity, and by the number of different grammatical verb forms (in each task) for syntactic variety and by the Guiraud Index of lexical richness for lexical variety (Vermeer, 2000). The

Address for correspondence: Yar-Sing Stacey Shiau, University of Auckland (MA graduate), 192 Norman Lesser Drive, St. Johns Park, Auckland 1072, New Zealand; Email: stacesing@gmail.com

University of Sydney Papers in TESOL, 6, 121-146.

results indicated that increasing task complexity in terms of reasoning demands had no significant effect on accuracy and syntactic complexity, but it led to significantly more lexically and syntactically varied learner production. These findings suggest manipulation of different task complexity variables may direct L2 learners' attentional resources to different aspects of language, which may help inform syllabus design for task-based language teaching.

INTRODUCTION

Task-based language teaching (TBLT) is increasingly advocated as a holistic approach to second language teaching (Bygate, Skehan, & Swain, 2001; Ellis, 2003; Samuda & Bygate, 2008; Willis & Willis, 2007), aiming to develop L2 proficiency by engaging L2 learners in meaning-focused communication through pedagogic tasks that have real-world relevance (Ellis, 2003). Tasks create a semantic space for L2 learners to engage in cognitive processes, providing learners with opportunities to both stretch and refine their interlanguage resources (Ellis, 2003). One of the central questions for task-based teaching is that of syllabus design; how should communicative tasks be organised in a coherent teaching programme that promotes language learning? Robinson (2001b) suggests that this should be done through task complexity. Under this perspective, by initially engaging in simple tasks and gradually moving towards more complex versions of real-world target tasks, learners should be able to transfer language abilities, developed in classrooms through the pedagogic tasks, to the real-world contexts (Long, 1998). Basing syllabus design on task complexity could create a need to develop criteria for grading and sequencing tasks to determine how tasks can be made more or less complex.

A currently influential proposal for grading production tasks is Robinson's (2001a) Cognition Hypothesis. Influenced by the information processing models, the Cognition Hypothesis proposes that tasks be graded by complexity based on cognitive processes. Task complexity is then "the result of the attentional, memory, reasoning, and other information processing demands imposed by the structure of the task on the language learner" (Robinson, 2001b, p.29). Robinson (2001a) claims in his Cognition Hypothesis that sequencing tasks according to their cognitive demands (from less complex to more complex) promotes language learning, because more cognitively demanding tasks may draw on more cognitive resources and more effective use of learning mechanisms, thereby forcing the learners to reanalyse and restructure their interlanguage. Robinson's (2001a) Cognition Hypothesis has motivated a growing body of empirical inquiry on the effects of task complexity on language production (e.g., Gilabert, 2007a, 2007b; Ishikawa, 2007; Kuiken & Vedder, 2007; Michel, Kuiken & Vedder, 2007; Robinson, 2001b, 2007a; Skehan & Foster, 1997). The purpose of the present study is to examine the effect of manipulating task complexity through reasoning demands on accuracy and linguistic complexity in L2 monologic oral production.

LITERATURE REVIEW

Early research on the role of task complexity in language production and learning was based on cognitive models that assumed limited attentional capacity. For example, Skehan and Foster (2001) argue in their Limited Attentional Capacity Model that, due to humans' "limited information processing capacity", accuracy and complexity are unlikely to increase simultaneously when task complexity is increased (p.189). Skehan and Foster's (1997) study which examined the effect of task type and task processing conditions on foreign language performance provided evidence to support their claim on the trade-off effect between accuracy and linguistic complexity in task performance.

In contrast to Skehan and Foster's (2001) hypothesis, Robinson's (2001a) Cognition Hypothesis predicts that increases in task complexity along the resource-directing dimensions will push L2 learners to produce more accurate and more complex output simultaneously. This hypothesis is based on Givon's (1985) prediction that complexity in language structures should increase when discoursal functions become more complex. The multiple-resource model proposed by Wickens (1989) has questioned the validity of the limited attentional capacity model and has suggested no competition for attentional resources when tasks draw on separate resource pools. According to Robinson (2001a), when tasks are made more complex in ways that help learners direct their attentional resources, language production becomes more complex and more accurate. Examples of resource-directing complexity factors include here-and-now vs. there-and-then, few elements vs. many elements, low reasoning demands vs. high reasoning demands.

Through an argumentative writing task, Kuiken and Vedder (2007) examined the role of reasoning demands in task production, defining reasoning demands as making decisions based on task information. Learners were asked to make a decision on their choice of the most suitable type of accommodation based on the criteria given and support it with reasons and arguments in the writing task. The less complex version involved three criteria, whereas the more complex version involved six criteria. This study partially supported Robinson's (2001a) Cognition Hypothesis, in that greater accuracy was elicited in the more complex version than in the less complex version, but no significant effect of task complexity was found on syntactic complexity and lexical complexity.

In a further study of reasoning demands, Robinson (2007a) operationalised reasoning demands in terms of judgments of intentions and how intentions related to subsequent behaviours. His participants were given pictures that formed a story and asked to put the pictures in the correct sequence and narrate the story to their partners who then had to sequence the pictures accordingly. The less complex task version involved the character's simple intention to build a house. In the two more complex versions, the main character's intentions to act in a certain way were influenced by their perceptions towards other characters' thoughts and beliefs. The findings of the study did not overall support the Cognition Hypothesis, as he found mixed results on syntactic complexity, no significant effect on accuracy and fluency, and his results on lexical complexity ran opposite to the predictions of the Cognition Hypothesis.

Gilabert (2007a) examined the effect of manipulating task complexity along the reasoning demands dimension on accuracy in L2 oral production in an English as a Foreign Language (EFL) setting using a decision-making task. He measured accuracy in terms of errors and self-repairs. He found no significant differences between the task versions in errors per Analysis-of-Speech-unit [AS-unit, c.f., Foster, Tonkyn & Wigglesworth (2000) and discussion below], ratio of errors to total words, error-repairs per AS-unit, ratio of error-repairs to total words, all self-repairs (including error-repairs and non-error repairs) per AS-unit, and percentage of self-repairs and ratio of repaired to unrepaired errors, indicating at best a limited role for reasoning demands in the accuracy of the participants' production.

However, Gilabert found significant differences between the task versions in the ratio of all self-repairs to total words and in the corrected ratio of repaired/unrepaired errors. In terms of all self-repairs/words, his participants repaired significantly more often in the less complex version of the task. This indicates that they monitored the accuracy of their second language production more often when engaged in a less complex version of the task, contrary to the predictions of the Cognition Hypothesis. In contrast, the other significant difference (on the corrected ratios of repaired/unrepaired errors) did provide support for the Cognition Hypothesis, indicating that participants repaired a higher proportion of errors on the more complex task. Overall Gilabert's (2007a) findings did not provide strong support for the Cognition Hypothesis. Gilabert (2007a) argued that the participants might have focused their attention on sequencing their actions and justifying their decisions, in other words, on task content. Consequently, fewer cognitive resources may have been available for monitoring their linguistic accuracy. He further suggested that the participants' cognitive resources may have been allocated to structural complexity, but he did not examine the effect on structural complexity in his study.

Prior research by Skehan and Foster (1997) indicated that a decision-making task involving "more complex outcome[s]" tended to elicit greater linguistic complexity in L2 output than task versions involving simple problems with straight-forward solutions (p.206). Reasoning tasks with complex outcomes pose multiple and intertwined problems which require learners to engage in a greater degree of "on-line computation" to make decisions that maximise the success of outcomes (Skehan and Foster, 1997, p.206). The task used in Gilabert's (2007a) study is a good example of a task that is made more complex by increasing the reasoning demands; however, this study did not investigate whether increasing the complexity of task outcomes also enhanced the complexity of linguistic production, as suggested by Skehan and Foster (1997). The purpose of the present study is to partially replicate Gilabert's (2007a) study in terms of examining the effect of increased task complexity on accuracy (adopting the task, accuracy measures, and the affective perception questionnaire from the original study), but also to extend the analysis to investigate the effect on linguistic complexity.

RESEARCH HYPOTHESES

The study investigates the following hypotheses:

1. Hypothesis 1: Following the Cognition Hypothesis, increased task complexity through increasing reasoning demands [+reasoning demands] will have a positive effect on the accuracy of L2 production. It is predicted that L2 learners will produce fewer errors and more instances of self-repair in the [+reasoning demands] condition than in the [–reasoning demands] condition.
2. Hypothesis 2: Following the Cognition Hypothesis, increased task complexity through increasing reasoning demands [+reasoning demands] will have a positive effect

on the lexical variety, syntactic complexity and syntactic variety of L2 production. It is predicted that L2 learners will produce more lexically varied, syntactically complex and syntactically varied language in the [+reasoning demands] condition than in the [–reasoning demands] condition.

METHODS

Participants

Fifteen participants (13 females and 2 males) from six language schools in New Zealand were recruited for this study. Their ages ranged from 20 to 39 (*M* = 24.27). They were enrolled in intermediate to advanced English courses. They were speakers of Mandarin (n = 4), Korean (n = 7) and Japanese (n = 4). All had studied English prior to coming to New Zealand and had studied in New Zealand for less than 6 months (*M* = 2.63). In order to ensure that differences in proficiency among the learners did not affect the results and to be certain that learners would have sufficient oral English ability to complete the task, they first participated in a pre-test (described below). Based on the performance of the first participants, a minimum proficiency level for the study was selected. Learners who did not meet the minimum proficiency level were removed from the data analysis.

Pre-test materials

The pre-test materials included a listening test and a grammar test, and were extracted from Longman Complete Course for the TOEFL test (2001, p.519-520 and 524-527). The participants' ability to communicate freely might be associated with their ability to process L2 in real time, which is reflected through their listening scores (c.f., Yuan and Ellis, 2003). The grammar scores were selected following Ellis's (2005) rationale that they may reflect the participants' rule-based competence, which is related to accuracy and complexity. To qualify for the study, participants needed to achieve a score of 50% on the listening test and 44% on the grammar test. These percentages acted as the benchmark to eliminate low-proficiency learners from the study and to control the effect of language proficiency on language production measures, so effects on accuracy and complexity could with more certainty be attributed to differences in task complexity. Descriptive statistics for the pre-test scores are displayed in Table 1. The narrow score ranges and relatively low standard deviations indicate that the participants' proficiency was quite uniform.

TABLE 1
Descriptive statistics for the listening and grammar pre-test scores

	Listening (12 items)	Grammar (25 items)
Mean	8.20	14.87
Median	8	15
Standard Deviation	1.66	2.85
Range	5	8
Minimum	6	11
Maximum	11	19

Note: The full score of the listening test is 12 and the full score of the grammar test is 25

Three tasks were used in the present study, a model task and two research tasks. The model task served as an example for the participants, and did not form part of the data for the study.

Model task

An audio-taped recording of a model decision-making task was used as a demonstration to ensure the participants' understanding of the task instructions and familiarity with the task type. It had the same task instructions as the research tasks, but was set in a different situation (flood instead of fire). The model task was performed by an ESL learner with high English proficiency without intervention from the researchers.

Research tasks

The less and more complex versions of a monological decision-making task (See Appendices A and B) originated from Gilabert (2007a) and were used with his permission. This decision-making task required each participant to act as a local fire chief and organise his/her team to rescue the victims in a burning building. The participants were required to describe what was happening in the picture, decide what actions to take and in what order. They were also asked to explain the reasons behind their decisions. The two versions differed in several ways. In the [-reasoning demands] condition, the learners were presented with ample resources (personal/equipment), while in the [+reasoning demands condition] there were fewer staff and less equipment available. In the [-reasoning demands] version, all victims looked similar, while in the [+reasoning demands] version, some victims were more vulnerable (e.g., children, a pregnant woman, an elderly man and an injured person), and victims were placed in more precarious positions. Other differences included locations of ventilation shafts and smoke blowing into, rather than out of, the building. According to Gilabert (2007a), the [+reasoning demands] condition forced the participants to make decisions involving multiple stages, where one decision conditioned later decisions, whereas the [–reasoning demands] condition only required single-step decisions because there were fewer cause-effect relationships between the elements. Specific items and characters in the pictures were labelled in the participants' first languages to avoid confusion. Following pilot testing, the English word 'ventilation shaft' was included in both pictures because it was generally unknown. The task instructions and prompts were slightly revised after pilot-testing to ensure clear understanding of the task requirements and elicitation of sufficient oral data.

Pre-task planning time was provided prior to each research task to help the participants pre-organise their propositional content, as suggested by Yuan and Ellis (2003). Consequently participants could direct more attentional resources towards language use during task performance, which was the research focus of the study. Following Yuan and Ellis (2003), notes were taken away prior to task performance to avoid mixing oral and written production in performance. Moreover, following Ishikawa (2007), in order to elicit the use of articles, the participants were not allowed to name the victims in the pictures.

Post-task questionnaire

A questionnaire on affective perceptions was adopted from Gilabert (2007a; see Appendix C). It was given to ensure that the operationalisation of task complexity matched the participants' perception of difficulty.

Data collection

One of the researchers met for 75 minutes with each participant in a library meeting room. Each participant completed the pre-test and then listened to the model task recording. Participants' understanding of the task instructions was checked before proceeding to the research tasks.

Each participant carried out both research tasks individually with the researcher, with the order counter-balanced, following Gilabert's (2007a) data collection procedures. To maintain consistency for each participant, the researcher did not speak or indicate misunderstanding, but acted as a listener receiving an explanation. An independent sample t-test was performed to find out if there was any significant difference in English proficiency levels between the group who did the simple-complex task order (n = 7) and the group who did the complex-simple task order (n = 8). There were no significant differences between the two groups in terms of their English proficiency levels on either the listening ($t = .121$, $p = .91$) or the grammar test ($t = 1.53$, $p = .15$).

The participants' oral production for the research tasks was recorded using a digital recorder. Five minutes of pre-task planning time were provided for each task. Notes written during the planning time were removed before the recording began. The researcher did not guide learner language use or decision-making. However, if a participant indicated that they had finished the task after only carrying out part of the task instructions (e.g., describing pictures without making decisions and justifications), the researcher would silently point at the forgotten instruction(s) as a reminder. After completing each task, participants were asked to fill out the post-task questionnaire for the task they had just finished. Personal information was retrieved from the participants at the end of the session through an informal discussion with the researcher of their language learning experiences in New Zealand and their home country (see earlier description of participants).

Most participants met with the researcher on a one-to-one basis. However, a few chose to come in pairs. These participants took the pre-test at the same time and listened to the model task together, but were not allowed to talk or look at each other's work. They were carefully monitored by the researcher. They then did the research tasks, questionnaire and personal information retrieval individually with the researcher. The participant not doing the tasks was asked to leave the room for 30 minutes and not to enter the room unless summoned.

Analysis

All the transcripts were first coded by one of the researchers based on linguistic accuracy and complexity measures described below. Based on the same coding guidelines, the other researcher checked all transcripts. Any discrepancies were then resolved so that 100% agreement in all coding was reached.

This study adopted accuracy and complexity measures as dependent variables to assess the participants' oral performance. The Analysis-of-Speech-unit developed by Foster *et al.* (2000) was used as the unit of analysis. An AS-unit was coded as a single utterance containing "an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either." (p.365). The AS-unit was chosen to segment oral data over the Terminable-unit because the AS-unit takes the reality and nature of oral data such as dysfluency and incomplete utterances into account, as argued by Foster *et al.* (2000).

Accuracy measures

Following Gilabert (2007a), accuracy was measured in the following ways: the number of errors per AS-unit, the ratio of the number of total errors to the total number of words, the number of self-repairs per AS-unit, and the ratio of self-repairs to the total number of words. Two types of ratios, one based on the total number of AS-units and the other one based on the total number of words, were calculated to avoid disadvantaging participants who produced longer AS-units.

Complexity measures

Syntactic complexity was measured by the number of clauses per AS-unit and the number of words per clause. The measure of clauses per AS-unit was adopted from Michel *et al.* (2007); words per clause is suggested as a complexity measure by Ellis (2003).

Syntactic variety was measured by the number of different grammatical verb forms used in each task, following Yuan and Ellis (2003). The grammatical verb types chosen for analysis were: simple present-past-future, present-past-future perfect, present-past-future progressive, present-past-future perfect progressive, modal present-past and passive voice.

Lexical variety was measured by the Guiraud Index of lexical richness. The Guiraud Index was chosen over the Type-Token Ratio because the Guiraud Index corrects for the effect of text length (Vermeer, 2000). This procedure is done mathematically by using the square root of the number of tokens.

Statistical analysis

The study compared the performance of one sample of learners across two task versions; therefore paired-sample t-tests were used for the main analysis to determine whether task complexity impacted on the accuracy and complexity of language production, and to examine the effect of task complexity on the five affective perceptions of task difficulty. The alpha level for statistical significance was set at .05.

RESULTS

Questionnaire on affective perceptions

The affective questionnaire responses were analyzed to determine whether the learners perceived the more complex version of the task as more difficult. Paired-sample t-tests indicated that participants did not perceive difficulty, stress, confidence, interest and motivation across the task versions as shown in Table 2.

TABLE 2
Affective perceptions across task versions (paired-sample t-test)

Affective perceptions	**Simple Task**		**Complex Task**		**t**	**df**	**p**
	Mean	**SD**	**Mean**	**SD**			
Difficulty	6.07	1.58	6.80	1.42	-2.05	14	.060
Stress	4.87	1.30	4.33	1.99	1.42	14	.178
Confidence	3.67	1.35	3.40	1.59	.81	14	.433
Interest	6.93	1.79	7.20	1.70	-1.29	14	.217
Motivation	6.73	1.91	6.73	2.02	.00	14	1.000

Note. Number of the participants = 15; **p < .05,* two-tailed

However, there was a non-significant trend for the more complex task version to be perceived as more difficult (p = .060) than the less complex version. This finding suggests that the operationalisation of task complexity through increasing reasoning demands from the less complex to the more complex versions was successful.

Oral production measures

To investigate the hypotheses, paired-sample t-tests were used to determine whether task complexity influenced the accuracy and complexity of language production. These findings will be reported separately.

Accuracy

As shown in Table 3, the paired-sample t-test indicates that there were no significant differences across the task versions in all of the accuracy measures.

TABLE 3
Performance comparisons between the task versions on accuracy (paired-sample t-test)

Dependent variable	Simple Task		Complex Task				
	Mean	SD	Mean	SD	t	df	p
errors/AS unit	1.65	.52	1.75	.66	-.79	14	.443
errors/words	.16	.05	.17	.07	-.90	14	.385
SR/AS unit	.39	.23	.37	.22	.32	14	.755
SR / words	.04	.02	.04	.02	.36	14	.722

Note. *$*p < .05$*, two-tailed

The descriptive statistics show that, contrary to Hypothesis 1, more errors were generated in the more complex task version (*M* for errors/AS-unit = 1.75, *M* for errors/words = .17) than the less complex version (*M* for errors/AS-unit = 1.65, *M* for errors/words = .16). However, the differences were not statistically significant for either errors/AS unit (*p*= .443) or errors/total words (*p*= .385). For self-repairs, no significant differences were found when accuracy was measured either by self-repairs/AS-unit (*p*= .755) or self-repairs/total words (*p*= .722). The high *p*-values of the four accuracy measures, especially the measures on self-repairs, indicate that any differences here are very unlikely to be related to task complexity.

Complexity

As displayed in Table 4, the paired-sample t-test shows that significant differences across the task versions were found in the measures for lexical variety and syntactic variety, but not in the measures for syntactic complexity.

TABLE 4
Performance comparisons between the task versions on complexity (paired-sample t-test)

Dependent variable	**Simple Task**		**Complex Task**				
	Mean	**SD**	**Mean**	**SD**	**t**	**df**	**p**
GI	5.16	.60	5.61	.58	-3.74	14	.002*
GV	5.00	1.13	5.80	.94	-2.45	14	.028*
Clauses/ AS unit	1.41	.18	1.49	.15	-1.74	14	.104
Words / clause	7.21	.86	6.89	.88	1.66	14	.119

Notes: GI = Guiraud Index of lexical richness; GV = number of different grammatical verb forms used in the task; *$p < .05$*, two-tailed

As predicted by Hypothesis 2, the descriptive statistics indicate that the more complex task version (*M* for Guiraud Index = 5.61; *M* for number of Grammatical Verb forms used = 5.80) elicited more lexically varied and more syntactically varied language than the less complex task version (*M* for Guiraud Index = 5.16; *M* for number of Grammatical Verb forms used = 5.00) from the participants, with statistically significant differences found for both the Guiraud Index of lexical richness (*p*= *.002)* and the number of different grammatical verb forms used in each task (*p*= *.028)*. Together, these results suggest that increasing task complexity by increasing reasoning demands led to increased lexical and syntactic variety of learner language production. However, for syntactic complexity, no significant differences were found in either clauses/AS-unit (*p*= .104) or words/clause (*p*= .119) (although in both cases the results approached significance). These findings indicate that syntactic complexity was not influenced by task complexity in a statistically significant way.

Because of the small sample size in this study (which can allow individual performance to affect group measures), there was concern that task order may have played a role in the findings despite counterbalancing. In this case, for example, greater complexity of language production in the more complex task may have been caused by the performance of those who completed the more complex version last. In order to investigate this possibility, the data were split according to the order of task completion.

TABLE 5
Effects of task order and task complexity on lexical variety (2x2 repeated measures ANOVAs)

Source	**Between-subjects**			
	Df	**F**	**η²**	**p**
Task order	1	1.89	.127	.192
Error	13	(.55)		
Source	**Within-subjects**			
	Df	**F**	**η²**	**p**
GI	1	12.90*	.498	.003
GI x task order	1	.03	.003	.858
Error (GI)	13	(.12)		

Notes: The value enclosed in the parenthesis represents the mean square error. GI = Guiraud Index of lexical richness; **p < .05*

TABLE 6
Effects of task order and task complexity on syntactic variety (2x2 repeated measures ANOVAs)

Source	**Between-subjects**			
	Df	**F**	**η²**	**p**
Task order	1	.24	.018	.635
Error	13	(1.45)		
Source	**Within-subjects**			
	Df	**F**	**η²**	**p**
GV	1	6.28*	.326	.026
GV x task order	1	.96	.069	.345
Error (GV)	13	(.80)		

Notes: The value enclosed in the parenthesis represents the mean square error. GV = number of different grammatical verb forms used in the task; **p < .05*

This split allowed for a *post hoc* analysis using repeated measures ANOVAs which was performed to investigate if there was any significant effect of task order on the Guiraud Index of lexical richness and the number of different grammatical verb forms used in each task and on the perceived difficulty. The factors in the ANOVA were order of completion (first or second) and task complexity (less complex or more complex). As shown in Table 5 and Table 6, between-subjects tests indicated that task order did not influence lexical variety (*p*= .192) and syntactic variety (*p*= .635) of language production; within-subjects tests showed that there were no significant interaction effects of task complexity and task order on the measures of lexical variety (*p*= .858) and syntactic variety (*p*= .345). It can be concluded that the participants' task performance across task versions was not affected by the order of task presentation, but rather by differences in task complexity.

DISCUSSION

In this study, the effect of increasing reasoning demands on accuracy and complexity in L2 oral production in an ESL setting was investigated.

Hypothesis 1: accuracy

The results indicated that there was no significant effect on all the accuracy measures. These findings did not support the Cognition Hypothesis which claimed that more cognitively demanding tasks manipulated along the resource-directing dimensions could direct L2 learners' attentional and memory resources to focus more on accuracy (Robinson, 2007b). This study's findings on accuracy echo Robinson's (2007a) findings in the percentage of error free C-units and part of Gilabert's (2007a) findings in errors per AS-unit, the ratio of errors to the total number of words and self-repairs per AS-unit. However, in terms of the ratio of self-repairs to the total number of words, this study's findings do not coincide with Gilabert's (2007a) study, which found a significant difference in this ratio between the task versions but in the opposite direction of the Cognition Hypothesis.

Overall, the results of the present study do not support the prediction of Robinson's (2001a) Cognition Hypothesis that increasing task complexity will lead to greater accuracy in production. Accuracy in language production may have reflected participant interlanguage knowledge rather than any effect of task complexity. Participants did indicate that they found the tasks quite challenging. It is possible that variation in the accuracy of language production would be measurable if the study were conducted among learners with higher language proficiency. However, this explanation would not apply to accuracy measured through self-repair, as frequency of self-repair does not seem to be related to language proficiency (c.f., Kormos, 2006, cited in Gilabert, 2007a). The uniformity in self-repair behaviour across task versions suggests that engaging in a more complex task did not direct more attentional resources to accuracy.

Hypothesis 2: complexity

This study's findings for linguistic complexity measured through lexical variety and grammatical variety supported Robinson's (2001a) Cognition Hypothesis which predicted that increasing reasoning demands would increase linguistic complexity, and Gilabert's (2007a) prediction that the fire-chief tasks might direct L2 learners' attention towards linguistic complexity over accuracy. However, the uniformity in instructions across task versions (description-decision-justification) may have promoted similar degrees of syntactic complexity, e.g., the repetitive use of 'because' adverbial clauses and relative clauses. The present study's results on syntactic complexity mirrored Robinson's (2007a) finding that increased reasoning demands did not influence syntactic complexity.

Gilabert's (2007a) fire-chief task design seemed to be the key contributor in capturing the significant differences in lexical and grammatical variety across task versions in the present study. The inclusion of more distinguishing elements to the victims in the more complex task version increased reasoning demands of the task in ways that promoted a wider range of lexical items. In addition, the increased number of emergency situations, reduced number of rescue resources, and more complex fire conditions seemed to heighten reasoning demands of the task in ways that pushed the participants to use a wider repertoire of grammatical verb forms to convey their message content.

The recruitment of the participants from an ESL setting may also have contributed to the significantly higher lexical and grammatical variety found in the more complex task version. As Lafford (2004) suggests, ESL learners are likely to focus on conveying meaning over accuracy of forms because of the "pragmatic exigencies" imposed on these learners in their daily interactions with the native speakers of the target culture (p.213). Consequently, the participants may have been more willing than participants in earlier studies to push their linguistic boundaries and experiment with new language forms. As Skehan and Foster (2001) state, "greater complexity is taken to be a surrogate of a willingness to experiment, and to try to extend or make more elaborate the underlying IL [interlanguage] system" (p.191).

CONCLUSION

Similar to several studies including Michel *et al.* (2007), Robinson (2007a), and Ishikawa (2007), this study has found mixed results that only partially support Robinson's (2001a) Cognition Hypothesis in terms of accuracy and complexity. The small sample size, the participants' varied English-learning backgrounds, and the use of a single task type may have limited the power of the study to illustrate effects of task complexity on production (c.f., Ishikawa, 2007).

However, relatively few studies to date have been conducted on the effect of reasoning demands on L2 production, and few studies on task complexity to date have demonstrated an effect on the complexity of language production. This study provides evidence that, as Robinson (2001a) predicts, increasing task complexity through reasoning demands may influence the linguistic complexity of L2 production. The findings of the current study suggest that the manipulation of task complexity can push learners to experiment with language. While more research is required to explore this link, these findings may assist task-based syllabus designers to grade and classify tasks to moderate focus on linguistic complexity and help language teachers to select appropriate tasks to promote attempts to stretch interlanguage resources in language practice.

THE AUTHORS

Yar-Sing Stacey Shiau is a MA graduate in Language Teaching and Learning from the University of Auckland. Her research interests include task-based language teaching and scaffolding in second language teaching and learning.

Rebecca Adams is a Senior Lecturer in Applied Language Studies at the University of Auckland. Her research interests include peer interaction, task-based learning, and individual differences in language learning. Her recent research appears in *Language Learning, TESOL Quarterly*, and the *Modern Language Journal*.

REFERENCES

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

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford: Oxford University Press.

Ellis, R. (2005). *Principles of instructed language learning*. Retrieved from http://www.asian-efl-journal.com/May_2005_Conference_Ellis.php.

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

Gilabert, R. (2007a). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics, 45*(3), 215-240.

Gilabert, R. (2007b). The simultaneous manipulation of task complexity along planning time and ± here-and-now: Effects on L2 oral production. In M.P. García-Mayo (Ed.), *Investigating tasks in formal language learning* (pp.44-68). Clevedon: Multilingual Matters.

Givón, T. (1985). Function, structure, and language acquisition. In D. Slobin (Ed.), *The cross-linguistic study of language acquisition* (vol.1, pp.1008-1025). Hillsdale, NJ: Lawrence Erlbaum.

Ishikawa, T. (2007). The effect of manipulating task complexity along the (± Here-and-Now) Dimension on L2 written narrative discourse. In M.P. García-Mayo (Ed*.), Investigating tasks in formal language learning* (pp.136-156). Clevedon: Multilingual Matters.

Kuiken, F. & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M.P. García-Mayo (Ed.), *Investigating tasks in formal language learning* (pp.117-135). Clevedon: Multilingual Matters.

Lafford, B.A. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a Second Language. *Studies in Second Language Acquisition, 26*, 201-225.

Long, M.H. (1998). Focus on form in task-based language teaching. *University of Hawai'i Working Papers in ESL, 16*(2), 35-49.

Michel, M.C., Kuiken, F. & Vedder, I. (2007). The influence of complexity in monologic versus dialogic tasks in Dutch L2. *International Review of Applied Linguistics, 45*, 241-259.

Phillips, D. (2001). *Longman complete course for the TOEFL test: Preparation for the computer and paper tests.* New York, NY: Addison-Wesley Longman.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: a triadic framework for examining task influences on SLA. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.287–318). Cambridge: Cambridge University Press.

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

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

Robinson, P. (2007b). Criteria for classifying and sequencing pedagogic tasks. In M.P. García-Mayo (Ed.), *Investigating tasks in formal language learning* (pp.7-26). Clevedon: Multilingual Matters.

Samuda, V. & Bygate, M. (2008). *Tasks in second language learning.* Basingstoke: Palgrave Macmillan.

Skehan, P. & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185-211.

Skehan, P. & Foster, P. (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.183–205). Cambridge: Cambridge University Press.

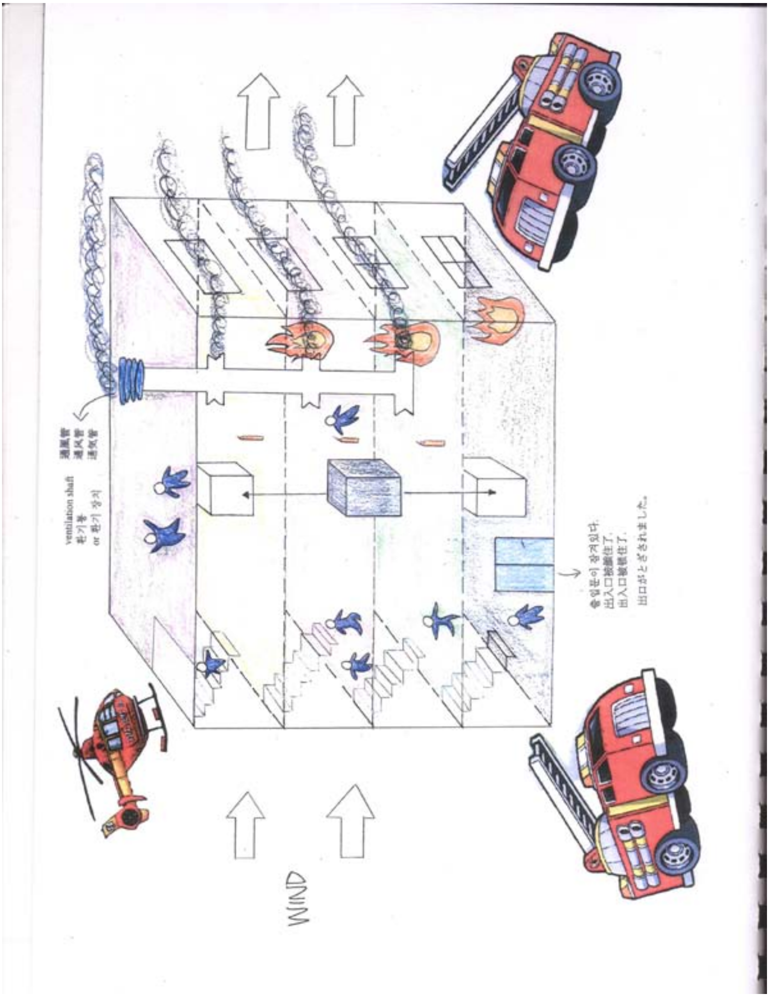
Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65-83.

Wickens, C. (1989). Attention and skilled performance. In D. Holding (Ed.), *Human skills* (pp.71-105). New York, NY: John Wiley.

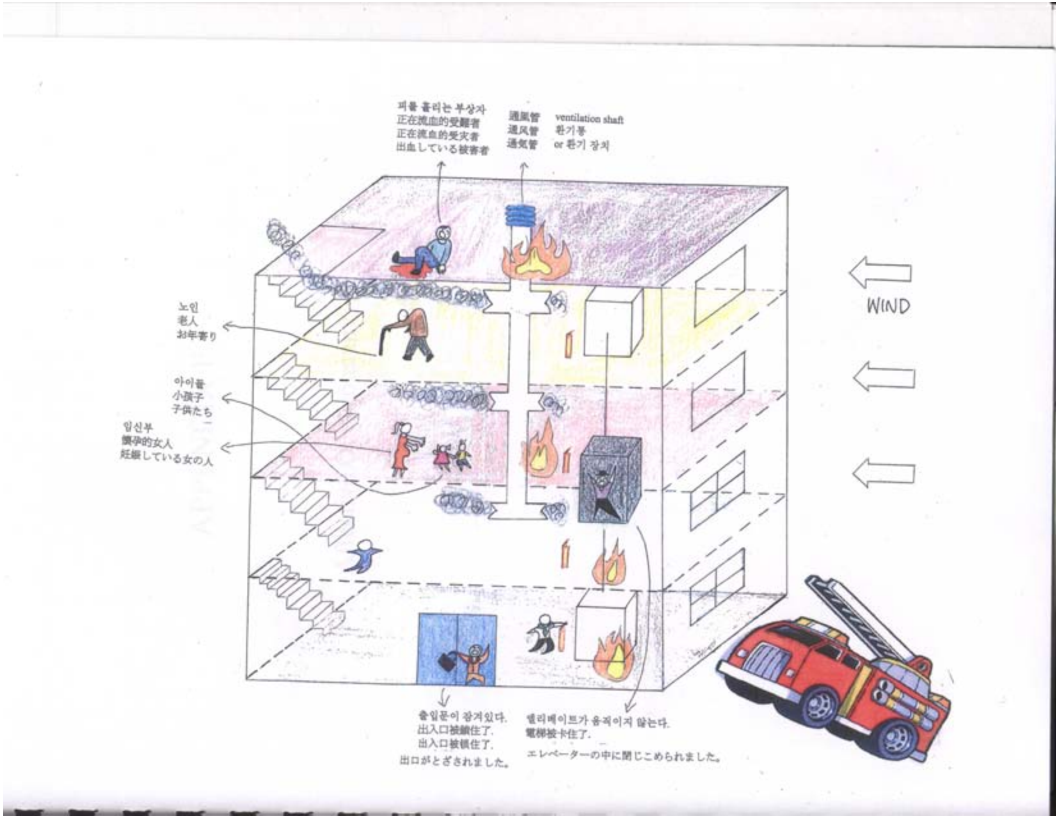
Willis, D. & Willis, J. (2007). *Doing task-based teaching: A practical guide to task-based teaching for ELT training courses and practising teachers.* Oxford: Oxford University Press.

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

APPENDIX A: RESEARCH TASK ([–REASONING DEMANDS] / LESS COMPLEX VERSION)



APPENDIX B: RESEARCH TASK ([+REASONING DEMANDS] / MORE COMPLEX VERSION)



APPENDIX C: AFFECTIVE QUESTIONNAIRE

Affective Variables Questionnnaire

Name: ____ Group: ___________

Consider the task you've just carried out.

Evaluate the task by circling the appropriate answer in each case:

I thought this task was easy	1	2	3	4	5	6	7	8	9	I thought this task was difficult
I felt frustrated doing this task	1	2	3	4	5	6	7	8	9	I felt relaxed doing this task
I did not do this task well	1	2	3	4	5	6	7	8	9	I did this task well
This task was not interesting	1	2	3	4	5	6	7	8	9	This task was interesting
I don't want to do more tasks like this	1	2	3	4	5	6	7	8	9	I want to do more tasks like this