Studies in Second Language Acquisition, 2015, **37**, 549–581.
doi:10.1017/S0272263114000618

THE INTERSECTION OF TASK-BASED INTERACTION, TASK COMPLEXITY, AND WORKING MEMORY

L2 Question Development through Recasts in a Laboratory Setting

YouJin Kim
Georgia State University

Caroline Payant
University of Idaho

Pamela Pearson
Linfield College

The extent to which individual differences in cognitive abilities affect the relationship among task complexity, attention to form, and second language development has been addressed only minimally in the cognition hypothesis literature. The present study explores how reasoning demands in tasks and working memory (WM) capacity predict learners' ability to notice English question structures provided in the form of recasts and how this contributes to subsequent development of English question formation. Eighty-one nonnative speakers of English completed three interactive tasks with a native speaker

This project was funded by the Cleon F. Arrington Research Initiation Grant at Georgia State University. We are indebted to four anonymous reviewers and the editor, who provided invaluable comments. Special thanks go to Melissa Baralt and Scott Crossley for their insightful comments and suggestions during the revision process. Any remaining errors are our own.

Correspondence concerning this article should be addressed to YouJin Kim, Department of Applied Linguistics and ESL, Georgia State University, 25 Park Place, Suite 1500, Atlanta, GA 30303. E-mail: ykim39@gsu.edu

interlocutor, one WM task, and three oral production tests. Prior to the first interactive task, participants were randomly assigned to a task group (simple or complex). During task performance, all learners were provided with recasts targeting errors in question formation. The results showed that learners' cognitive processes during tasks were in line with the cognitive demands of the tasks, at two complexity levels. The findings suggest that WM was the only significant predictor of the amount of noticing of recasts as well as of learners' question development. With regard to interaction effects between WM and task complexity, high WM learners who carried out a complex version of the tasks benefitted the most from task-based interaction.

Since Long's (1996) updated interaction hypothesis, there has been a surge in research conducted on the effects of conversational interaction on second language (L2) learning, to the degree that "a robust connection" between interaction and learning is now commonly accepted (Gass & Mackey, 2007, p. 176; see also Mackey, Abbuhl, & Gass, 2012, for a review). Learner internal and external factors mediating the positive relationship between interaction and L2 learning have been identified, and the role of task design features in both L2 performance and interaction-driven language learning is being increasingly examined in the field of instructed SLA. Researchers have explored the effects of task complexity on L2 development by testing the predictions of Robinson's (2011) cognition hypothesis and Skehan's (1998) trade-off hypothesis. However, very little research has looked into the ways individual learners' cognitive characteristics, such as working memory (WM) capacity, mediate the effects of task complexity on L2 learning. Moreover, from a methodological standpoint, researchers have yet to systematically document evidence of task complexity validation (e.g., Baralt, 2013; Gilabert & Barón, 2013; Norris, 2010; Révész, 2014). To address these issues and to test the predictions of the cognition hypothesis, the current study (a) assesses whether task complexity manipulations are in line with learner perceptions and cognitive processes during task completion; (b) examines the relationship between task complexity, WM, and noticing of recasts; and (c) investigates the development of English language questions by L2 learners.

BACKGROUND

Task Complexity and the Cognition Hypothesis

The construct of task complexity has motivated a large body of research in the fields of SLA and task-based language teaching. Both Skehan's

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

trade-off hypothesis and Robinson's cognition hypothesis predict the impact of task complexity on language performance (see the meta-analysis done by Jackson & Suethanapornkul, 2013). Whereas Skehan's trade-off hypothesis predicts that learners' attentional resources are limited during task performance and that cognitively demanding tasks consume learners' attentional resources, Robinson's cognition hypothesis (2001a, 2003, 2005) predicts that learners are able to access multiple and noncompetitional pools of attention. Both Skehan and Robinson present specific claims about language production and development. In particular, Robinson highlights the ways task design features and learner factors such as WM affect interlanguage development and language performance. Robinson also connects input and interaction to the cognitive and conceptual demands of tasks and particularly addresses interaction-oriented language learning contexts (Long, 1996; Mackey et al., 2012). Because the current study focuses on interaction-driven language learning through task performance, we examined the relationship between task complexity, WM, and L2 development on the basis of the cognition hypothesis.

According to Robinson (2001a, 2001b, 2005, 2007a), a triadic componential framework for pedagogic task classification differentiates among three sources of cognitive demands: (a) task features, (b) interactive features, and (c) learner factors. The first source, task features, includes inherent features of tasks that impact the level of cognitive complexity, classified as either resource-directing or resource-dispersing variables. Resource-directing variables direct attention and memory resources to the linguistic features necessary for successful task performance and so may make greater demands on attention and memory. On the other hand, resource-dispersing variables make increased performative or procedural demands as tasks increase in complexity (Robinson, 2001a, 2005). Increases in task complexity along resource-directing dimensions can be achieved by manipulating reasoning demands, the number of elements, and/or narrating events that are displaced in time or space (Robinson, 2001a). Those along resource-dispersing dimensions can be achieved by drawing learner attention to nonlinguistic areas—namely, requiring learners to perform more than one task simultaneously or to carry out a task with no prior knowledge. The second task classification, interactive features, comes from the task settings and interaction conditions. Robinson suggests that although tasks need to be sequenced on the basis of cognitive complexity (from simple to complex via resource-dispersing and then resource-directing variables), characteristics of task conditions are "held constant and replicated each time more cognitively complex pedagogic versions are attempted so as to help ensure development" (Robinson, 2011, p. 13). The third task source takes into account learner factors, such as affective variables (e.g., motivation or anxiety)

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

and ability variables (e.g., aptitude or WM), which can influence performance on tasks and L2 development.

Drawing on the claim that L2 pedagogic tasks should be sequenced from simple to complex to maximize L2 learning (Baralt, Gilabert, & Robinson, 2014; Robinson, 2003, 2011), several predictions about the interaction among task complexity, learner factors, L2 performance, and L2 development have emerged. In conversational contexts, it is predicted that the cognitive and conceptual demands of a task will differentially affect the amount of interactional features (e.g., language-related episodes [LREs]) and the uptake of forms made salient through task input. Specifically, Robinson and Gilabert (2007) propose that more cognitively complex tasks will facilitate "greater attention to, and uptake of, forms made salient during the provision of reactive focus on forms techniques such as recasts" (p. 167). Robinson also theorizes that increasing task complexity will lead to L2 development, especially with developmentally more advanced forms (Robinson, 2007a, 2007b; Robinson & Ellis, 2008). In terms of interaction among variables, Robinson (2011) suggests that learner factors can contribute to our understanding of between-learner variations in task performance and argues that learner factors such as WM, anxiety, and aptitude are more apparent during complex tasks than during simple tasks. Given that the current study tested the predictions concerning the role of task complexity during interactive tasks, the subsequent section reviews empirical studies that were conducted in conversational interaction contexts.

Studies Exploring Task Complexity, Interaction, and L2 Development

With the promotion of interactive tasks in educational contexts, researchers have studied the role of task complexity on interaction-driven L2 development during learner-learner and native speaker (NS)–learner interaction. One strand has explored the effects of task complexity on the incidence of interactional features during learner-learner interaction, with a subset of these studies further examining subsequent L2 learning (e.g., Baralt, 2014; Gilabert, Barón, & Llanes, 2009; Kim, 2009; Nuevo, 2006; Révész, 2011; Robinson, 2001b, 2007b). Another strand of research has examined researcher-learner interaction in lab-based contexts, wherein the researcher provides oral corrective feedback during task performance (Baralt, 2013; Révész, 2009; Révész, Sachs, & Mackey, 2011).

Research reporting on learner-learner interaction during collaborative tasks in classroom contexts has provided insights into the effects of task complexity on interaction-driven L2 learning. In a study examining the role of task complexity in L2 interaction, uptake, and perception of

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

task difficulty with 42 English as a foreign language (EFL) learners, Robinson (2007b) devised three levels of task complexity by manipulating the [+/– reasoning] variable during a picture-narration task. Results showed that reasoning demands resulted in a significant increase in the number of turns taken, clarification requests, and confirmation checks. With 60 EFL learners, Gilabert et al. (2009) examined the role of task complexity in the occurrence of interactional features, employing three tasks: a narrative reconstruction task, a decision-making task, and an instruction-giving map task. They found that the frequency of interactional features increased with more complex tasks, particularly with the narrative reconstruction and instruction-giving map tasks.

Révész (2011) also tested the impact of manipulating the [+/– reasoning] variable on interactional features. Although her study, involving 43 English as a second language (ESL) learners, did not reveal statistically significant group differences for confirmation checks, clarification requests, or recasts, it did show that the more complex tasks promoted a significantly higher number of LREs and metalinguistic talk. Finally, Kim (2009) examined the effect of task complexity on the occurrence of LREs using a picture-narration and a picture-difference task with 34 ESL participants at two proficiency levels. She found that the complex narration task elicited more LREs by the higher proficiency learners than the simple narration task, whereas the opposite was true for lower proficiency learners. For the picture-difference task, the lower proficiency learners produced significantly more LREs during the complex task than during the simple task. Overall, with regard to the effects of task complexity on the amount of interaction-driven learning opportunities, previous studies only partially support the claim that task complexity positively affects the incidence of interactional features. Furthermore, these findings show that other task design and learner factors can mediate the relationship between task complexity and interaction-driven learning opportunities.

Studies implementing a pretest-posttest research design in learner-learner interaction contexts provide further insights into how task complexity may or may not promote L2 learning. For instance, Nuevo (2006) manipulated the [+/– reasoning] variable with two interactive task types targeting locative prepositions and the past tense. The results did not show any relationship between task complexity and L2 development. Kim (2012) explored multiple levels of task complexity by examining the [+/– reasoning] variable with three complexity levels (i.e., simple, +complex, ++complex) in question development with 191 EFL university learners. She found that the ++complex group achieved the greatest advancement in question development, which may have been attributable to a greater amount of LREs targeting more advanced questions during learner-learner interaction. Kim and Tracy-Ventura (2011) also found that the ++complex group outperformed

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

the +complex group, followed by the simple group, in the development of the English past tense.

The second research strand has examined dyads made up of researcher-learner interaction in controlled laboratory contexts (Baralt, 2013; Révész, 2009; Révész et al., 2011). Typically, the researcher provides corrective feedback (e.g., recasts) during performance and then examines the extent to which learners acquire the target linguistic forms across different complexity levels. In Révész's (2009) study with EFL learners in which she examined the relationship among task complexity, recasts, and the development of past progressive forms, learners were randomly assigned to a feedback/complexity group: recast/simple, recast/complex, no recast/simple, and no recast/complex. Those in the recast/complex group demonstrated a greater amount of L2 development in terms of past progressive forms than those in both simple task conditions. In a follow-up study, Révész et al. (2011) examined the relationship among task complexity, uptake of recasts, and L2 development. They discovered that task complexity did not influence the rate of uptake, yet uptake was a positive predictor of L2 development in the simple condition.

Baralt (2013) added another dimension to interaction-based task complexity research—that is, face-to-face (FTF) versus computer-mediated-communication (CMC) environments. She examined how these two interactional learning environments mediate the efficacy of recasts in promoting the learning of the Spanish subjunctive. The participants (n = 84) carried out tasks with different complexity levels that were manipulated by [+/– intentional reasoning] with a researcher. In the FTF mode, findings showed that performing the cognitively complex task while receiving recasts led to the most learning, whereas in the CMC mode, the cognitively complex task with the provision of recasts did not lead to L2 learning.

In sum, the past decade has witnessed an increasing number of studies testing the predictions of the cognition hypothesis, particularly in interactional contexts. These studies have analyzed interactional feedback, uptake, and LREs as indicators of learning opportunities and have explored the acquisition of task-induced linguistic features using pretest-posttest-delayed posttest designs. In general, the role of task complexity remains enigmatic: Although it appears to promote interaction-driven learning opportunities and some linguistic development, it does so inconsistently. The research implies that this is due to differences in operationalizations of task complexity and in learner-internal variables. The previous mixed findings regarding the effects of task complexity on L2 learning support the need to further investigate cognitive complexity with more robust designs that confirm the validity of task complexity (e.g., Révész & Gilabert, 2013) and that explore the moderating role of individual learner factors.

Task Complexity, Working Memory, and Interaction-Driven L2 Development

Individual difference factors have been hypothesized to influence the impact of task complexity (Robinson, 2011); among these factors, cognitive abilities (e.g., WM) are presumed to be of particular importance (e.g., Révész et al., 2011). Thus, the current study sought to examine whether WM moderates the relationship between task complexity and interaction-driven language learning opportunities. In fact, previous SLA research has shown that WM, or "the ability to maintain information in an active and readily accessible state, while concurrently and selectively processing new information" (Conway, Jarrold, Kane, Miyake, & Towse, 2007, p. 3), is thought to be one of the main cognitive factors affecting interaction-driven L2 learning overall (e.g., Goo, 2012; Li, 2013; Mackey, Adams, Stafford, & Winke, 2010; Mackey, Philp, Fujii, Egi, & Tatsumi, 2002; Mackey & Sachs, 2012; Miyake & Friedman, 1998; O'Brien, Segalowitz, Collentine, & Freed, 2006; Révész, 2012; Sagarra, 2007; Trofimovich, Ammar, & Gatbonton, 2007; Yilmaz, 2013).

The most widely accepted WM model, developed by Baddeley and Hitch (1974), involves a multicomponent memory system composed of a central executive system (i.e., an overall supervisor of information) and two domain-specific slave systems (i.e., a phonological loop and a visual-spatial sketchpad). The phonological loop is responsible for the temporary storage of phonological information, and the visual-spatial sketchpad stores and processes visual and spatial information. Baddeley (2000) extended the original model to include a fourth component called an episodic buffer, which holds and integrates visual, spatial, and verbal information. The phonological loop and central executive systems are the two most widely investigated components of WM. Given that articulation takes place in real time, resulting in a limited span of immediate memory, the phonological loop is of limited capacity. This component is often measured by immediate serial recall of numbers or words (Baddeley, 2003). Alternatively, complex WM, involving the functioning of the central executive system, can be measured with reading and listening spans (Daneman & Carpenter, 1980) as well as running span tests (Broadway & Engle, 2010). These tasks typically require that participants store information while processing new information.

To date, most interaction-oriented SLA research has explored the extent to which WM plays a role in the noticing of feedback and interaction-driven L2 learning. For instance, in arguably the first study on WM capacity as a moderator of task-based interaction learning, Mackey et al. (2002) found that, during conversational interaction, learners with higher WM capacity noticed recasts better than those with lower

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

WM capacity. Additionally, Mackey et al. (2010) showed that WM was positively correlated with the amount of modified output produced during collaborative tasks. A study by Mackey and Sachs (2012) noted that older learners with higher WM demonstrated question development through interactive tasks. Furthermore, Goo (2012) revealed that although recasts and metalinguistic explanations were equally effective on learners' acquisition of the *that*-trace filter in English, WM was what significantly mediated the effectiveness of recasts. These findings strongly imply that executive attention is involved in the noticing of recasts.

Considering the claims of the cognition hypothesis regarding maximizing learning, there is very little research that examines how WM modulates the relationship between task complexity and L2 development. Among the few studies that have examined this relationship, Kormos and Trebits (2011) investigated how WM mediated 44 EFL secondary school students' oral production—in terms of complexity, accuracy, and fluency—during two narrative tasks (complex picture- vs. simple cartoon-narration tasks). The results showed that students produced significantly more diverse vocabulary during the simple cartoon picture task than the complex picture-narration task, but no difference was found in other areas. In terms of WM effects measured by a backward digit span test, only the cartoon-narration task showed a significant effect of WM on syntactic complexity, suggesting a limited role of WM in L2 oral language production. Kormos and Trebits concluded that WM capacity may not affect language production but, rather, may affect the amount of attention that learners can devote to noticing various linguistic features presented to them in the input (measured indirectly by the accuracy of linguistic structures).

In another study, Baralt (2010) explored the role of task complexity and WM in the development of the Spanish past subjunctive through recasts in both the FTF and online CMC modes. Task complexity was operationalized as [+/– intentional reasoning]. Learners in the less complex group had to retell a story (–intentional reasoning), whereas learners in the more complex group, in addition to retelling a story, also had to hypothesize why a character in the story performed a certain action (+intentional reasoning; those in the simple group were provided with that information). Learners in all conditions received recasts during interaction. With regard to WM effects, the findings suggested that WM did not moderate the relationship between task complexity and the production knowledge development of the Spanish past subjunctive. However, high WM was significantly associated with the improvement of receptive knowledge in the simple group and in the FTF mode only.

In sum, these two studies reporting on WM and task complexity did not provide evidence for significant WM effects in language production (i.e., complexity, accuracy, or fluency) or in learning from recasts

in conjunction with cognitively complex tasks during FTF interaction. These findings are not in line with previous interaction studies that did show a positive effect for high WM in the noticing of feedback during interaction and L2 development. The key difference between these two studies and previous interaction studies is whether or not task complexity levels are manipulated following theoretically informed criteria to promote learning from recasts (Robinson, 2007a, 2007b, 2011).

Independent Measures of Learners' Noticing of Recasts in Task Complexity Research

Another methodological issue that is critical to address is how noticing is measured in these studies as this measure affects the ability to make claims on the moderating effects of learners' WM capacity. To date, Robinson's (2011) claim that task complexity impacts learners' noticing of linguistic forms has not been examined with direct measures of noticing. In previous interaction studies, noticing of feedback in the form of recasts has been informed by uptake (i.e., immediate response to recasts) or modified output (i.e., attempts to modify nontargetlike utterances using interactional feedback) after receiving feedback. This type of evidence, however, does not clearly account for learner-internal processes (Egi, 2007), and so it remains unclear whether learners are focusing on the corrective nature of the recast or on other conversational responses.

As rightfully argued by Kormos and Trebits (2011), by Norris (2010), and by Leow (2012), learners may divide their attention among aspects of task performance in ways that are not intended by researchers. There is thus a pressing need for task complexity research to implement direct measures of noticing of interactive feedback and target linguistic features. Task complexity research may benefit from adopting methods from interaction studies that tap into learner-internal processes, including (a) immediate cued recall—that is, asking students to immediately recall corrective feedback followed by a salient cue such as knocking (e.g., Bigelow, Delmas, Hansen, & Tarone, 2006; Egi, 2004; Philp, 2003); (b) concurrent think-alouds, in which learners express their thoughts as they process task input (e.g., Bowles, 2010; Gurzynski-Weiss, Al-Khalil, Baralt, & Leow, in press); (c) stimulated recall—that is, asking students what they were thinking at the time of receiving feedback based on video or audio stimuli (e.g., Egi, 2010; Gass & Mackey, 2000); and (d) eye-tracking technology (Smith, 2012). Each method has its strengths and weaknesses but should be used depending on the context of the study, its needs, and its participants (e.g., think-aloud protocols would not be possible with FTF task-based interaction). The current

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

study adopted the immediate cued recall method to measure learners' noticing of recasts, given its capacity to prevent memory decay per empirical findings in past research (Egi, 2004; see the Method section).

Purpose of the Study and Research Questions

Drawing on the ideas that (a) WM is considered to be among the most important individual difference factors moderating the noticing of corrective feedback in task-based interaction (e.g., Révész et al., 2011) and that (b) the cognition hypothesis claims that WM capacity will moderate learning to a greater degree during complex tasks, research is needed that explores the relationship between task complexity and WM. As such, the present study manipulated [+/– reasoning] demands (task complexity) and WM (task difficulty) and investigated how these variables interacted in the noticing of recasts targeting question formation and subsequent question development. Noticing of recasts was measured by immediate cued recall methods (a direct measure). Also, building on recent studies that included independent measures of task complexity (Baralt, 2010, 2014; Gilabert & Barón, 2013), the present study examined whether task complexity manipulation was in fact reflected in the participants' cognitive processes during task performance using stimulated recall. The study was guided by the following three research questions:

1. Independent measure of task complexity: Are different levels of task complexity reflected in learners' cognitive processes during task performance, as measured by stimulated recall?
2. Noticing: To what extent is noticing of recasts predicted by task complexity, WM, and their interaction?
3. Learning: To what extent is L2 question development predicted by task complexity, WM, and their interaction?

METHOD

Participants

The participants were 81 English language learners (41 females and 40 males) enrolled in an intensive English program (IEP), also known as an English-for-academic-purposes program, at a large public university in the United States. This IEP has five levels (high beginning to advanced), and students are assigned to a level on the basis of their performance on an in-house placement exam. The majority of students were enrolled full-time, which is equivalent to 18 hours of instruction per week.

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

All classes in the program are taught in English. The participants came from 15 different countries, ranged in age from 17 to 52 (M = 26.20), and had spent, on average, 6.8 years studying English (including instruction in their home countries). At the time of the study, all participants were enrolled at the intermediate level (i.e., Levels 3–4).

The interlocutors were two native English speakers who, at the time of the study, were in their 30s and were pursuing doctoral studies in applied linguistics. They each had more than 10 years of language teaching experience in North America and abroad, and both had training in task-based instruction and interaction-based research. They met three times with the principal investigator to discuss the research and to practice the treatment tasks, thereby ensuring consistency in treatment conditions (i.e., providing recasts and immediate cued recall).

Design

The study employed a pretest-posttest-delayed posttest design to examine the relationship between task complexity, WM, and English question development. The independent variables were task complexity and WM, and the dependent variables were (a) stimulated recall responses for testing the validity of task complexity manipulation, (b) noticing of recasts measured by immediate cued recall (Egi, 2004; Philp, 2003), and (c) question development based on the developmental stages for question formation (Pienemann and & Johnston, 1987). Participants were randomly assigned to either a simple or a complex group and carried out three tasks within their assigned complexity level with a NS interlocutor in a laboratory environment. Following Robinson's task complexity framework (2001a), [+/– reasoning demand] along resource-directing dimensions was used to manipulate two degrees of task complexity (see the Treatment Tasks section for more information on task complexity manipulation). To test the validity of task complexity manipulation through learners' cognitive processes (Research Question 1), a stimulated recall protocol was carried out immediately after the third treatment session. Stimulated recall is one of the most prominent methods used in SLA to gauge the cognitive and thought processes of learners during a task (see Gass & Mackey, 2000, for a review). During the stimulated recall session, learners were instructed to verbalize what their thought processes were at the time of the interaction episode, which was prompted by the video stimulus.

During each treatment task, the interlocutors provided recasts not only following erroneous question formation but also after other linguistic errors (e.g., tense or pronunciation) to distract participants from uncovering the target structure of the study. All recasts with the

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

target structure in the current study involved full questions. Immediate cued recall was used to measure noticing of recasts, assuming "what was already detected and entered in WM was available for immediate recall" (Philp, 2003, p. 109). The example in (1) provides one instance of the immediately cued recall cycle: The participant makes an error when forming a question, and the NS interlocutor recasts the target structure and knocks twice, which serves as the cue for the participant to repeat the last-heard utterance (i.e., the recast):

(1) Immediate cued recall for the recast
Learner: *Dormitory rooms.*
NS: *Yes*
Learner: *How many person we have to share with?*
NS: *How many people do we have to share with?* [2 knocks]
Learner: *How many people do we . . . have to share with?* [recall of the recast]
NS: *Two people*

The immediate cued recall method was chosen over modified output to measure noticing of recasts to directly measure the level of learners' attentional resources to target forms. This was to control for modified output, given that previous interaction studies have shown that modified output may directly impact L2 learning (e.g., McDonough, 2005). Also, it was methodologically imperative to balance the amount of modified output opportunities between the two task complexity groups to avoid any additional mediating variables, following Goo (2012).

Target Linguistic Forms

Question formation was selected as the target linguistic focus for several reasons. First, because the cognition hypothesis claims that increasing the cognitive demands of tasks results in the learning of developmentally more advanced forms, a structure with an empirically proven developmental sequence was considered appropriate (Robinson, 2007a, 2007b; Robinson & Ellis, 2008). Second, Robinson (2001a) claims that learners are expected to learn linguistic structures while carrying out cognitively more demanding tasks, and the required condition is that the linguistic targets should be relevant to how task complexity was manipulated (i.e., +/– reasoning). To make a decision (+reasoning condition) using the information from both the learner and researcher, the participants had to ask specific questions involving advanced question forms (i.e., *wh*-questions with inversion). The targeted forms (i.e., questions) were therefore inherent to the task demands; they are a fundamental linguistic structure needed for carrying on conversation.

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

The present study used Pienemann and Johnston's (1987) question developmental sequence, which is governed by processing mechanisms (see also empirical work by Loewen & Nabei, 2007; Mackey, 1999, 2006; Mackey & Philp, 1998; McDonough, 2005; McDonough & Mackey, 2006; Philp, 2003). Along this scale, learners' question stages progressively increase from Stage 1 to Stage 6. The current study examined learner development between Stage 3 (in which learners' questions have *do*-fronting and *wh*-fronting but no inversion, e.g., *where you went?*), Stage 4, (which is marked by *wh*-questions and a copula, e.g., *where were the teachers?*), and Stage 5 (in which learners' questions display inversions in *wh*-questions with both an auxiliary and a main verb, e.g., *how long did it take?*).

Materials

The materials used for this study included three treatment tasks, a running span WM test, and three oral production tests for the testing phases.

Treatment Tasks. A total of three 30-min, two-way collaborative information gap tasks following Ellis's (2003) criteria for tasks were designed based on authentic U.S.-based needs. The task topics included (a) traveling in the United States, (b) college life, and (c) cell phones. To investigate the impact of task complexity, a simple and a complex version of each task were designed. For the simple group, [–reasoning demand], participants exchanged information. For the complex group, [+reasoning demands], participants exchanged and evaluated the information on the basis of four predetermined criteria and selected the best travel destination, college, and cell phone. During the cell phone task, for example, participants in both groups had information about the phone plans of three separate companies (e.g., phone features and applications, billing plans and associated costs, and customer reviews). In the simple condition, participants simply exchanged that information (i.e., information gap) to introduce these different options to incoming students in their language programs. The activity was complete once each aspect had been discussed for the various cell phone options. In the complex condition, participants not only exchanged the information but also had to compare and evaluate it to make a decision about which cell phone plan to purchase (i.e., information gap + consensus). Under the complex condition, the activity was complete once both participants had reached an agreement. The same format was used for the other two topics.

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

Working Memory Test. Following Baddeley's (2003) WM model, the current study examined the ability to simultaneously process and store information as measured by an aural running span test. The running span test has been validated in many psychology studies (e.g., Broadway & Engle, 2010). It was delivered using E-Prime 2.0 (Schneider, Eschman, & Zuccolotto, 2012), a suite of applications for designing and conducting experiments in the field of psychology. The running span test was the preferred method for ESL learners, as it mitigated a potential confounding effect of variation in L2 proficiency on WM. For this test, participants heard a series of letters and were instructed to recall the last *n* items from lists that were *m* + *n* items long (Broadway & Engle, 2010). Span length was predetermined; however, participants were unaware of the length (e.g., the message "remember the last 3 letters" would appear on the screen, but the participant was not informed of the total number of letters in the series). The span length ranged from three to six, and there were six sets of each span, for a total of 108 letter items in the test. Following Broadway and Engle (2010), one point was assigned for each item correctly chosen in correct serial position "with respect to the set of the last *n* targets, not the whole *m* + *n* input sequence" (p. 565). Thus the highest possible total score on the WM test was 108.

Oral Production Tests for Questions. The participants' question development was measured on the basis of their performance on three oral production tasks. To elicit a range of question types from the participants, three types of oral production tasks were created: (a) an icebreaker, in which participants were given four statements (three truths and a lie) about the interlocutor, and they had to ask 20 questions to determine which one was the lie; (b) six short role plays, wherein the participant and interlocutor coconstructed a dialogue based on pictures and unique scenarios (e.g., Bill got a new job. Nancy is asking Bill about his new job.); and (c) five short interview scenarios, in which the participant was instructed to help the interlocutor practice for an interview (e.g., a scholarship interview for graduate school). Three versions of each task, with different pictures and scenarios, were designed and counterbalanced for the pretest, immediate posttest, and delayed posttest.

Procedure

As illustrated in Figure 1, in a 3-week period, participants completed a pretest, three collaborative tasks, a WM test, an immediate posttest, a stimulated recall session, a delayed posttest, and a questionnaire.

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

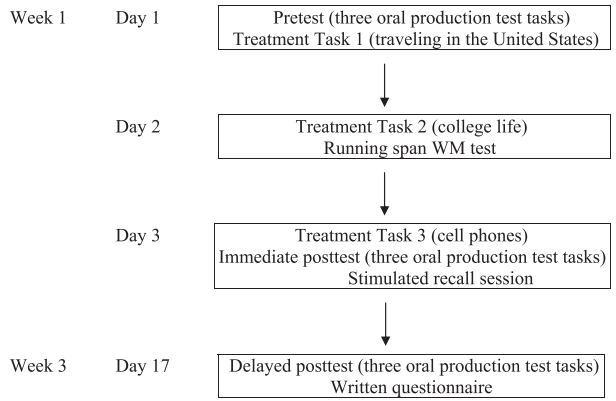

Figure 1. Procedure of the study.

Day 1 involved the pretest (an oral production test for questions) and Treatment Task 1 (traveling in the United States).[1] Participants carried out an oral production test that consisted of three sections with an interlocutor for 30 min. All treatment tasks, also carried out with an interlocutor in a dedicated lab, were audio and video recorded. A total of 30 min was set aside to carry out each treatment task. Day 2 included Treatment Task 2 (college life) and the running span WM test, which was administered on a personal computer in a sound-attenuated lab and took approximately 15 min to complete. Day 3 involved Treatment Task 3 (cell phones) and the immediate posttest (a different version of the pretest/oral production tasks). While learners performed the immediate posttest, the principal investigator prepared the stimuli for the stimulated recall session, during which the investigator would pause the video and ask the participant to describe what he or she was thinking. The average duration of all participants' stimulated recall sessions was 30 min. This methodology was employed to gauge learners' perspective on the construct of task complexity. Because the session was aimed at understanding original thoughts during task performance, participants were encouraged to comment on what they were thinking at the time of task performance, as opposed to what they were currently thinking (see Gass & Mackey, 2000, for more details; Polio, Gass, & Chapin, 2006). The researcher stopped the video to target at least four feedback episodes or when there was an indication of learners' mental effort during task performance (e.g., pauses before utterances). Participants were also encouraged to stop the video whenever they wanted to share their thoughts while carrying out the tasks. Finally, Day 4, completed two weeks after Day 3, involved the delayed posttest (the remaining version of the pretest) and a written exit questionnaire.

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

Data Coding and Analysis

The data were captured via digital recorder in all sessions (i.e., tests, treatments, and stimulated recall) as well as by digital camera for the treatment tasks, although the video was used only for the stimulated recall. To prepare the data for coding and analysis, two research assistants performed a verbatim transcription of the entire set of audio data. Working memory data were automatically scored by E-Prime (Schneider et al., 2012) using a partial-credit load scoring.

Task Performance. For the treatment task data, the number of recasts per task was calculated, and all questions produced were assigned stages per Pienemann and Johnston (1987). Additionally, as a measure of noticing of recasts, the participants' responses on the immediate cued recall were coded for recall of target question forms using a modified version of Philp's (2003) categories: (a) full repetition (i.e., repetition of the recast utterance including the target structure; see the example in [1]); (b) partial repetition (i.e., repetition not featuring the targetlike question forms provided in recasts), as shown subsequently in the following example in (2); and (c) no repetition (i.e., failure to repeat the recast).

(2) Partial repetition without the target structure
Learner: *Why does stadium isn't round?*
Researcher: *Why isn't the stadium round?* [2 knocks]
Learner: *Why this stadium is not round?*

Oral Production Tests. To ascertain participants' L2 question development, individual questions produced in the oral production tests (i.e., pretest, posttest, and delayed posttest) were coded for stage according to Pienemann and Johnston's (1987) developmental sequences. When identifying learners' stages on each test, a conservative emergence criterion was used: Each learner was assigned to the highest level on the scale at which the learner produced two unique questions in a minimum of two production tasks out of three. This was similar to the criterion used in other SLA studies involving question development (e.g., Mackey & Philp, 1998; McDonough, 2005; McDonough & Mackey, 2006; Philp, 2003; Spada & Lightbown, 1993). After determining each learner's question stage on each test, learners were then categorized as *developed* or *not developed*, with development operationalized as a movement from an initial stage on the pretest to a higher stage on the posttest and maintenance of that stage on the delayed posttest. Learners who did not advance to a higher stage on one or both posttests were classified as *not developed*.

Stimulated Recall Data. Following previous oral feedback studies (e.g., Egi, 2010), the stimulated recall data were analyzed through thematic

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

analysis, "a method for identifying, analyzing and reporting patterns (themes) within data" (Braun & Clarke, 2006, p. 79). The researchers adopted a theoretical thematic analysis such that the themes were driven by analytic interest—namely, how task complexity manipulation impacted the participants' cognitive processes. Participant comments were thus analyzed for themes related to the ways task complexity impacted cognitive processes by (a) grouping together comments addressing the same theme and then (b) tallying the number of participants who addressed the theme.

Working Memory Test. The running span tests were scored in E-Prime (Schneider et al., 2012) using a partial-credit load scoring according to Broadway and Engle (2010). For example, if the last four items were reported from an input sequence that was four items long (e.g., J P K T), a response of only "P K T" would receive 3 points.

Intercoder Reliability and Statistical Analyses. Intercoder reliability was established by a second rater (one of the coauthors), who coded 25% of the oral production, recast, and stimulated recall data sets. The percentage agreement was 94% for the oral production tasks, 95% for the treatment tasks, and 92% for the stimulated recall sessions. Disagreements in coding were resolved through discussion.

As for the statistical analyses used in the study, the first research question was answered based on descriptive statistics and qualitative analyses from the participants' stimulated recall sessions. Standard multiple regression analyses were utilized in answering the second research question, which examined the extent to which task complexity and WM predict noticing of recasts targeting questions. Finally, logistic regression analysis was used to answer the third research question, which inquired into the relationship between task complexity and WM in relation to English question development.

RESULTS

Preliminary Results: Recasts, Working Memory, and the Pretest

Prior to answering the three research questions, participant eligibility and group comparability were examined by analyzing the number of recasts received by each group, WM scores between the two task groups, and their question stage measured at the pretest. First, the number of recasts during task performance data was determined for each question stage to ensure that both groups received similar amounts of recasts. The number of recasts was then compared between the two complexity groupings for the three tasks.

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

In looking at recasts, it was found that the simple group received an average of 44.93 recasts during the three tasks, whereas the complex group received an average of 39.00. The independent *t* test did not report significant differences between the groups in terms of the number of recasts targeting all question stages, $t = 1.74$, $p = .06$. Given that recasts targeting Stage 4 and 5 questions could potentially lead to a higher stage classification on posttests, the amount of recasts with Stage 4 and 5 questions were also compared ($M = 36.63$ for the simple group vs. $M = 31.54$ for the complex group). The results again showed that there was no significant difference in the number of recasts targeting Stage 4 and 5 questions, $t = 1.69$, $p = .13$.

With regard to WM, the mean score by the simple group was 69.56 ($SD = 17.39$), whereas the mean score by the complex group was 63.95 ($SD = 14.87$; the total possible score for the running span WM test was 108). Independent *t* test results suggest that the two groups were comparable in terms of their WM capacity, as there was no significant group difference between the two groups on their running span scores, $t = 1.56$, $p = .12$.

Finally, pretest results were analyzed for learners' question stage to ensure participant eligibility and group comparability. We used the following criteria for the inclusion of participants in the final analyses: (a) no Stage 5 learners at the time of the pretest are included in the analysis and (b) there is no preexisting difference between the two groups in their ability to produce advanced question stages. The results showed a similar number of students at different stages between the simple and complex groups: 25 versus 27 for Stage 3, seven versus nine for Stage 4, and seven versus six for Stage 5, respectively. Learners who were at Stage 5 on the pretest were excluded because they had already acquired the target structure. Therefore, only the results from the participants at Stage 3 or 4 (simple: $n = 34$, complex: $n = 34$) could be included in our subsequent analyses. These are reported in the following sections ($N = 68$).

Research Question 1: Task Complexity and Cognitive Processes

The first research question asked how task complexity manipulation impacted the participants' cognitive processes during task performance. This question also addressed the participants' perspective to determine whether task complexity was manipulated as intended by the researchers. As mentioned in the Method section, the learners participated in a stimulated recall session with the principal investigator on Day 3, after having completed the third task and the immediate posttest. They were asked to describe what they were thinking at the times at which the video was stopped. Additionally, the learners were instructed to pause the video whenever they wanted to share

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

what they were thinking at that time. Three major themes emerged from the stimulated recall data: comparisons/evaluations, language use, and task procedure.

The first of these (i.e., comparisons/evaluations) included learners' comments on the information exchanged during task performance, such as in (3).

(3) *So here I was thinking about best option for us. Because we have to buy a phone and we have the three option and I am interest how international calls because I wanna call my family. Some phones, they don't have international calls so I thinking about good or best option.* (Complex, ID 65)

Two types of comparisons were identified: personal experiences (i.e., comparing the information from the task input to their own life) and options within tasks (i.e., comparing information among options presented in the task input).

The second theme, language use, referred to comments regarding language production, as in (4).

(4) *Ask question. I have this problem. Sometimes I ask question, I ask verb before. I always do this.* (Complex, ID 56)

Finally, task procedure related to comments about task management, such as in (5).

(5) *I look for something helpful for the paper for the example Metro PCS, Verizon, you can say something about.* (Simple, ID 43)

Comments that did not contain specific content were categorized as *other*, as in (6).

(6) *I think nothing.* (Simple, ID 2)

The first category, comparisons/evaluations, reflected cognitive processes that distinguish complex tasks from simple tasks (i.e., reasoning demands). Therefore, the number of learners who described such processes during the stimulated recall session was compared between the simple and the complex groups. A total of 16 of 34 learners (47%) in the simple group and 24 out of 34 (71%) in the complex group mentioned that they compared personal experiences with task situations. Additionally, 27 learners (79%) in the complex group stated that they were evaluating cell phone options provided in the task input while carrying out the task; only five learners (15%) in the simple group described such cognitive processes. These findings suggest that, during task performance,

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

more learners from the complex group used reasoning processing, as intended, compared to those from the simple group.

Research Question 2: Task Complexity, Working Memory, and Noticing of Recasts

Research Question 2 examined the extent to which task complexity and WM, individually or collectively, predicted learners' noticing of recasts for the target structure (operationalized as immediate cued recall of recasts). As discussed in the data-coding section, immediate cued recall was coded as either full recall, partial recall, or no recall. Table 1 shows the raw frequency and the proportion scores of recall of recasts for all question forms. The descriptive statistics for learners' noticing of recasts focusing on developmentally advanced questions (Stages 4 and 5) are provided separately. Because each group had a different total number of recasts for cued recall, proportion scores of correct recall of recasts were used for multiple regression analyses.

The descriptive statistics showed a similar pattern between the two groups: Both groups recalled about 80% of the recasts correctly, with about 14%–16% of recasts partially recalled. To examine whether task complexity, WM, or their interaction predicted the amount of noticing of recasts, a multiple regression was performed on the percentage of full recall of recasts as the dependent variable. Working memory scores, task complexity level, and their interaction were computed as independent variables. The multiple regression analysis resulted in a multiple correlation of $R = .331$, $F = 2.63$, $p = .05$. The coefficient of multiple determination was $R^2 = .11$, indicating that 11% of the variability in noticing of recasts was explained by task complexity, WM, and the interaction between the two variables. Only WM was found to contribute significantly to the amount of learners' noticing of recasts targeting all

Table 1. Raw frequency and the proportion scores of recall of recasts

	Simple task (n = 34)						Complex task (n = 34)					
	Full		Partial		No		Full		Partial		No	
Linguistic targets	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All questions	35.62	11.99	7.06	5.88	2.91	3.32	31.47	9.78	5.79	4.75	2.68	3.93
	.79	.13	.16	.13	.06	.06	.80	.14	.14	.10	.06	.08
Stages 4 & 5	28.00	9.39	6.21	5.16	2.18	2.59	25.26	8.42	5.59	4.64	1.79	2.43
	.78	.13	.17	.13	.05	.05	.79	.14	.16	.11	.05	.07

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

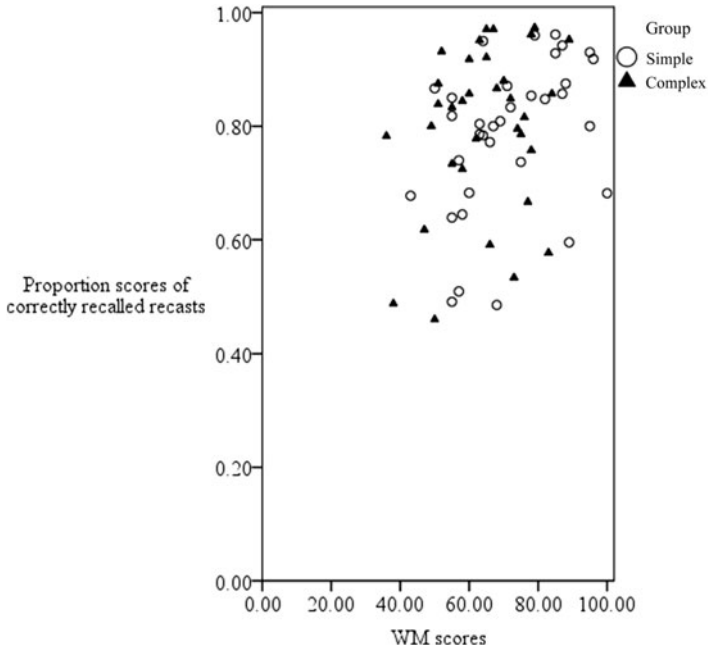

Figure 2. Task complexity, WM, and noticing of recasts.

questions: β = .33, t = 2.68, p = .009. The squared semipartial (sr^2 = .10) indicated that about 10% of the variance in noticing of recasts was uniquely predictable from WM. Figure 2 visually demonstrates the trend of the relationship between task complexity, WM, and learners' noticing of recasts.

Figure 2 demonstrates that about 72% of learners scored higher than .75, especially those learners who were in the middle and high range of WM scores (60 and above) and who obtained high proportion scores of correctly recalled recasts. In other words, this finding suggests that learners in both groups were able to recall questions provided through recasts correctly, regardless of the complexity level of the task they were performing. Next, we ran a follow-up analysis to examine the recall of recasts for Stage 4 and 5 questions only, in light of Robinson's (2011) claim that task complexity facilitates the acquisition of developmentally advanced forms. The descriptive statistics showed that both the complex group and the simple group correctly recalled about 78%–79% of recasts targeting Stage 4 and 5 questions. The multiple regression analysis resulted in a multiple correlation of R = .48, F = 6.32, p = .001. The coefficient of multiple determination was R^2 = .23, which suggests that 23% of the variability in noticing of recasts targeting advanced question forms was explained by task complexity, WM, and their

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

interaction. However, only WM was again found to be a significant predictor of learners' noticing of recasts targeting Stage 4 and 5 questions: $\beta = .47$, $t = 4.08$, $p < .001$. On the basis of the squared semipartial ($sr^2 = .21$), it was found that about 21% of the variance in noticing of recasts was uniquely predicted by WM. This implies that WM capacity is significantly associated with learners' ability to recall advanced question forms provided through recasts, irrespective of task complexity level.

Research Question 3: Task Complexity, WM, and Question Development

The final research question addressed the extent to which task complexity and WM during interaction individually or collectively predict question development. As discussed previously, question development was operationalized on the basis of question stage changes between the pretest and the two posttests. To be identified as developed, the learners had to maintain stage increases between the pretest and the two posttests. The results showed that a total of 20 and 21 learners from the simple and the complex groups, respectively, advanced to a higher question stage.

To find the most appropriate model to describe the relationship among task complexity, WM, and learners' question development (i.e., developed vs. not developed), a logistic regression was conducted. The results of the logistic regression revealed that the model was significant, $\chi^2(3, 68) = 23.29$, $p < .001$, indicating that the predictors, as a set, reliably distinguished between learners who advanced to a higher stage of questions and those who did not. The model was also evaluated on the basis of its goodness of fit and its success at predicting group membership. According to Nagelkerke R^2, 39% of the variance in the dependent variable (i.e., question development) was accounted for by the model. For group membership, the model successfully predicted about 56% of no development and 73% of development cases, with an overall success rate of 66%. Three independent variables in the model were analyzed to determine the strength of their relationship to question development. Tests of significance indicated that WM was the only significant predictor of question development, Wald statistics = 14.78, Exp(B) = 1.12, $p < .001$.

Robinson (2011) claimed that learners' cognitive individual differences will mediate the role of task complexity in language development. Even though task complexity was not found to be one of the significant predictors for question development, we examined how learners with different WM capacities might benefit from carrying out tasks with varying complexity levels. To do so, a scatterplot was created with individual scores (see Figure 3). The x-axis shows learners' WM scores, and

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

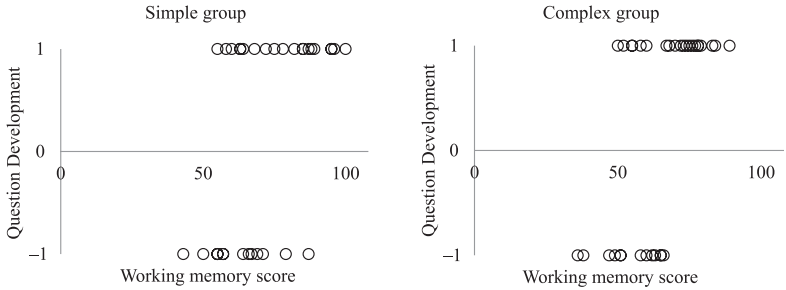

Figure 3. Question development by WM scores in simple and complex groups; –1 = no development, and 1 = development.

the y-axis indicates their question development (–1 = no development, 1 = development).

As shown in Figure 3, question development seems to be associated with higher WM capacity learners, especially for those who performed cognitively demanding tasks. More specifically, out of 15 learners from the complex group who had higher WM capacity (i.e., WM scores higher than the mean and median of 67), 13 learners (87%) advanced to a higher stage, whereas only 11 out of 19 learners (58%) in the complex group who had lower WM capacity (i.e., WM scores lower than the mean and median of 67) showed any question development. The benefits of higher WM capacity among the simple group on question development were not as detectable. For instance, out of 19 learners who had higher WM capacity (i.e., WM scores higher than the mean and median of 67), 12 learners (63%) showed development. Among the 15 learners who had lower WM capacity (i.e., WM scores lower than the mean and median of 67) in the simple group, eight learners (53%) showed improvement in their question formation.

In summary, the results suggest that our operationalization of task complexity was valid. Learners who performed the more demanding cognitive tasks were pushed to engage in higher cognitive processes, such as simultaneous comparison of information presented on task input while making a decision. Our results also found that WM capacity was the only significant predictor of the amount of noticing of recasts and question development. More specifically, the data suggested that higher WM was beneficial for interaction-driven language learning (i.e., question development), especially with more complex tasks.

DISCUSSION

Drawing on Robinson's cognition hypothesis, the current study examined how task complexity, WM, or their interaction predicts learners'

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

noticing of recasts during learner-learner interaction. It further examined whether their question development as a result of interactive tasks can be predicted by task complexity, WM, or their interaction. Before examining these questions, it was necessary to first validate the construct of task complexity in light of learners' cognitive processes. Learners demonstrated two types of comparisons/evaluation processes: comparing the given situation in the task input to their own experiences (i.e., personal evaluations) and evaluating options in the task input by using reasoning processes (i.e., options within tasks). The results showed that learners, regardless of complexity grouping, tended to engage in personal evaluations, which supports the authenticity of tasks; however, it was only in the complex group that we observed the evaluations of the options in the task input, a process directly related to task complexity manipulation. In sum, the stimulated recall data were taken as evidence that the learners in the complex group were involved in the expected reasoning processes, whereas the simple group showed little reasoning during performance, suggesting that the tasks were implemented as intended by the researchers.

The second research question tested Robinson's cognition hypothesis by examining the extent to which task complexity, WM, or the interaction between task complexity and WM is associated with the amount of learners' noticing of question forms made salient during the provision of reactive focus-on-form techniques such as recasts. Regarding the role of WM as a task difficulty factor (i.e., individual differences), we examined Robinson's (2011) hypothesis that the benefits of high WM will be more apparent during a complex version of a task than during a simple version.

Overall, learners correctly recalled approximately an average of 80% of the recasts. On the basis of the multiple regression analysis, we found that only WM capacity was a significant predictor of the amount of noticing of recasts involving questions. Because Robinson (2011) made specific claims about the benefits of carrying out more complex tasks in regard to the acquisition of developmentally advanced forms, which corresponds to the conceptual demand of a task, recasts targeting Stages 4 and 5 questions were separately analyzed (e.g., Robinson & Ellis, 2008). Similarly, the results showed that task complexity was not associated with the amount of correct recall of developmentally advanced questions. These results echo those of Révész et al. (2011), who found that, during task-based interaction, more complex tasks did not lead to a significantly higher rate of uptake of recasts compared to simple tasks. Révész et al. posited that having to process additional task input (i.e., photos) could also result in learners dividing their attention, thereby allocating fewer attentional resources to incoming linguistic information via recasts. In the present study, the learners in both task complexity groupings received the same task input and the same amount

of control in attention to task input during interaction. Thus, it was expected that the differences in the deployment of attentional resources would be mainly from the interaction of WM and task complexity during the storing and processing of the aural and written input. However, our findings suggested that WM was the only significant predictor of the amount of learners' noticing of recasts. Such findings were in line with the results of several interaction studies that suggested the benefits of high WM on noticing of recasts and/or learning outcome (e.g., Mackey et al., 2002; Révész, 2012).

The results relating to WM capacity being the only significant predictor of noticing of recasts could be accounted for by the way noticing of recasts was operationalized—namely, through correct immediate cued recall (Philp, 2003). The immediate cued recall process captures mere attention to information in the oral L2 input during interaction that is being stored briefly in WM. However, the extent to which the target structures were cognitively processed during interaction, as the increasing reasoning demands of the tasks directed learners' attentional and memory resources to them (e.g., noticing the differences between the original questions that they produced and the recasts), cannot be confirmed through immediate cued recall (i.e., repeating recasts without providing negative evidence). The role of task complexity level and the interaction effects between task complexity and WM on learners' attention to form during interaction may be better determined by measuring different levels of noticing or attention (e.g., detection or detection plus rehearsal; see Leow's, 2012, work on different levels of awareness). For instance, as shown in the current study, both simple and complex conditions may equally encourage learners to detect and rehearse advanced questions, but whether learners are explicitly conscious of the differences between their interlanguage and target question forms (attention plus awareness, per Schmidt's, 1990, definition of noticing) may differ between the two complexity conditions. At this juncture, the current study directly measures learners' noticing of recasts by using immediate offline response prompts to avoid memory decay; however, it is important to further examine this question. From a theoretical and methodological point of view, the investigation of different levels of noticing of linguistic codes using various data collection methods such as stimulated recall and eye-tracking is warranted in future studies when testing predictions related to the noticing of feedback (see Robinson, Mackey, Gass, & Schmidt, 2012, for future directions).

In light of the observed trend in the data—namely, that higher WM learners from both groups recalled recasts (i.e., full recall) at a similar rate—we believe that WM appears to play an important role during interaction. These findings are in line with previous interaction research. Recently, a growing number of interaction studies have supported the idea that WM (particularly the central executive function) plays a greater

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

role in directing learners' attentional resources to L2 linguistic codes during recasts (e.g., Goo, 2012; Mackey et al., 2002). In the current study, however, it is important to note that WM explained only 11% and 23% of the variability in the noticing of recasts targeting all questions and those targeting developmentally advanced question forms (i.e., Stages 4 and 5), respectively. As a result, this leaves us with certain questions about the role of other variables, such as the accuracy of questions produced by learners during task performance and the impact of other learner factors such as anxiety in the noticing of recasts.

The third research question focused on question development. A growing number of studies have suggested the positive effects of more complex tasks in language development during FTF interaction contexts (Baralt, 2010; Kim, 2012). However, the current study found that WM was the only significant predictor of question development during interactional tasks with provision of recasts. These results again support the findings of some previous interaction studies that suggest the benefits of high WM on the learning of morphosyntactic features during interactional tasks while receiving feedback (e.g., Mackey & Sachs, 2012; Révész, 2012).

One way to account for a lack of task complexity effects is the high level of recall of recasts regardless of the level of task complexity. More specifically, both task groups were able to correctly recall about 80% of the recasts targeting questions. These findings echo McDonough (2005), who showed that the amount of learners' modified output during interaction (i.e., correctly recalling recasts, in the current study) was a stronger indicator of English question development. In the current study, because both the simple and complex groups were provided with recasts involving the target structure, the role of task complexity needs to be interpreted in this particular condition. To examine task complexity effects independently from other factors (e.g., the provision of recasts), future studies may want to include additional groups who performed tasks without feedback (Révész, 2009).[2]

The current study provided additional insights into how WM may mediate the role of task complexity in question development (Baralt, 2010). Although there was no significant task complexity effects on the number of students who advanced to a higher question stage (20 vs. 21 of 34, respectively), post hoc analyses suggested that learners in the complex group with higher WM showed a noticeable degree of question development (13 out of 15 learners), whereas only eight of the 19 learners with lower WM from the complex group advanced to a higher stage. This suggests that learners with lower WM capacity may not benefit as much from carrying out cognitively demanding tasks compared to those with higher WM capacity, especially when recasts are provided. As a result, the findings suggest that the role of task complexity may be complemented by individual learner abilities such as WM (Robinson, 2011).

What remains perplexing is the finding that seven learners from the simple group with higher WM capacity (i.e., those who scored higher than the mean and median of 67) did not show question development, despite having approximately 80% of correct recasts recall rates. This suggests that immediate recall of cued recasts may not be directly measuring what learners actually internalized in their interlanguage system on the basis of noticing the gap between the original utterance and the recasts. These findings support Baralt (2010), who suggested that awareness at the level of understanding (i.e., tasks with a high level of processing such as hypothesis testing and rule formation; Leow, 2012) may be more strongly related to L2 development. Furthermore, although the current study measured the learning of question forms using the production of target forms, the role of WM in interaction-driven, task-based instruction may be different in receptive recognition of the correct target forms versus the production of target forms, as noted in Baralt (2010) as well as Révész (2012). Because the learners were not given learner-initiated modified output opportunities during interaction, the degree of the noticing of recasts may impact the recognition knowledge of questions more than the production knowledge. In sum, further exploration of different constructs related to noticing and awareness is warranted, as is the inclusion of measures that address both recognition and production knowledge.

The current study offers important theoretical and methodological implications for the field of task-based instruction. First, building on previous cognitive-interaction and task complexity research, the current study addressed task complexity and individual learner variables along multiple sources of Robinson's triadic componential framework and showed a complex relationship among these variables. The findings suggest the importance of taking individual cognitive ability such as WM into consideration in both task complexity and interaction research. Furthermore, the potentially competing roles of task complexity and WM in interaction-driven language learning observed in this study further suggest a need for closely examining synergetic effects among variables along complexity and difficulty axes. Finally, the study also suggests a need for specifying the level of noticing of linguistic forms in research studies when testing the cognition hypothesis.

From a methodological viewpoint, the current study's design was noteworthy in that it used stimulated recall data to test the validity of task complexity manipulation from learners' perspectives of their cognitive processes. Moreover, analyzing recasts for question stages and carrying out further analysis with developmentally more advanced questions allowed us to test Robinson's hypothesis, which predicts that task complexity would lead to the acquisition of developmentally more advanced forms (Robinson & Ellis, 2008).

The current study had some limitations that should be acknowledged and addressed in future research. First, this investigation tested two levels of task complexity. However, as some recent studies have argued (Kim, 2012), the construct of task complexity may be more appropriately operationalized on a continuum. Doing so can especially accommodate the different variables of task complexity, task conditions, and task difficulty factors present in classrooms. Thus future studies should include multiple levels of task complexity operationalizedwith a variety of task complexity variables other than [+/– reasoning demands]. Second, the current study measured noticing of recasts using immediate cued recall of recasts, which did not allow for the examination of the amount of uptake or modified output initiated by learners during natural conversation. As discussed previously, different levels of noticing and awareness—which can be observed through other data sources, such as comments made during the task, hypothesis testing, and rule formation (e.g., Baralt, 2010)—need to be explored in future studies (Robinson et al., 2012). Another concern relates to the construct of task complexity and the potential role of immediate cued recall of recasts in learners' posttests results. As an anonymous reviewer pointed out, learners' successful repetition of recasts indicates that they had modified output opportunities, which provide additional learning opportunities. However, because there was no significant difference between the two task groups in the amount of noticing of recasts, we would argue that both groups received the same amount of modified output opportunities. Nevertheless, the posttest results in the current study need to be interpreted with some caution given that the cued recall procedures may have confounded learners' performance on the posttests.

To examine learners' cognitive processes and confirm the construct of task complexity, the current study carried out stimulated recall. Because of the learners' diverse first language backgrounds, this was done in English. Although the participants were high-intermediate learners, conducting this in their nonnative language may have hindered their ability to describe their thinking processes. This is, arguably, another limitation to the study. Finally, with regards to language development, we focused only on question formation. Future studies are warranted that examine diverse linguistic forms (pronunciation, vocabulary, etc.).

CONCLUSION

The current study examined the relationship between task complexity, WM, noticing of recasts targeting questions, and learners' question development. It provides a useful step toward greater explanatory validity by

https://doi.org/10.1017/S0272263114000618 Published online by Cambridge University Press

using stimulated recall methods to determine whether or not the intended task complexity did indeed create different levels of cognitive processes in tasks. Additionally, detailed analysis of immediate cued recall with different question stages provided insights into learners' attentional allocation during task-based interaction depending on WM. Furthermore, although the current study did not support the cognition hypothesis in terms of the independent role of task complexity in the acquisition of developmentally advanced forms, the findings of post hoc analyses suggest that WM might mediate question development, especially during more complex tasks. Over the last few years, an increasing number of studies have tested various predictions of the cognition hypothesis and have provided useful insights into task design in task-based instruction. One of the ultimate goals of the cognition hypothesis is to advance task-based instruction in various teaching and learning contexts. Building on the findings of the previous studies, future research needs to explore interaction effects between variables presented in the triadic componential framework in various instructional contexts.

Received 9 June 2013
Accepted 24 April 2014
Final Version Received 11 July 2014

NOTES

1. To use the same task (i.e., cell phones) for the stimulated recall session, the order of the three tasks was not counterbalanced. However, the order of the oral production tests for question development was counterbalanced.

2. One anonymous reviewer mentioned that the cued recall procedures are likely to have confounded participants' performance on the posttest. We agree with the reviewer's concern and would like to suggest that the learning outcome of this study was based on task performance with the provision of recasts as well as the recall opportunities of recasts.

REFERENCES

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*, 417–423.

Baddeley, A. D. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*, 829–839.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (Vol. 8, pp. 47–89). New York, NY: Academic Press.

Baralt, M. (2010). *Task complexity, the cognition hypothesis and interaction in CMC and FTF environments* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.

Baralt, M. (2013). The impact of cognitive complexity on feedback efficacy during online versus face-to-face interactive tasks. *Studies in Second Language Acquisition*, *35*, 689–725.

Baralt, M. (2014). Task complexity and task sequencing in traditional versus online classes. In M. Baralt, R. Gilabert, & P. Robinson (Eds.), *Task sequencing and instructed second language learning*. London, UK: Bloomsbury.

Baralt, M., Gilabert, R., & Robinson, P. (2014). *Task sequencing and instructed second language learning*. London, UK: Bloomsbury.

Bigelow, M., Delmas, R., Hansen, K., & Tarone, E. (2006). Literacy and the processing of oral recasts in SLA. *TESOL Quarterly*, *40*, 665–689.

Bowles, M. (2010). *The think-aloud controversy in language acquisition research*. New York, NY: Routledge.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*, 77–101.

Broadway, J. M., & Engle, R. W. (2010). Validating running memory span: Measurement of working memory capacity and links with fluid intelligence. *Behavior Research Methods*, *42*, 563–570.

Conway, A. R., Jarrold, C. E., Kane, M. J., Miyake, A., & Towse, J. N. (2007). Variation in working memory: An introduction. In A. R. Conway, C. E. Jarrold, M. J. Kane, A. Miyake, & J. N. Towse (Eds.), *Variation in working memory* (pp. 3–18). Oxford, UK: Oxford University Press.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*, 450–466.

Egi, T. (2004). Verbal reports, noticing, and SLA research. *Language Awareness*, *13*, 243–264.

Egi, T. (2007). Interpreting recasts as linguistic evidence. *Studies in Second Language Acquisition*, *29*, 511–537.

Egi, T. (2010). Uptake, modified output, and learner perceptions of recasts: Learner responses as language awareness. *Modern Language Journal*, *94*, 1–21.

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

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

Gass, S. M., & Mackey, A. (2007). Input, interaction and output in second language acquisition. In J. Williams & B. VanPatten (Eds.), *Theories in second language acquisition* (pp. 175–199). Mahwah, NJ: Erlbaum.

Gilabert, R., & Barón, J. (2013). The impact of increasing task complexity on L2 pragmatic moves. In K. McDonough & A. Mackey (Eds.), *Second language interaction in diverse contexts* (pp. 45–70). Amsterdam, the Netherlands: Benjamins.

Gilabert, R., Barón, J., & Llanes, À. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics in Language Teaching*, *47*, 367–395.

Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. *Studies in Second Language Acquisition*, *34*, 445–474.

Gurzynski-Weiss, L., Al-Khalil, M., Baralt, M., & Leow, R. (in press). The roles of type of feedback and type of linguistic item on L2 awareness in computer mediated communication. In R. Leow, M. Baralt, & L. Cerezo (Eds.), *Technology and second/foreign language learning: A psycholinguistic approach*. Berlin, Germany: de Gruyter.

Jackson, D. O., & Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, *63*, 330–367.

Kim, Y. (2009). The effects of task complexity on learner-learner interaction. *System*, *37*, 254–268.

Kim, Y. (2012). Task complexity, learning opportunities, and Korean EFL learners' question development. *Studies in Second Language Acquisition*, *34*, 627–658.

Kim, Y., & Tracy-Ventura, N. (2011). Task complexity, language anxiety and the development of past tense. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 287–306). Amsterdam, the Netherlands: Benjamins.

Kormos, J., & Trebits, A. (2011). Working memory capacity and narrative task performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cogni-*

tion Hypothesis of language learning and performance (pp. 267–286). Amsterdam, the Netherlands: Benjamins.
Leow, R. (2012). Explicit and implicit learning in the L2 classroom: What does the research suggest? *European Journal of Applied Linguistics and TEFL*, *2*, 1–14.
Li, S. (2013). The interactions between the effects of implicit and explicit feedback and individual differences in language analytic ability and working memory. *Modern Language Journal*, *97*, 634–654.
Loewen, S., & Nabei, T. (2007). Measuring the effects of oral corrective feedback on L2 knowledge. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 361–377). Oxford, UK: Oxford University Press.
Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. C. Ritchie & T. K. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). San Diego, CA: Academic Press.
Mackey, A. (1999). Input, interaction, and second language development: An empirical study of question formation in ESL. *Studies in Second Language Acquisition*, *21*, 557–587.
Mackey, A. (2006). Feedback, noticing and second language development: An empirical study of L2 classroom interaction. *Applied Linguistics*, *27*, 405–430.
Mackey, A., Abbuhl, R., & Gass, S. (2012). Interactionist approach. In S. Gass & A. Mackey (Eds.), *The Routledge handbook of second language acquisition* (pp. 7–23). New York, NY: Routledge.
Mackey, A., Adams, R., Stafford, C., & Winke, P. (2010). Exploring the relationship between modified output and working memory capacity. *Language Learning*, *60*, 501–533.
Mackey, A., & Philp, J. (1998). Conversational interaction and second language development: Recasts, responses, and red herrings? *Modern Language Journal*, *82*, 338–356.
Mackey, A., Philp, J., Fujii, A., Egi, T., & Tatsumi, T. (2002). Individual differences in working memory, noticing of interactional feedback and L2 development. In P. Robinson (Ed.), *Individual differences and instructed language learning* (pp. 181–208). Amsterdam, the Netherlands: Benjamins.
Mackey, A., & Sachs, R. (2012). Older learners in SLA research: A first look at working memory, feedback, and L2 development. *Language Learning*, *62*, 704–740.
McDonough, K. (2005). Identifying the impact of negative feedback and learners' responses on ESL question development. *Studies in Second Language Acquisition*, *27*, 79–103.
McDonough, K., & Mackey, A. (2006). Responses to recasts: Repetitions, primed production, and linguistic development. *Language Learning*, *56*, 693–720.
Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. F. Healy & J. Lyle E. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Mahwah, NJ: Erlbaum.
Norris, J. M. (2010, September). *Understanding instructed SLA: Constructs, contexts, and consequences*. Paper presented at the annual conference of the European Second Language Association (EUROSLA), Reggio Emilia, Italy.
Nuevo, A. (2006). *Task complexity and interaction: L2 learning opportunities and development* (Unpublished doctoral dissertation). Georgetown University, Washington, DC.
O'Brien, I., Segalowitz, N., Collentine, J., & Freed, B. (2006). Phonological memory and lexical, narrative, and grammatical skills in second-language oral production by adult learners. *Applied Psycholinguistics*, *27*, 377–402.
Philp, J. (2003). Constraints on "noticing the gap": Nonnative speakers' noticing of recasts in NS-NNS interaction. *Studies in Second Language Acquisition*, *25*, 99–126.
Pienemann, M., & Johnston, M. (1987). Factors influencing the development of language proficiency. In D. Nunan (Ed.), *Applying second language acquisition research* (pp. 45–141). Adelaide, Australia: National Curriculum Resource Centre, Adult Migrant Education Program.
Polio, C., Gass, S., & Chapin, L. (2006). Using stimulated recall to investigate native speaker perceptions in native-nonnative speaker interaction. *Studies in Second Language Acquisition*, *28*, 237–267.

Révész, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition*, *31*, 437–470.
Révész, A. (2011). Task complexity, focus on L2 constructions, and individual differences: A classroom-based study. *Modern Language Journal*, *95*(Suppl. 1), 162–181.
Révész, A. (2012). Working memory and the observed effectiveness of recasts on different L2 outcome measures. *Language Learning*, *62*, 93–132.
Révész, A. (2014). Towards a fuller assessment of cognitive models of task-based learning: Investigating task-generated cognitive demands and processes. *Applied Linguistics*, *35*, 87–92.
Révész, A., & Gilabert, R. (Chairs). (2013, March). *Methodological advances in TBLT research: Measurement of task demands and processes*. Colloquium conducted at the annual meeting of the American Association of Applied Linguistics, Dallas, TX.
Révész, A., Sachs, R., & Mackey, A. (2011). Task complexity, uptake of recasts, and second language development. In P. Robinson (Ed.), *Researching task complexity: Task demands, task-based language learning and performance* (pp. 203–238). Amsterdam, the Netherlands: Benjamins.
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). New York, NY: Cambridge University Press.
Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57.
Robinson, P. (2003). The cognition hypothesis, task design, and adult task-based language learning. *Second Language Studies*, *21*, 45–105.
Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics in Language Teaching*, *43*, 1–32.
Robinson, P. (2007a). Criteria for classifying and sequencing pedagogic tasks. In M. del Pilar García Mayo (Ed.), *Investigating tasks in formal language learning* (pp. 7–26). Clevedon, UK: Multilingual Matters.
Robinson, P. (2007b). 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 in Language Teaching*, *45*, 193–213.
Robinson, P. (2011). Second language task complexity, the cognition hypothesis, language learning, and performance. In P. Robinson (Ed.), *Second language task complexity: Researching the Cognition Hypothesis of language learning and performance* (pp. 3–38). Amsterdam, the Netherlands: Benjamins.
Robinson, P., & Ellis, N. C. (2008). An introduction to cognitive linguistics, second language acquisition, and language instruction. In P. Robinson & N. C. Ellis (Eds.), *Handbook of cognitive linguistics and second language acquisition* (pp. 2–24). New York, NY: Routledge.
Robinson, P., & Gilabert, R. (2007). Task complexity, the cognition hypothesis and second language learning and performance. *International Review of Applied Linguistics in Language Teaching*, *45*, 161–176.
Robinson, P., Mackey, A., Gass, S., & Schmidt, R. (2012). Attention and awareness in second language acquisition. In A. Mackey & S. Gass (Eds.), *The Routledge handbook of second language acquisition* (pp. 247–267). New York, NY: Routledge.
Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer delivered recasts and working memory on L2 development. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 229– 248). Oxford, UK: Oxford University Press.
Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158.
Schneider, W., Eschman, A., & Zuccolotto, A. (2012). *E-Prime user's guide*. Pittsburgh, PA: Psychology Software Tools, Inc.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, UK: Oxford University Press.
Smith, B. (2012). Eye tracking as a measure of noticing: A study of explicit recasts in SCMC. *Language Learning & Technology*, *16*, 53–81.

Spada, N., & Lightbown, P. M. (1993). Instruction and the development of questions in L2 classrooms. *Studies in Second Language Acquisition*, *15*, 205–224.

Trofimovich, P., Ammar, A., & Gatbonton, E. (2007). How effective are recasts? The role of attention, memory, and analytic ability. In A. Mackey (Ed.), *Conversational interaction in second language acquisition: A collection of empirical studies* (pp. 171–195). Oxford, UK: Oxford University Press.

Yilmaz, Y. (2013). Relative effects of explicit and implicit feedback: The role of working memory capacity and language analytic ability. *Applied Linguistics*, *34*, 344–368.