THE ROLE OF TASK COMPLEXITY IN THE LINGUISTIC COMPLEXITY OF NATIVE SPEAKER OUTPUT

Qualifying Paper (1)

David P. Ellis

January 17, 2011

Committee Chair:

Mike Long I approve this qualifying paper ______________________________
signature

Committee Members:

Robert DeKeyser I approve this qualifying paper ______________________________
signature

Cathy Doughty I approve this qualifying paper ______________________________
signature

Introduction

Task-based language teaching (TBLT) has garnered considerable attention since Long (1985) introduced the concept of *task* to the Second Language Acquisition (SLA) field. As Richards and Rodgers (2001) have pointed out, however, "there have been few large-scale practical applications of it and little documentation concerning its implications or effectiveness as a basis for syllabus design, materials development, and classroom teaching" (p. 224). Nearly a decade later, the situation remains largely unchanged. While TBLT has been implemented in some government-funded foreign language programs (e.g., the US National Foreign Language Initiative Flagship programs, the US Border Patrol program, and a Dutch program in Flanders, Belgium – see Van den Branden (2006)), it remains conspicuously absent from the average foreign language classroom. One reason for this seems to be the difficulty curriculum designers and practitioners have creating and sequencing level-appropriate tasks. In an effort to bridge this gap, Skehan (2003) and Robinson (2007a) have each proposed a model that delineates what they believe to be the main components of task complexity and how such complexity influences task performance. While the models differ in some respects, both predict that an increase in task complexity can/will result in an increase in the linguistic complexity of the output. To date, however, there is a dearth of evidence to support this prediction. The purpose of this paper is to determine why.

Task Complexity Models in SLA

The construct *task complexity* is not unique to SLA. There is an abundance of literature in other fields that operationalize and test the construct, from Biology (Durik & Matarazzo, 2009) to Business (Haerem & Rau, 2007), Psychology (Marsh, 2006) to Surgery (Berguer & Smith, 2006). Over the past two decades, numerous SLA scholars have tried to do the same within the context of learning a second/foreign language (Breen, 1987; Brown, Anderson, Shilcock, & Yule, 1984; Bygate, Skehan, & Swain, 2001; Candlin, 1987; Crookes, 1986, 1989; Kuiken, Mos, & Vedder, 2005; Long, 1985, 1989; Nuevo, 2006; Nunan, 1989; Prabhu, 1987; Willis, 1996). Two SLA researchers have taken important additional steps, creating models of task complexity that not only attempt to disambiguate the elements of a task but also try to predict how task complexity might drive L2 development. The first of these two researchers is Skehan.

In Skehan's (1998; 2001; 2003) model, shown below in Table 1, Skehan divides task complexity into three primary components: 1) *code complexity* (the language required); 2) *cognitive complexity* (the thinking required), which can be subdivided into *cognitive familiarity* and *cognitive processing*; and 3) *communicative stress* (the performance conditions). Based on this taxonomy and the work of Van Patten (1990), Skehan's fundamental prediction is that increasing task complexity can induce an increase in the performance of one linguistic dimension (accuracy, fluency, or complexity), but only at the expense of the other two (e.g., an increase in linguistic complexity will result in a decrease in accuracy and fluency).

Table 1. Skehan's Limited Attention Capacity model of task complexity

Code Complexity	**Cognitive Complexity**		**Communicative Stress**
	Cognitive Familiarity	**Cognitive Processing**	
Linguistic complexity & variety	Topic & its predictability	Information organization	Time limits/pressure
Vocabulary load & variety	Discourse genre	Amount of 'computation'	Speed of presentation
Redundancy & density	Task	Clarity & sufficiency of info given	# Participants
		Information type	Length of texts used
			Type of response required
			Chances to control interaction

The other task complexity model within SLA is the *Triadic Componential Framework* (Robinson, 2007a), recently renamed the *SSARC Model* (Robinson, 2010).[1] Robinson's model also comprises three primary components: 1) *task complexity* (cognitive factors, which are subdivided into *resource-directing* dimensions and *resource-dispersing* dimensions); 2) *task conditions* (interactional factors); and 3) *task difficulty* (learner factors). However, based in part on L1 acquisition research conducted by Givon (1985), Robinson predicts an increase in task complexity along resource-directing dimensions will not only result in an increase in linguistic complexity but also in accuracy (Robinson, 2006).

Table 2. Robinson's (2007a) *Triadic Componential Framework* (a.k.a., the *SSARC Model*)

Task Complexity (Cognitive Factors)		**Task Conditions (Interactional Factors)**		**Task Difficulty (Learner Factors)**	
Resource-Directing	**Resource-Dispersing**	**Participation**	**Participant**	**Affective Variables**	**Ability Variables**
+/- Here-and-Now	+/- Planning	+/- Open solution	+/- Same proficiency	H/L Openness	H/L Working memory
+/- Few elements	+/- Prior knowledge	+/- One-way flow	+/- Same gender	H/L Control of emotion	H/L Reasoning
+/- Spatial reasoning	+/- Single task	+/- Convergent solution	+/- Familiar	H/L Task motivation	H/L Task-switching
+/- Causal reasoning	+/- Task structure	+/- Few participants	+/- Shared content knowledge	H/L Processing anxiety	H/L Aptitude
+/- Intentional reasoning	+/- Few steps	+/- Few contributions needed	+/- Equal status and role	H/L Willingness to communicate	H/L Field Independence
+/- Perspective-taking	+/- Independency of steps	+/- Negotiation not needed	+/- Shared cultural knowledge	H/L Self-efficacy	H/L Mind-reading

Given the many similarities between the two frameworks, it appears the fundamental difference in prediction is based in how Skehan and Robinson view attentional resources. While Skehan sees attention as a single mechanism with all cognitive demands competing for the same finite resource, Robinson sees it as comprising multiple resources that can operate separately

[1] It is not entirely clear when or why Robinson changed the name of his task complexity model. The only reference this author can find to the acronym is in Robinson's (2010) chapter in the edited volume *Cognitive Processing in Second Language Acquisition: Inside the Learner's Mind* (eds. M. Putz & L. Sicola). On p. 248, the acronym is defined in this manner: SS = simple/stabilizing interlanguage; A = automatizing access to interlanguage; and RC = restructuring and complexifying. These phrases are preceded by a series of pseudo-mathematical formulas that apparently are to be followed by practitioners when trying to sequence tasks in the order of increasing complexity (p. 244).

and/or simultaneously through a *central executive* (Baddeley, 1986, 1996). As a result, in Robinson's view, it is possible for both linguistic accuracy and complexity to increase simultaneously without conflict, at least along resource-directing dimensions.[2]

Task Complexity Studies in SLA

Both Skehan and Robinson have cited several studies that allegedly provide support for their respective models. For example, Skehan has cited Foster and Skehan (1996), which examined how planning and task type impact performance. In their study, 32 English as a Foreign Language (EFL) students from a variety of L1 backgrounds were divided into three groups (no planning, limited planning, and detailed planning) and asked to complete three tasks (personal information exchange, narrative, and decision-making). True to the model's predictions, the authors found fluency generally decreased as task complexity increased and planning time decreased. However, none of the complexity measures exhibited a statistically significant difference regardless of task or planning condition. Moreover, accuracy exhibited a U-shaped pattern, whereby the narrative task induced the least accurate language despite being the intermediate task in terms of complexity. These findings therefore provide only partial support for Skehan’s model.

In an effort to obtain more robust support, Skehan and Foster (1999) examined the effects of task structure and processing conditions on narrative retellings, with the rationale being that more fine-grained analysis of a single task type (the narrative) would yield better results than the “crude” distinctions made among the personal information exchange, narrative, and decision-making tasks of the earlier study (p. 115). To this end, 47 EFL learners, again from a wide

[2] More specifically, Robinson claims that increasing demands along resource-directions dimensions should increase both accuracy and complexity, whereas increasing demands along resource-dispersing dimensions will have a negative impact on all three linguistic aspects of production.

variety of L1 backgrounds, completed two narrative tasks (structured sequence of events and unstructured sequence of events) under one of four different processing conditions (watch a video clip and narrate simultaneously; read a storyline, then watch and narrate simultaneously; watch once, then watch again and narrate simultaneously; and watch once, then narrate without viewing again). As anticipated, the authors found support for their hypotheses, namely that fluency was affected by the degree of task structure, complexity was affected by the processing conditions, and accuracy was affected by the interaction of task structure and processing condition, though not all of these effects were statistically significant. Nevertheless, they imply that based on the results of their study, Skehan's single-resource attention model may at least partially be able to predict the influence of task complexity on learner performance.

Like Skehan, Robinson has cited several studies that allegedly support his multiple-resource attention model. Among others (e.g., Niwa, 2000), Robinson (1995) investigated the oral production of 12 East Asian participants working in dyads on three narrative tasks (either two *Here-and-Now* and one *Then-and-There* task or vice versa), based on the assumption Then-and-There tasks are cognitively more complex than Here-and-Now tasks. Unfortunately, Robinson found support for only one of his six hypotheses (increasing task complexity would lead to increased lexical density).

Citing limitations of his 1995 study, including a small sample size, low proficiency participants, and questionable operationalizations of the fluency, accuracy, and complexity constructs, Robinson (2001b) tried to improve his design in the hopes of finding more supportive results. In a study of 44 Japanese undergraduate student participants, Robinson again used two tasks, but this time participants completed direction-giving tasks rather than narratives. For the two conditions, he used giving directions in a familiar place (the university campus) and giving directions in an unfamiliar place (a particular area of downtown Tokyo). Perhaps not unexpectedly, Robinson found more support for his model in this later study, much like Skehan and Foster (1999) did. Specifically, he found that both accuracy and fluency were influenced by task complexity. However, changes in linguistic complexity were not statistically significant.

More recently, Cadierno & Robinson (2009) and Robinson, Cardieno, and Shirai (Robinson, Cadierno, & Shirai, 2009) conducted time and motion event studies to determine the influence of task complexity on oral performance. In the Cardieno and Robinson (2009) study, the authors operationalized task complexity along the *Here-and-Now* vs. *Then-and-There* resource-directing dimension, examining the oral performance of two groups whose L1s differ in their typological distance from English. Twenty Danish and 20 Japanese L1 speakers participated in the study, which involved two picture strip oral narrative tasks. With motion verb complexity and motion clause complexity the dependent variables (among others), Cadierno and Robinson found that ANCOVA analyses yielded results that approached (but did not achieve) significance with respect to their complexity measures. In fact, the only statistically-significant finding was the correlation between proficiency and target-like performance, which of course is to be expected and was openly acknowledged as such by the authors (p. 264).

The findings of Robinson et al (2009) are very similar. In this study,[3] the complexity variables were operationalized as the percentage of participants who produced non-prototypical uses of tense-aspect markings and the raw frequencies of tense-aspect morphology use to mark specific semantic categories. Once again, the findings approached but did not achieve statistical significance.

[3] Two studies were reported in this paper, but the second is a recapitulation of the Cadierno and Robinson (2009) study.

Aside from these self-run studies, other SLA researchers have tested these task complexity models. For example, Gilabert (2007) investigated the effects of task complexity on self-repair. In a study of 42 Spanish-speaking EFL learners who performed three oral narrative tasks (comic strip, map, and decision-making), the author found partial support for the relationship between task complexity and the amount of self-repair and error-free production. Interestingly, Gilabert cites these findings as support for Robinson's Cognition Hypothesis (2005), but it would appear the data simultaneously support Robinson's framework but Skehan's prediction because only one dimension of production (accuracy, as operationalized by the amount of self repair and error-free production) increased with a change in task complexity. In other words, increasing task complexity improved the performance along only one linguistic dimension (accuracy), not accuracy *and* complexity, as Robinson's Cognition Hypothesis (2005) predicts, provided the increase in task complexity is along a resource-directing dimension, which it was in this study (Task 1: +/- Here-and-Now; Task 2: +/- Few Elements). In this sense, Robinson's *framework* finds some support in Gilabert's study because manipulating tasks along resource-directing dimensions enabled some systematic changes in accuracy, but because no measures of complexity were examined, there can be no claims of support for Robinson's *prediction*. The author even acknowledges this failure to test Robinson's model in his conclusion:

> Certainly, subsequent studies on the same data should look at how accuracy interacts with the dimensions of fluency and lexical and structural complexity. It remains an issue how task design may force learners to make strategic decisions, in Wickens' (this volume) terms, as to which dimension of production (i.e., fluency, complexity, or accuracy) they allocate attention to (p. 237).

As shown, the author looked only at accuracy and makes no claims about simultaneous increases in complexity. Moreover, he implicitly advocates testing Skehan's limited-attention model by virtue of seeking to determine *which dimension of production...[learners] can allocate attention to*, words which imply making a choice of one dimension over another rather than support for the multiple-resource model Robinson advocates.

In a follow-up study, Gilabert, Baron, and Llanes (2009) employed the same three tasks with 60 Spanish-speaking EFL learners, but the authors modified them to make them dialogic rather than monologic tasks. Using negotiation of meaning (via clarification requests, confirmation checks, and comprehension checks), recasts, and self-repair as the dependent variables, the authors once again found mixed results, with some of the measures exhibiting statistically-significant relationships while others did not. Ironically, the authors once again cite support for Robinson's Cognition Hypothesis (2005) because there was greater negotiation of meaning on, e.g., the more complex comic strip task (i.e., a greater number of clarification requests and comprehension checks were generated on the more complex task). However, this interpretation is backward: if a greater number of clarification requests and comprehension checks are needed to complete the task, then the accuracy of the output is presumably *lower*. In other words, accuracy actually *decreased* as task complexity increased, which is opposite of what Robinson's Cognition Hypothesis would predict. Moreover, even if greater negotiation of meaning is hypothesized to facilitate acquisition/development, as Robinson suggests it does, there is still no support for his model in this study's findings because the experiment was cross-sectional, not longitudinal, a design problem that precludes any claims about developmental impact, as explained later in the discussion section of this paper.

Revesz (2009) looked at the effects of recasts on learner output, in particular their effect on *focus on form* (Doughty, 2001). In this study, 90 EFL learners completed a series of picture description tasks manipulated along two resource-directing dimensions (+/- Contextual Support and +/- Here-and-Now). While the author found a large effect size between performance and the combination of recasts and increased task complexity, she found only a small effect size when comparing oral performance to task complexity alone. This finding is congruous with Nuevo's (2007) null finding for the effects of task complexity on oral performance, leading the author to conclude that "…task features, in and of themselves, without being combined with some type of external intervention, may only have a limited impact on L2 learning" (p. 465).

This conclusion finds further support in Kim (2009a), where mixed support was found for the conclusion that turn-taking increases when task complexity increases. In this study, 17 dyads divided into high- and low-proficiency groups performed two tasks (picture description and picture difference identification) to determine the number of language-related episodes (LREs) they would generate when the tasks were manipulated along the resource-directing dimensions +/- Complex Reasoning and +/- Few Elements. On the picture description task, Kim found that the low-proficiency group generated more LREs on the simple version of the task, while the high-proficiency group generated more LREs on the more complex task. The opposite was true on the picture differences task, where the low-proficiency group generated more LREs on the more complex task while the high-proficiency group exhibited no statistically-significant difference between the two tasks. A synthesis of these conflicting findings suggests an interaction between task complexity and some other factor(s), as Revesz (2009) surmised.

Perhaps more interesting than the mixed results found in Kim (2009) is that the low-proficiency group generated a greater number of LREs on all tasks, regardless of task type or level of complexity, which is in line with Gilabert et al's (2009) findings that turn-taking is greatest among less accurate speakers. If true, this finding would run contradictory to what Robinson's model would predict with respect to the accuracy of learner output in the face of increasing task complexity. At the very least, it calls into question whether proficiency interacts with task complexity, and it is this interaction that has an effect on linguistic complexity.

Although oral performance is the focus of most task complexity studies, Kuiken and Vedder (2007; 2008) examined the effects of cognitive complexity on writing output. In their 2007 study, 84 Dutch students of Italian and 75 Dutch students of French were grouped into low- and high-proficiency groups and tasked with writing a letter to a friend about holiday travel plans. In the repeated-measures design, the letter-writing task was manipulated along the resource-directing dimensions +/- Few Elements and +/- Complex Reasoning, and like virtually every other task complexity study, findings were mixed. For example, fewer lexical errors on the more complex task were found in the output of both groups, but more orthography, “appropriateness,” and “other” errors were found only among the French learners on the more complex task (p. 275).

In the 2008 study, which appears to be a reanalysis of the same data with a few additional participants, 91 Dutch students of Italian and 76 Dutch students of French performed the same letter-writing task that was manipulated along the +/- Few Elements and +/- Complex Reasoning resource-directing dimensions. For this study, the authors examined how measures of syntactic complexity, syntactic accuracy, and lexical variation changed with an increase in task complexity. Findings yet again illustrate mixed results, with an increase in accuracy, but no statistically significant difference in either syntactic complexity or lexical variation. Oddly, the authors cite the improvement in accuracy as support for Robinson’s Cognition Hypothesis (pp. 56-7), yet they themselves acknowledge Robinson’s claim is an expected increase in both accuracy *and* complexity, not only one or the other (p. 50).

There have been several other studies in recent years that test these models (Michel, Kuiken, & Vedder, 2007), including many PhD dissertations (Kaneko, 2008; Lee, 2002; Medina, 2008), but the outcome of each is more of the same – mixed or null findings and often misinterpreted conclusions. In short, there is not one single study in publication that provides clear support for either model.

Critique of the Models & Studies

While there are many reasons these frameworks lack empirical support, one is flawed experimental design in the early studies. For example, in Foster and Skehan (1996), the authors had four treatments and two experimental conditions (limited planning and detailed planning) but only 16 dyads, yielding an N-size of only two dyads per condition. Needless to say, it is unsurprising few effects were found. This study had other problems, too, including how the authors operationalized complexity, accuracy, and fluency (e.g., the summation of pauses as a measure of fluency), the rationale for choosing the tasks they used in the study (i.e., because similar tasks are commonly found in ESL textbooks), the low level of proficiency of their participants, and the decision to analyze only group mean data rather than looking at changes within subjects, thereby masking potentially important data.

Skehan and Foster (1999) also suffered from many of the same methodological weaknesses. For example, their operationalizations of complexity, fluency, and accuracy are again questionable (e.g., they oddly eliminated their only statistically significant measures of fluency from their Foster and Skehan (1996) study). Moreover, they still examined only group means rather than individual changes in performance.

Robinson has also cited studies fraught with methodological flaws. For example, Robinson (1995) suffered from a very small participant population (N=12) and weak construct measures (e.g., target-like use of articles as a measure of accuracy and the frequency of 2-second pauses as a measure of fluency). Moreover, and as Robinson (1995) acknowledges, the distinction between his tasks may have been insufficient, and the proficiency level of his students was too low, resulting in problems similar to those of Skehan and Foster (1996). Similar problems haunted Robinson (2001) despite attempts to the contrary.

More recent research has been sound methodologically (Cadierno & Robinson, 2009; Gilabert et al., 2009; Revesz, 2009), but the findings were still mixed because of a more fundamental problem – how the dimensions of task complexity are defined (or not). As Kuiken and Vedder (2007) explain:

> Of these two models Robinson's Framework constitutes the most elaborate attempt at developing a model of task complexity. However, the Triadic Componential Framework also raises a number of questions. Intuitively it may be assumed that the variables distinguished by Robinson do play a role in determining task complexity, but it is far from clear how these variables have to be operationalised, which of them are predominant, how they interact and how fine-grained they should be…With other additions and substitutions the 2007 version of the Triadic Componential Framework comprises 36 variables instead of the 18 variables in the earlier version of the model. One may wonder how all these variables can be operationalised and differentiated and how for instance the supposedly different kinds of reasoning should be tested in an experimental setting (p. 265).

In short, the authors of previous studies have lacked clear guidance on how to operationalize the dimensions of task complexity within either framework. As Kuiken and Vedder (2007) noted, Skehan's model is all but devoid of elaboration. The same cannot be said for Robinson, who has elaborated on his model in multiple publications (1995; 2001a; 2001b; 2005; 2006; 2007a; 2007b; 2010). Nevertheless, clarity remains elusive. Take Robinson's resource-directing dimensions. Setting aside for the moment why these six dimensions should be assumed to promote increases in both accuracy and complexity, it is unclear how to operationalize many of them (e.g., +/- Intentional Reasoning, +/- Perspective Taking). The names alone conjure a multitude of interpretations, and there is precious little practical guidance from Robinson despite his abundant publications on the Cognition Hypothesis and Triadic Componential Framework/SSARC Model. It is little wonder +/- Few Elements and +/- Here-and-Now are the two dimensions manipulated most often in the research; their names are fairly self-explanatory, or so it would seem until actually trying to operationalize them. Take +/- Few Elements. How many elements characterize a “few”? Three? Six? Ten? And how does one count the number of elements in a task? All of the visual elements in a picture, each with an equal weight with respect to salience? Or should only those elements designed to be elicited during oral production be counted? If the latter, how would one know which elements will actually emerge until after the task has been administered?

The point here is not to nitpick the efficacy of the +/- Few Elements as a dimension, but to highlight just how difficult it is to operationalize any purported dimension of task complexity. Put simply, the dimensions as posited by Robinson and Skehan are too ambiguous and/or under-defined at this point to be tested rigorously. Even seemingly transparent dimensions like +/- Few Elements are much more opaque than they first appear, so operationalizing any of the other dimensions is challenging indeed.

The second problem with previous research is its myriad operationalizations of the dependent variable. As illustrated earlier, operationalizations of linguistic complexity have ranged from the number of S-nodes per T-unit (Robinson, 1995) to the amount of self-repair (Gilabert, 2007) to the number of language-related episodes (Kim, 2009b) to the number of recasts (Revesz, 2009) to the amount of negotiation of meaning (Gilabert et al., 2009). And within many of these studies, the broader dependent variables (e.g., amount of negotiation of meaning) were each operationalized in several ways (e.g. number of confirmation checks, comprehension checks, clarifications requests, etc). The same is true for the other dependent variables of interest – accuracy and fluency.

The fact is, there are seemingly innumerable means of operationalizing these three constructs. In the L2 writing domain, for example, Wolfe-Quintero et al (1998) identified over 100 measures of linguistic performance. Acknowledging there are many equivalent measures in the L2 speaking domain (if not even more when both monologic and dialogic production is considered), one can quickly see why findings to date have been mixed.

In sum, there has been no consistent operationalization of either the independent variables (dimensions of task complexity) or dependent variables (dimensions of linguistic output) across task complexity studies. The result is a complete lack of replication studies within this line of research, yielding virtually no corroborating evidence for any statistically significant findings that have been found. This no doubt is due in part to the lack of unambiguous findings in the research, but the field as a whole has taken a seemingly unsystematic approach to this research. If it is to ever see any success, there must be a much more organized and theoretically sound approach to not only operationalizing task complexity, but also to operationalizing the dependent variables.

The third but most easily overlooked flaw of the previous task complexity research is the participant population of each study. In all but one (Tavakoli & Foster, 2009), the participants have been second language learners. At first blush, this methodological choice seems to make perfect sense: if L2 development is the models' raison d'etre, then their hypotheses/predictions should be tested using L2 learners as participants. However, as explained earlier, the current research paradigm is wrought with numerous other more-immediate challenges, not the least of which is a clearer and more finite operationalization of the independent and dependent variables.

Adding the immense variation in L2 participant profiles (i.e., their variations in proficiency, age, L1, educational background, etc) serves only to further confound results. Recall that the goal is to establish a systematic relationship between task complexity and linguistic complexity (if not also accuracy). Until a stable, replicable relationship can be found between at least one measure of task complexity and one measure of linguistic complexity, the individual differences among L2 learner profiles further confounds the testing of each model's hypotheses/predictions.

To overcome this design challenge, there is a simple solution – test native speakers of the language. The advantages of doing so are clear. First, native-speaker-like production is the putative goal of second language learning. Although some researchers would dispute this claim (Birdsong, 2005; Doughty, 2008), native-like production underpins several lines of SLA research. One need look no further than the research on ultimate attainment, age effects, and the Critical Period Hypothesis. Whether one supports (DeKeyser, Alfi-Shabtay, & Ravid, 2010) or refutes (Flege et al., 2006) the existence of a critical period for language learning, researchers on both sides of the argument cite evidence for their position by juxtaposing (either explicitly or implicitly) their L2 subjects' performance with native-like language production.

Second, and of far greater practical importance, is that testing native speakers provides a methodologically-induced control for accuracy, fluency, and processing demands because nearly all native speakers, regardless of educational background or intelligence, have complete command of their L1 grammars.[4] In other words, not only can linguistic complexity be isolated as the lone dependent variable, thereby leaving its operationalization as the "only" design

[4] The claim being made assumes the "average" native speaker of a language, one who suffers from no hearing or speech impediments like those who stutter or who were born with hearing intact but became deaf shortly after birth, thereby affecting their pronunciation, fluency, and perhaps accuracy and complexity of their speech.

challenge on the dependent variable side of the equation, but task *processing* demands can be clearly segregated from task *complexity* demands.

In sum, at least one stable, replicable relationship between one independent variable (dimension of task complexity) and one dependent variable (dimension of linguistic output) must be demonstrated empirically before this line of research can make any forward progress. In an effort to facilitate the likelihood this will happen, testing native speakers seems to be the most promising means of doing because it eliminates many of the problems that had plagued previous studies. Unfortunately, as pointed out earlier, only one other research team to date has recognized the value in taking this approach (Foster & Tavakoli, 2009).

In their study, Foster and Tavakoli tested 40 first-year university students majoring in either Literature or Psychology using four narratives manipulated along the dimensions of tight-loose narrative structure and +/- complex storyline. Using number of clauses per AS-unit and mean length of utterance (MLU) as the dependent variables,[5] the authors found support for the hypothesis that increased storyline complexity would lead to more linguistically complex language use among the native-speaking participants. Interestingly, they found a similar result among their non-native-speaking participants who performed the same tasks in an earlier study, but this correlation broke down with respect to the fluency and accuracy of non-native speaking output (Tavakoli & Foster, 2008).

Summary

Numerous studies on the effects of task complexity on L2 learner output have been conducted over the past 15 years, and yet no single stable finding has emerged. While there may be many factors contributing to this problem, three seem to be most salient:

[5] See (Foster, Tonkyn, & Wigglesworth, 2000) for an explanation of the use of AS-units as a theoretically-motivated operationalization of the dependent variable.

1) a lack of guidance on how to operationalize the dimensions of task complexity as outlined in the frameworks
2) a lack of consistent operationalizations of the dependent variables of interest (linguistic complexity, accuracy, and fluency)

a widely-varying participant population that obscures whatever relationships may have otherwise been found.[6]

Research Question

In light of the literature review above, the following research question has emerged as the sole focus of this study: *Can the linguistic complexity of speaker output be shown to vary systematically with changes in task complexity?*

Methods

Participants

Thirty-six native speakers of English participated in the study – 18 males and 18 females. Participant ages ranged from 21 to 49. All are US citizens who speak U.S. English.[7]

Tasks

Two task types were employed to minimize task type effect and increase the amount of output that could be collected and analyzed. One task is a car accident narration task (see Appendix A), where participants had to describe a car accident they just witnessed. In addition to reporting what they saw, they were asked to decide who was at fault and why. The second task is

[6] See Samuda and Bygate (2008) for an alternative critique of Robinson's framework.

[7] Institutional Review Board (IRB) approval was acquired prior to recruiting participants, and all required IRB procedures were followed during the study, including the receipt of signed consent forms from all participants prior to commencement of the experiment.

a map directions task (see Appendix B). Participants were told they work in an information booth at Golden Gate Park in San Francisco and have to describe to a tourist how to get from Point A to Point B. They were also told they were giving the tourist directions over the telephone so they had to be as explicit as possible while explaining the route.

In light of the issues raised earlier regarding the number of task dimensions purported to increase linguistic complexity and the lack of clarity on how to operationalize them, only one dimension was chosen for this study, that being Robinson's (2007) +/- Few Elements. This dimension was chosen for two reasons. First, it is one of the six resource-directing dimensions Robinson claims can be manipulated to induce more linguistically complex (and accurate) output. Second, it is one of the most self-evident resource-directing dimensions with respect to how to operationalize it.

Both tasks were also manipulated on three levels rather than the more common two, with the hope greater insight could be gained into how linguistic complexity is related to task complexity, if at all. Because no objective means of verifying the complexity of the tasks other than by merely counting the visual elements within the task, three experienced teachers of second/foreign languages were asked independently to sequence the tasks in order of increasing complexity. For both task sets, 100% agreement was reached with respect to the relative complexity of each version of each task.

Procedure

Using a repeated measures design, each participant was asked to complete all six tasks in a prescribed order. Counterbalancing measures were employed both within and across tasks to minimize the likelihood of a sequencing effect. Table 3 illustrates the counterbalancing method employed in this study:

Table 3. Sample of the counterbalancing of task order and task complexity

Participant	Gender	Map Task	Car Task	Task Order
1	M	312	321	MC
2	M	321	312	CM
3	M	132	231	MC
4	M	123	213	CM
5	M	231	132	MC
6	M	213	123	CM
7	M	312	123	MC
8	M	321	132	CM
9	M	132	312	MC

1 = least complex task; 2 = mid-complex task; 3 = most complex task

All participant responses were recorded and transcribed using *Express Scribe*, an open-source transcription application. Data were then parsed and coded using the linguistic measures described below.

In addition to completing the six tasks, each participant completed a background questionnaire (Appendix C) and C-test (Appendix D) so that native speaker performance could be compared to non-native speaker performance in a subsequent study. Total participation took about 45 minutes and each participant was paid $10 for their time.

Linguistic Measures

As noted in the literature review, there are numerous problems with the way researchers have operationalized linguistic complexity in the past, not the least of which is that many operationalizations do not seem to be measures of linguistic complexity at all (e.g., the number of self-repairs made). To avoid a similar problem, one of the two measures of linguistic complexity used in Robinson (2001) – the number of dependent clauses per C-unit (CPC)[8] – was used because it is both theoretically-motivated and theoretically-justifiable (i.e., subordination is

[8] *C-unit* is defined here as an utterance that consists of a single complete sentence, phrase, or word and has a clear semantic and pragmatic meaning in the context in which it occurs. In effect, it is the same as a T-unit except that it includes elliptical utterances (Ellis, 2004).

a widely-accepted indicator of linguistic complexity, both in spoken and written output, and syntactic complexity is widely-accepted as a subcomponent of linguistic complexity).[9]

Results

Table 4 illustrates the mean number of clauses per C-unit generated during the completion of each task.

Table 4. Mean number of clauses per C-unit (CPC)
M = Map task; C = Car accident task; 1 = least complex task; 2 = intermediate task; 3 = most complex task.

	Mean CPC	Std. Deviation	N
M1	1.6113	.22950	36
M2	1.5013	.18939	36
M3	1.5750	.24553	36
C1	2.3600	.49745	36
C2	2.2275	.26927	36
C3	2.2100	.42962	36

As shown, the mean number of clauses per C-unit exhibited a U-shape pattern on the map task (1.6113 → 1.5013 → 1.575) and a decrease on the car accident task (2.36 → 2.2275 → 2.21). These results are contrary to the predictions of both models.

Because group results can sometimes mask effects, individual task results were analyzed to determine whether any participants exhibited the expected behavior on either/both of the tasks. Table 5 displays the participants who did.

[9] The other measure of complexity used in Robinson (2001) was type-token ratio (TTR), which is a measure of lexical variation. To keep the focus of this research as narrow as possible, only CPC was measured. Whether lexical variation – or TTR, as it was operationalized in Robinson (2001) – is a measure of complexity is also more open for debate than is subordination, which is generally accepted/uncontroversial among SLA researchers.

Table 5. Participants who exhibited anticipated behavior on either of the tasks.
M = Male; F = Female. Mean and Range are for entire participant population.

	Clauses per C-Unit							
	Map 1	Map 2	Map 3	Mean	Car 1	Car 2	Car 3	Mean
M4	1.33	1.25	1.48	1.36	**1.91**	**2.05**	**2.43**	2.13
M10	1.57	1.65	1.51	1.58	**1.83**	**2.23**	**2.32**	2.13
F13	**1.37**	**1.46**	**1.68**	1.50	3.22	2.31	2.25	2.59
F14	**1.42**	**1.49**	**1.70**	1.53	1.33	1.33	2.67	1.78
F15	1.82	1.59	1.26	1.56	**2.40**	**2.44**	**2.53**	2.46
F18	**1.11**	**1.33**	**1.52**	1.32	**1.64**	**1.70**	**2.13**	1.82
F19	1.83	2.13	1.42	1.79	**1.67**	**1.82**	**2.85**	2.11
F22	1.53	1.27	1.37	1.39	**2.08**	**2.20**	**2.40**	2.23
M Mn	1.44	1.44	1.40	1.43	2.14	2.09	2.01	2.08
F Mn	1.51	1.47	1.46	1.48	**2.07**	**2.09**	**2.20**	2.12
Mean	1.47	1.46	1.43	1.45	2.11	2.09	2.11	2.10
Range	1.00	1.07	1.05		1.29	1.25	1.20	
	2.00	1.88	1.92		3.22	2.56	3.00	

As illustrated, three of the participants exhibited the predicted increase in linguistic complexity on the map task (P13, P14, P18), while six did on the car task (P4, P10, P15, P18, P19, P22). Moreover, females as a group exhibited an increase in linguistic complexity on the car task. Unfortunately, none of these results is statistically significant. Perhaps more importantly, the linguistic patterns exhibited in the data cover literally every possible combination: U-shape, inverted U-shape; increase in complexity, and decrease in complexity. In short, there were no predictable outcomes.

Two-way Analysis of Variance (ANOVA) with task type and task complexity as the independent variables and gender as a covariate was also conducted. As shown in Table 6, task type was significant (F(1, 34) = 174.301, *p* < .001) and gender approached significance (F(1,34) = 3.722, *p* = .057), but neither task complexity nor its interaction with task type was statistically significant (*p* = .970 and .918, respectively).

Table 6. Results of 2-way ANOVA

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	22.042(a)	6	3.674	29.710	.000
Intercept	56.849	1	56.849	459.756	.000
Gender	.460	1	.460	3.722	.055
Task Type	21.552	1	21.552	174.301	.000
Task Complexity	.008	2	.004	.031	.970
Type * Complexity	.021	2	.011	.086	.918
Error	25.843	209	.124		
Total	717.565	216			
Corrected Total	47.885	215			

(a) R Squared = .460 (Adjusted R Squared = .445)

Visual inspection of the two tasks provides some insight into this cross-task difference. Put simply, a great deal of subordination is necessary to describe each version of the car accident whereas the map task does not. For example, the similar color of the vehicles in the most complex car task requires numerous relative clauses and modifying phrases to differentiate the vehicles. The same is true for the spatial relationships in each scene. All of the vehicles and pedestrians have to be described in relation to each other, which requires numerous prepositional phrases and relative clauses. In contrast, the map task appears to require far less subordination (e.g., the spatial relationship among the elements appears far less intertwined). As a result, far fewer locative prepositions, modifying phrases, and relative clauses are needed to complete the task. Instead, it appears a wider lexicon is needed to succeed on the map tasks.[10]

[10] Whether lexical range/variation/diversity should be considered a measure of linguistic complexity at all was questioned earlier and is addressed in more detail in the Discussion section of this paper.

Discussion

As illustrated, there was a significant difference in linguistic complexity across tasks but not within tasks. This is an important finding, but not without a caveat: while it is certainly useful to know curriculum developers can design tasks that will elicit particular morpho-syntactic features like subordination, the question remains how to determine which features of linguistic complexity are more (or less) complex relative to each other. In this case, for example, if a teacher had to sequence the map and car tasks within a curriculum, would it make more sense to have the map tasks precede the car tasks because subordination is assumed to be more complex than lexical range? Or is the opposite true – that lexical range is more complex than subordination, thereby dictating the map tasks follow the car tasks? Or do subordination and lexical range need further subdivision to ascertain their relative levels of complexity?[11] This issue is important to consider because both measures have appeared frequently in the literature as measures of linguistic complexity (e.g., Ortega, 2009; see Malvern & Richards (2002) for a sample of lexical richness/diversity studies), yet no one has explored their relative levels of complexity. Without some clarification, sequencing target tasks within a curriculum remain problematic.

Ultimately, relative linguistic complexity across tasks may be a moot point. It has been demonstrated here that different tasks types clearly elicit different linguistic features and can in fact be designed to induce the production of particular features. However, linguistic features per se are not the focus of a task-based syllabus (Long, 1985). Instead, the focus is on developing sufficient competence to complete the target tasks identified during the needs analysis. Because it is assumed learners will be unable to perform most or all of the target tasks at the start,

[11] See Norris and Ortega (2009b) for a fuller treatment of this issue.

practitioners need to know how to create and sequence pedagogic tasks that approximate the target task to aid their learning (Long, 2005). For this reason, the null findings with respect to within-task variation are troubling. Trying to sequence tasks within a target task type using this study's findings is all but impossible because participants did not exhibit a consistent pattern of linguistic complexity within either of the tasks. Quite the opposite, participants exhibited every possible order of complexity despite the fact each set of tasks was independently rated by three experienced ES/FL teachers, with 100% agreement on the sequence of the tasks. There are at least three possible explanations:

1) The independent variable was not operationalized appropriately.

One possible explanation for the null findings within tasks is how +/- Few Elements was operationalized in this study. Because the dimension is fairly self-explanatory, the only immediately obvious way to alter the operationalization of +/- Few Elements is by increasing the difference in the number of elements that separates a less-complex task from a more-complex task. In this light, it could be argued the more-complex versions of the tasks simply did not contain enough additional elements to sufficiently differentiate them from the less-complex tasks. Recall, however, that +/- Few Elements was manipulated not once but twice. As a result, the middle task was removed and the data reanalyzed using only the bookend tasks, thereby further widening the gap in number of elements between the less and more complex tasks. Unfortunately, doing so still yielded neither discernible patterns nor statistically significant differences in CPC for either task. As a result, this subsequent finding yields three possible interpretations with respect to the independent variable used in this study:

1) +/- Few Elements is a core dimension of task complexity but needs further refinement
2) +/- Few Elements is not a core dimension of task complexity for these task types
3) +/- Few Elements is not a core dimension of task complexity for any task type

As suggested in the literature review, it could be possible viewing +/- Few Elements only in terms of the number of elements is misguided. It could be that not all elements are created equal, meaning it is possible certain elements are more important than others and should carry more weight when manipulating them along this dimension. For example, it may be more important to manipulate the number of *salient* elements – presumably only those critical to task completion – because learners are likely to ignore others to keep cognitive processing load to a minimum, thereby freeing up attentional resources for other task demands. In this study, then, perhaps more cars and/or pedestrians could have been added to/near the intersection where the accident tasks place, which in turn may have necessitated additional elaboration/explanation to complete the task. The same is true for the map task. Were, e.g., obstacles and/or more landmarks added to the paths of the more complex tasks, they might have induced the need for greater elaboration and in turn greater subordination.

Another aspect of elements to consider is their relationship to one another. Perhaps it is more important to establish links between/among additional elements than it is to merely add to their volume. Consider adding one person and one car to the most complex car task: it may be more beneficial to establish a relationship between the two elements rather than just adding them independently. Having the person positioned in front of an oncoming car, for example, might increase the complexity of the language needed to describe their proximal relationship (e.g., *the accident probably could have been avoided, but the dark blue car had to swerve suddenly into oncoming traffic to avoid hitting the sky blue car that was rapidly approaching a pedestrian* *crossing the street*). In short, without some connection, the two additional elements may do nothing other than add visual clutter to the scene.

Whatever the merit of these two particular hypotheses, it seems likely merely adding more elements to a task is not a sufficient condition to induce a systematic change in the complexity of learner output, so a further refinement/description/elaboration of this (and all) resource-directing dimensions is necessary.

A second possible interpretation of the null finding is that +/- Few Elements is not an appropriate dimension to manipulate for these particular tasks. In other words, it may be certain task dimensions are better matches for certain task types (Jackson, 2009). This could very well be the case and should be explored empirically to determine whether this hypothesis can be supported. Using this study's two basic task types, for example, one could manipulate all six of Robinson's (2007) resource-directing dimensions (separately) to see whether one or more produces systematic changes in the dependent variable.

The third possible interpretation of the null finding is that the number of elements in a task simply has no relationship to its complexity, at least not in isolation. As mentioned in the Methods section, +/- Few Elements was chosen because it is one of the few resource-directing dimensions seemingly easy to operationalize and test. It goes without saying ease of operationalization is not a valid means of choosing a variable of interest. Nevertheless, both task sets were independently rated by three experienced curriculum developers and sequenced with 100% agreement, so there was at least a *perceived* increase in complexity.

While perceptions are reality in some contexts, Norris and Ortega (2003) would argue to the contrary in this case, claiming all independent variables must be validated before being introduced into research. i.e., Even if a predictable relationship between linguistic complexity and +/- Few Elements were established in this or any other study, it would remain unclear whether task complexity per se was manipulated because +/- Few Elements has never been validated as a dimension of task complexity. Instead, Norris (2009a) argues researchers must first validate all potential independent variables by gathering and analyzing one or more of the following sets of data:

1) self-reported data from subjective judgments of task difficulty or subjective time estimations after completing a task
2) behavioral data from reaction times on a secondary task (the rationale being that more cognitively complex tasks will be susceptible to greater deterioration in performance when a dual task condition is introduced)
3) physiological data from measurement of heart rate variability, skin conductivity, pupillary response, brain activation and blood flow patterns, and even the pressures which study participants apply to a computer mouse during computer

While Norris (2009a) and Norris and Ortega (2003) make a compelling argument, they seem to equate erroneously cognitive complexity with task complexity. Note that Robinson (2005) makes an important distinction between task *complexity* and task *difficulty*. Whereas difficulty is relative to the learner (what one learner finds difficult, another finds easy), the complexity of a task is absolute (external to/independent of the learner). As a result, measuring the cognitive load a learner experiences while completing a task is not a measure of its complexity but its difficulty.

The current study is a good example of the problem one would encounter if using measures of cognitive load to rate the complexity of a task. Neither the car task nor the map task was designed to be particularly cognitively challenging for the native-speaking participants (in fact, they struggled more with the C-test than they did the tasks). For this reason, it seems reasonable to conclude that, e.g., brain scans of neural activity would be no different for these participants regardless of which version of the task they were completing. On the other hand, the results of this study do demonstrate that much more syntactically complex language was needed to complete the car task than the map task, thereby reinforcing the necessary distinction between complexity and difficulty. In short, task difficulty/cognitive load should not be equated with task complexity.

All this being said, there is still no evidence +/- Few Elements is in fact a core dimension of task complexity. Further research along the lines suggested earlier (manipulation of the more salient elements and/or more intricate relationships among the elements, as well as testing the dimension against a greater variety of task types) need to be conducted before the validity of +/- Few Elements as a core dimension of task complexity can be confirmed or refuted. The same is true for all six resource-directing dimensions, in fact.

2) The dependent variable was not operationalized appropriately.

A second reason for the within-task null findings could be that the dependent variable was not operationalized appropriately/sufficiently. One way to examine its appropriateness is to look at the findings. In this case, the data do illustrate subordination was relied upon heavily in the completion of the car task, if not the map task. In this respect, clauses per C-unit (CPC) as the dependent variable seems to have been a reasonable choice, at least for the car task. More important than this post-hoc confirmation, however, is that it was theoretically motivated. As noted in the literature review, stating subordination is an indication of more complex language is uncontroversial among SLA researchers, as is claiming CPC is a reasonable measure of subordination. The frequent use of CPC as a measure of syntactic complexity in task complexity research further supports this claim (e.g., Iwashita, McNamara, & Elder, 2001; Robinson, 2007b). Nevertheless, there are a few potential concerns with this choice of dependent variable. The first is whether linguistic complexity (or the even narrower construct, syntactic complexity) can be seen as unidimensional. Norris and Ortega (2009b) argue it cannot, suggesting it has to be analyzed in several distinct but complementary ways:

> In the case of syntactic complexity…SLA researchers should measure global or general complexity, complexity by subordination, and complexity via phrasal elaboration, as well as possibly coordination if early proficiency data are also included. That will demand the use within single studies of metrics chosen to tap at least overall complexity (e.g. mean length of T-unit), complexity by subordination (e.g. mean number of clauses per T-unit), and complexity by subclausal or phrasal elaboration (e.g. mean length of clause) (p. 574).

In this light, one could argue it was ill advised to examine only one aspect syntactic complexity and label it linguistic complexity. However, if one concedes CPC is a valid instantiation of "complexity by subordination," whether measuring one measure of syntactic complexity or ten, *any* null finding is troubling provided the measure was motivated theoretically as it was in this study. i.e., it would be inappropriate to disregard null findings in one measure even if statistically significant findings were found in all of the other measures. This might not be the case if all measures were related to a unidimensional construct, but as Norris & Ortega (Norris & Ortega, 2009b) suggest, utilizing multiples measures that are designed to measure the same (unidimensional) construct is both redundant and unnecessary (p. 560).

Regardless, it could be argued linguistic complexity comprises not only syntactic complexity but also lexical complexity. However, it is argued here that lexical range/diversity is not a measure of linguistic complexity. Put another way, syntactic complexity and lexical range/diversity – whether unidimensional or multidimensional themselves – are separate constructs and therefore should not be combined under the umbrella construct linguistic complexity. Data support this argument, too, as syntactic complexity and lexical range/diversity have been shown to load on separate factors in numerous studies (Wolfe-Quintero et al., 1998).

Even on the surface, it seems misguided to associate lexicon with complexity per se. Take the word *auteur*. It is no more complex conceptually than its synonym *director* – they are both concrete nouns describing the person responsible for leading the development of a movie. The fact that auteur is less common in the input does not make it more complex, only less common. Therefore, it is argued that *low frequency* and *complexity* should not be considered two sides of the same coin.[12] As a result, it is argued here syntactic complexity is a reasonable and full characterization of linguistic complexity for the purposes of this study. Moreover, it is argued clauses per C-unit is a sufficient operationalization of syntactic complexity for the reasons cited in Norris and Ortega (2009b).

3) Linguistic complexity, however operationalized, is the wrong choice for dependent variable. A third and final interpretation of the null findings is that linguistic complexity, no matter how it is operationalized, is the wrong construct on which to base any model of task complexity. Consider once again the genesis of this line of research: researchers interested in Instructed SLA were (and still are) concerned primarily with determining how to best sequence pedagogic tasks to promote L2 development. The fundamental problem with all of the task complexity findings to date is they say nothing about acquisition, only online production. Stated differently, if a study's design is cross-sectional and/or employs repeated measures, the output elicited is merely

[12] The argument could be made that certain classes of words are more complex than others (e.g., abstract nouns vs. concrete nouns). However, most measures of lexical range in the literature involve type-token ratios and rarely if ever classify word classes as more or less complex than others, even when particular word classes (e.g., range of verb types) are the input for the TTR measure.

a reflection of what the participants were able to access from their existing L2 knowledge base at the time of task completion. In other words, even when findings suggest a relationship between task complexity and linguistic complexity, participants made no developmental progress; they were merely accessing more complex aspects of their existing interlanguage.[13]

Take for example the research design of nearly all task complexity studies to date (including this one). Almost invariably, a task is manipulated along some alleged dimension of task complexity and administered to participants in a counterbalanced, repeated-measures design. In the worst case, findings are null across the board, regardless of how the variables are operationalized and tested (Nuevo, 2007). In the best case, one or more measures show some systematic relationship (Gilabert et al., 2009). The question is, to what avail? All that has been demonstrated is researchers can manipulate some aspect of the task and elicit a particular linguistic feature. The problem is, no conclusions can be drawn with respect to L2 development because development is gradual, not instantaneous, and cross-sectional and repeated-measures studies elicit only instantaneous output.

One other subtle but important flaw in this line of research is the implicit assumption linguistically-complex language is the ultimate goal, when in fact the ability to communicate clearly, concisely, and with culturally-appropriate word choice is more desired.[14] Even at lower levels of proficiency, where such lofty goals are unattainable, the goal is not linguistically-complex language but simply successful task completion (Long, 1985). In short, focusing on

[13] Being able to design tasks that encourage the use of particular linguistic features of course is useful in its own right, in that it can help learners consolidate their existing L2 knowledge and improve the automaticity of their access to it. On the other hand, previous findings still lack the ability to provide support for the real impetus of this research, namely L2 development.

[14] See the Interagency Language Roundtable Skill Level Descriptions for Speaking at http://www.govtilr.org/skills/ILRscale2.htm#5 for an example of what is considered most important in advanced speakers of a foreign language.

linguistic complexity as the dependent variable of interest is misguided both for methodological and pedagogical reasons. The focus instead needs to be on designing and sequencing tasks that induce *task failure*, not necessarily the elicitation of linguistically complex language. In other words, gaps in knowledge need to be induced and then "filled" via either implicit or explicit feedback (see Doughty & Long, 2003; and Long, 2007 for the advantages and disadvantages of each) in order to spur L2 development. Such feedback can help learners stretch their linguistic repertoires, assuming the gaps induced and feedback provided are in sync with the learners' interlanguage development. Repeating this process throughout a course of instruction should lead to successful L2 development, which can be tested longitudinally for confirmation.

Conclusion

A considerable number of task complexity studies have emerged over the past 15 years in an attempt to help validate a framework that can help practitioners create and sequence pedagogic tasks in a manner that maximizes L2 development. Unfortunately, clear replicable findings remain elusive due to apparent flaws in both theory and methodology. While this study's findings may not necessarily help get the field closer to its elusive goal, it is hoped they will at least spur a shift in the current train of thought regarding linguistic complexity and its place within the realm of task complexity and task-based language teaching.

References

Baddeley, A. D. (1986). *Working Memory*. Oxford: Clarendon Press.

Baddeley, A. D. (1996). Exploring the central executive. *Quarterly Journal of Experimental Psychology, 18*(1), 119-129.

Berguer, R., & Smith, W. (2006). An Ergonomic Comparison of Robotic and Laparoscopic Technique: The Influence of Surgeon Experience and Task Complexity. *Journal of Surgical Research, 134*(1), 87-92.

Birdsong, D. (2005). Nativelikeness and non-nativelikeness in L2A Research. *International Review of Applied Linguistics in Language Teaching, 43*(4), 319-328.

Breen, M. (1987). Contemporary paradigms in syllabus design: Parts 1 and 2. *Language Teaching, 20*(1), 91-174.

Brown, G., Anderson, A., Shilcock, R., & Yule, G. (1984). *Teaching Talk: Strategies for Production and Assessment*. Cambridge: Cambridge University Press.

Bygate, M., Skehan, P., & Swain, M. (2001). *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing*. Essex, England: Pearson.

Cadierno, T., & Robinson, P. (2009). Language typology, task complexity and the development of L2 lexicalization patterns for describing motion events. *Annual Review of Cognitive Linguistics, 7*, 245-276.

Candlin, C. (1987). Towards task-based language learning. In C. Candlin & D. Murphy (Eds.), *Language Learning Tasks*. Englewood, NJ: Prentice Hall.

Crookes, G. (1986). *Task classification: A cross-disciplinary review (Technical Report #4)*. Honolulu: University of Hawaii, Second Language Teaching and Curriculum Center.

Crookes, G. (1989). Planning and interlanguage variation. *Studies in Second Language Acquisition, 11*(3), 367-385.

DeKeyser, R., Alfi-Shabtay, I., & Ravid, D. (2010). Cross-linguistic evidence for the nature of age effects in second language acquisition. *Applied Psycholinguistics, 31*, 413-438.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206-257). New York: Cambridge University Press.

Doughty, C. (2008). Feedback on an earlier draft of this paper. College Park, MD.

Doughty, C., & Long, M. (Eds.). (2003). *Handbook of Second Language Acquisition*. Oxford: Blackwell Publishing.

Doughty, C., & Williams, J. (1998). Pedagogical choices in focus on form. In C. Doughty & J. Williams (Eds.), *Focus on Form in Classroom Second Language Acquisition* (pp. 197-261). New York: Cambridge University Press.

Durik, A., & Matarazzo, K. (2009). Revved up or turned off? How domain knowledge changes the relationship between perceived task complexity and task interest. *Learning and Individual Differences, 19*, 155-159.

Ellis, R. (2004). The effects of planning on fluency, accuracy, and complexity in second language narrative writing. *Studies in Second Language Acquisition, 26*(1), 59-84.

Flege, J., Birdsong, D., Bialystok, E., Mack, M., Sung, H., & Tsukada, K. (2006). Degree of foreign accent in English sentences produced by Korean children and adults. *Journal of Phonetics, 34*, 153-175.

Foster, P., & Skehan, P. (1996). The influence of planning and task type on second language performance. *Studies in Second Language Acquisition, 18*(2), 299-323.

Foster, P., & Tavakoli, P. (2009). Native speakers and task performance: comparing effects on complexity, fluency, and lexical diversity. *Language Learning, 59*(4), 866-896.

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

Gilabert, R. (2007). Effects of manipulating task complexity on self repairs during L2 oral production. *International Review of Applied Linguistics, 45*(2), 215-240.

Gilabert, R., Baron, J., & Llanes, A. (2009). Manipulating cognitive complexity across task types and its impact on learners' interaction during oral performance. *International Review of Applied Linguistics, 47*, 365-395.

Givon, T. (1985). Function, structure and language acquisition. In D. Slobin (Ed.), *The Cross-Linguistic Study of Language Acquisition* (Vol. 1, pp. 1008-1025). Hillsdale: Lawrence Erlbaum.

Haerem, T., & Rau, D. (2007). The influence of degree of expertise and objective task complexity on perceived task complexity and performance. *Journal of Applied Psychology, 92*(5), 1320-1331.

Iwashita, N., McNamara, T., & Elder, C. (2001). Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information-processing approach to task design. *Language Learning, 51*(3), 401-436.

Jackson, S. (2009). Feedback at an oral presentation of this paper. College Park: University of Maryland.

Kaneko, E. (2008). *An analysis of performance by Japanese learners of English.* Unpublished Dissertation, University of Wisconsin-Milwaukee, Milwaukee.

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

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

Kuiken, F., Mos, M., & Vedder, I. (2005). Cognitive task complexity and second language writing performance. In S. Foster-Cohen, M. Garcia-Mayo & J. Cenoz (Eds.), *EUROSLA Yearbook 2005* (Vol. 5, pp. 195-222). Amsterdam: John Benjamins.

Kuiken, F., & Vedder, I. (2007). Cognitive task complexity and linguistic performance in French L2 writing. In M. Pilar Garcia-Mayo (Ed.), *Investigating Tasks in Formal Language Learning* (pp. 7-26): Multilingual Matters.

Kuiken, F., & Vedder, I. (2008). Cognitive task complexity and written output in Italian and French as a foreign language. *Journal of Second Language Writing, 17*, 48-60.

Lee, Y. (2002). *Effects of task complexity on the complexity and accuracy of oral production in L2 Korean.* Unpublished Dissertation, University of Hawaii-Manoa, Manoa.

Long, M. (1985). A role for instruction in second language acquisition: Task-based language teaching. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and Assessing Second Language Acquisition* (pp. 77-99). Clevedon: Multilingual Matters.

Long, M. (1989). Task, group, and task-group interaction. *University of Hawaii Working Papers in English as a Second Language, 8*(20), 1-26.

Long, M. (2005). Methodological issues in learner needs analysis. In M. Long (Ed.), *Second Language Needs Analysis* (pp. 19-78): Cambridge University Press.

Long, M. (2007). *Problems in SLA*. Mahwah, NJ: Lawrence Erlbaum.

Marsh, J. K. (2006). Order effects in contingency learning: The role of task complexity. *Memory & Cognition, 34*(3), 568-577.

Medina, A. (2008). *Concurrent verbalization, task complexity, working memory: Effects on L2 learning in a computerized task.* Unpublished Dissertation, Georgetown University, Washington, DC.

Michel, M., 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.

Niwa, Y. (2000). Reasoning demands of L2 tasks and L2 narrative production: Effects of individual differences in working memory, intelligence, and aptitude. Aoyama Gakuin University.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. In C. Doughty & M. Long (Eds.), *Handbook of Second Language Acquisition* (pp. 716-761). London: Blackwell.

Norris, J., & Ortega, L. (2009a). Personal Communication: discussion about relationship between independent and dependent variables in SLA.

Norris, J., & Ortega, L. (2009b). Towards an organic approach to investigating CAF in instructed SLA: the case of complexity. *Applied Linguistics, 30*(4), 555-578.

Nuevo, A. (2006). Task Complexity and Interaction: L2 Learning Opportunities and Development. Unpublished Dissertation. Georgetown University, Department of Linguistics.

Nuevo, A. (2007). *Task Complexity and Interaction: L2 Learning Opportunities and Development.* Unpublished Dissertation, Georgetown University, Washington, DC.

Nunan, D. (1989). *Designing Tasks for the Communicative Classroom*. Cambridge: Cambridge University Press.

Prabhu, N. S. (1987). *Second Language Pedagogy*. Oxford: Oxford University Press.

Revesz, A. (2009). Task complexity, focus on form, and second language development. *Studies in Second Language Acquisition, 31*(437-470).

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge: Cambridge University Press.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language Learning, 45*(2), 283-331.

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), 22-57.

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *International Review of Applied Linguistics, 43*(1), 1-32.

Robinson, P. (2006). The Cognition Hypothesis of Task-based Language Syllabus Design: Implications for Focus on Form and Aptitude/Individual Differences Research. Unpublished Paper. Aoyama Gakuin University.

Robinson, P. (2007a). Criteria for classifying and sequencing pedagogic tasks. In M. Pilar Garcia-Mayo (Ed.), *Investigating Tasks in Formal Language Learning* (pp. 7-26): 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, 45*, 193-213.

Robinson, P. (2010). Situating and distributing cognition across task demands: The SSARC model of pedagogic task sequencing. In M. Putz & L. Sicola (Eds.), *Cognitive Processes in Second Language Acquisition: Inside the Learner's Mind* (pp. 239-264). Amsterdam/Philadelphia: John Benjamins.

Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics, 30*(4), 533-554.

Samuda, V., & Bygate, M. (2008). *Tasks in Second Language Learning*. London: Palgrave Macmillan.

Skehan, P. (1998). *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Skehan, P. (2001). Tasks and language performance assessment. In M. Bygate, P. Skehan & M. Swain (Eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Harlow: Pearson Education.

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

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning, 49*(1), 93-120.

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

Tavakoli, P., & Foster, P. (2009). Native speakers and task performance: comparing effects on complexity, fluency and lexical diversity. Unpublished Paper.

Van den Branden, K. (2006). Introduction: Task-based language teaching in a nutshell. In K. V. d. Branden (Ed.), *Task-Based Language Education* (pp. 1-16). Cambridge: Cambridge University Press.

Van Patten, B. (1990). Attending to content and form in the input: An experiment in consciousness. *Studies in Second Language Acquisition, 12*(3), 287-301.

Willis, J. (1996). *A Framework for Task-Based Learning*. Essex, England: Longman.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: measures of fluency, accuracy, and complexity (Technical Report No. 17)*: Second Language Teaching and Curriculum Center: University of Hawaii.

Appendix A

Participant #: ________

CAR ACCIDENT REPORTING TASK – SET 1

Scenario: Attached are three series of four photographs, with each set depicting a car accident that you witnessed. After each accident, the police have asked you to explain what you saw. Using the pictures as a visual aid, describe to the police how each crash happened. Be sure to include as much detail as possible in your descriptions so that the police can get a very clear understanding of what took place. Also be sure to explain <u>why</u> you think each crash happened. i.e., which driver(s) and pedestrian(s) were most responsible in each accident and why?

Appendix B

Participant # ________

MAP DIRECTIONS TASK – SET 1

Scenario: Attached are three versions of a real map of Golden Gate Park in San Francisco. In this task, you will simulate being a park travel and tourism phone volunteer. Your job is to provide directions when people call the information center to ask for directions to a particular destination within the park. For each of the versions, your job is to explain how the tourists can get from their current location (Point A) to the their goal destination (Point B). Because the tourists do not have maps and are on foot, it is very important to be as explicit as possible in explaining how to get from A to B, citing landmarks and any other information (e.g., cross streets, the type of road or trail, etc.) you think would help them visualize and understand how to get to their final destination.

Version 1A

Explain to the tourist how to get from the top of Strawberry Hill in the middle of Stow Lake (A) to Anglers Lodge (B). You are free to choose the route you send the tourist, but remember to be as explicit as possible, citing landmarks and other helpful information, because he/she does not have a map and will have only your directions to guide him/her (NOTE: Please trace your route with a pen while you are describing it verbally. i.e., There is no time allotted for planning your route before you begin describing it, so please trace the route and describe it simultaneously without pre-planning it).

Version 1B

Explain to the tourist how to get from the Barbecue Pits (A) to the front entrance the of the deYoung Museum (B) following the green-highlighted route. Remember to be as explicit as possible, citing landmarks and other helpful information, because the tourist does not have a map and will have only your directions to guide him/her.

Version 1C

Explain to the tourist how to get from the back side of Metson Lake (A) to the Lawn Bowling fields (B) following the orange-highlighted route. Remember to be as explicit as possible, citing landmarks and other helpful information, because the tourist does not have a map and will have only your directions to guide him/her.

Appendix C

Background Questionnaire

1. Gender (circle one): Male Female 2. Date of Birth: Month _____ Year 19____

3. Place of Birth: US Other: _____________

4. If born outside the US, at what age did you move here? ______ years _______ months

5. What language did your family speak to when you were an infant? _____________________

If your primary language was other than English when you were an infant, when and in what context were you first exposed to English extensively? (examples: Age 3, attended daycare center; Age 5, attended kindergarten, etc)

Age of Exposure: ______ Context: ____________________________

6. List the languages you speak to some degree (1 = most comfortable; 4 = least comfortable).

1:______________ 2: ____________ 3: ____________ 4: _____________
(most comfortable) (least comfortable)

7. Have you ever lived in a foreign country? Yes / No

Year	Length of Stay	Purpose

8. Have you taken any foreign language classes? Yes / No

Year	Length of instruction & How often did you meet?	Location and Class Title

9. Highest Degree Completed: High School Associate's Master's Doctorate

If you attended college, what was your major: ___________________________

Appendix D

Directions

The following tests have been developed by removing the second half of every second word in a text. You are supposed to reconstruct the texts.

Example: My name is Tom. I'm t__ oldest ch__ in m__ family. I ha__ a sister a__ two brot___.

Answer: My name is Tom. I'm the oldest child in my family. I have a sister and two brothers.

Writing Development

The representation of thought was achieved by means of oral signs, mutually understood by the group who recognized the same system of representation. This or___ manifestation w___ later o___ preserved i___ the fo___ of draw___ and writ___, so th___ each comm____ left beh___ a record o___ its cul___. But wri___ is n___only a w___ to pres___ memory; i___ is al___ the sym___ of a cul___. This c___ be cle___ observed i___ the sys___ of wri____, which were historically developed. Writing was later developed into artistic and aesthetic forms of knowledge and communication and whether it developed so do calligraphy.

Send Me a Postcard

Postcards always spoil my holidays. Last sum___, I we___ to It___. I vis___ museums, a___ sat i___ public gar___. A frie___ waiter tau___ me a f___ words o___ Italian. H___ lent m___ a bo___. I re___ a f___ lines, b___ I d___ not under___ a wo___. Every d___ I tho___ about post___. My holi___ passed qui___, but I did not send any cards to my friends. On the last day I made a big decision. I got up early and bought thirty-seven cards. I spend the whole day in my room, but I did not write a single card!

Antismoking campaigns

Some people believe that cigarette smoking is dangerous and should be considered a health hazard. They wa___ their gover___ to cre___ antismoking prog___. People dif___as t___ how st___these antis___ campaigns sho___ be. So___ of stro___ campaigns wo___ try t___ completely elim___cigarette smo___. Supporters o___ these prog___ would t___ to b___ cigarette smo___completely i___ public pla___. Others wo___ try on___ to rest___ the number of places where people could smoke. Such restrictions would not try to eliminate public smoking completely, but only to curb smoking by reducing cigarette consumption.

Grandparents and grandchildren

Recent studies indicate that grandparents and grandchildren are better off when they spend large amount of times together. Grandparents gi__ children lo___ of affe___ with n__ strings atta___, and t__ children ma___ the grandp___ feel lo___ and nee___ at a ti___ when t__ society m___ be tel___ the ol___ people th___ they a___ a bur___ Grandparents a___ a sou___ of stre___ and wis___ and he___ ease t___ pressure bet___ children and their parents.