Applied Linguistics 30/4: 510–532

doi:10.1093/applin/amp047 Advance Access published on 8 December 2009

Modelling Second Language Performance: Integrating Complexity, Accuracy, Fluency, and Lexis

PETER SKEHAN

Chinese University of Hong Kong

Complexity, accuracy, and fluency have proved useful measures of second language performance. The present article will re-examine these measures themselves, arguing that fluency needs to be rethought if it is to be measured effectively, and that the three general measures need to be supplemented by measures of lexical use. Building upon this discussion, generalizations are reviewed which focus on inter-relationships between the measures, especially between accuracy and complexity, since positive correlations between these two areas have been less common in the literature. Some examples of accuracy–complexity correlations are reviewed. The central issue here is how to account for these correlations, and so the discussion explores rival claims from the Cognition and Trade-off Hypotheses. It is argued that such joint raised performance between accuracy and complexity is not a function of task difficulty, as the Cognition Hypothesis would predict, but that instead it reflects the joint operation of separate task and task condition factors. Extending the theoretical discussion, connection is made with the Levelt model of first language speaking, and it is proposed that the results obtained in the task-based performance literature can be linked to this model, modified to take account of differences between first and second language processing, particularly as these stem from differences in the underlying mental lexicons.

INTRODUCTION

There are a range of approaches to accounting for performance on language-learning tasks. It is useful therefore to set out the assumptions that underlie the analysis presented here. First, it is assumed that attentional capacity and working memory are limited. This is a fairly standard account from contemporary cognitive psychology, with a long tradition of relating working memory operations to attentional availability (Baddeley 2007). Secondly, if one relates this general account of attentional limitations to second language performance, some specific consequences follow. Successful performance in task-based contexts has often been characterized as containing:

- more advanced language, leading to *complexity*;
- a concern to avoid error, leading to higher *accuracy* if this is achieved; and
- the capacity to produce speech at normal rate and without interruption, resulting in greater *fluency*.

If performance in each of these areas, complexity, accuracy, and fluency (CAF), requires attention and working memory involvement, then committing attentional resources to one may have a negative impact on others. In particular, one can propose that there is a tension between form (complexity and accuracy), on the one hand, and fluency, on the other. Then, within form, one can contrast attention directed to using challenging language (complexity) relative to conservative, less advanced language, but greater accuracy (Skehan 1998). One could portray these tensions in the form of a Trade-off Hypothesis, which would predict that committing attention to one area, other things being equal, might cause lower performance in others.

To express the Trade-off Hypothesis in this way, though, is to propose a default position, which simply argues that there are constraints on performance which have to be accounted for, *post hoc*, on a case-by-case basis. If the Trade-off Hypothesis said no more than this, then it would be, as Robinson (2007) claims it is, vacuous. For the hypothesis to have substance, more needs to be said about the precise ways in which the performance areas enter into competition, and what influences there are which mediate this competition. Explicating the nature of the competition for attentional resources, and how it can be overcome, will be one of the central aims of this article. As we will see, one can address this issue at a theoretical level, and also by examining available empirical findings. A brief account of previous research helps set the scene for the contributions of the present article. The focus is on the independent variables in task-based performance, which have systematic influences on CAF.

Earlier research within a complexity–accuracy–fluency framework supported generalizations such as the following:

- accuracy and fluency, but not complexity, are raised in personal information exchange tasks;
- there is higher complexity, but lower accuracy and fluency, on narrative tasks; and
- pre-task planning produces greater complexity and fluency.

These findings are broadly consistent with a trade-off interpretation, since they suggest higher level performance in two out of three areas, but not typically in all three.

This earlier research, operating with an underspecified characterization of task types, gave way to finer grained approaches, showing systematic influences of task characteristics and task conditions. The newer research studies (e.g. reviewed in Skehan (2001)) offered generalizations such as:

- tasks based on concrete or familiar information advantage accuracy and fluency;
- tasks containing clear structure advantage accuracy and fluency;
- interactive tasks advantage accuracy and complexity;
- tasks requiring information manipulation lead to higher complexity; and

- post-task conditions such as public performance or transcription of one's own performance raise accuracy.

This research is supportive of a Trade-off Hypothesis, in that it indicates that simultaneously advantaging all three (CAF) performance areas is unusual. It is also clarifying, in that most 'two influence' results do indeed suggest that fluency can be accompanied by either accuracy or complexity, but not both. The research, though, is limited, in that while it offers generalizations consistent with attentional limitations, the explanatory force of these accounts is lacking.

It is the purpose of this article to address these shortcomings, and to offer a theoretically motivated and empirically grounded account of CAF in second language performance. The article is organized into four main sections: two somewhat preliminary and empirical, and two more theoretical. First, new measures of task-based performance will be discussed. It will be proposed that it is essential to be more sophisticated in the measurement of fluency, and vital to incorporate some measure of lexis into task performance. Progress in each of these areas is important to understand the psycholinguistics of second language speech production and the inter-relationship between the performance areas. Secondly, there are emerging generalizations which go beyond those given above. We now have additional interesting claims that can be made about the effects of task characteristics especially. In particular, there are findings of accuracy–complexity relationships, which might pose some problems for any trade-off-based account (Skehan 1998), and, contrastingly, provide support for the Cognition Hypothesis (Robinson 2001).

This more empirical and descriptive coverage leads to the third section, more theoretical in nature, and which explores the plausibility of Trade-off and Cognition accounts of these generalizations. It will be argued that it is task characteristics and task conditions, in particular combinations, which predict accuracy–complexity correlations when they occur, and that it is not task complexity, *per se*, as Robinson argues, that accounts for this relationship. Fourthly, as a broad explanatory framework, Levelt's model of speaking is used to organize the research findings, and to locate the different effects within the Conceptualizer and Formulator stages of the model.

NEW MEASURES OF TASK PERFORMANCE

Researching task performance and measuring CAF is a relatively recent aspect of applied linguistics, arguably starting with Crookes (1989), but there is clear scope to develop additional measures. Here the focus will be on two measures of *general* performance, specifically of fluency and lexis.

Regarding fluency, a range of measures are available, broadly examining:

- breakdown (dys)fluency, indexed by pausing;

- repair (dys)fluency, indexed by measures such as reformulation, repetition, false starts, and replacements; and
- speed, with measures such as syllables per minute.

One can also use higher order measures such as length-of-run, as an indicator of automatization in language performance (Towell *et al*. 1996; Towell and Dewaele 2005).

An interesting feature of the measurement of fluency is to compare second language speakers to first language speakers, who, of course, also pause (see Skehan and Foster 2008 and Skehan 2009a for discussion). It has been argued (Davies 2003) that what distinguishes native and non-native pausing is more likely to be *where* pauses occur rather than *that* they occur. Davies (2003) suggests that it is fruitful to distinguish between pauses that occur at mid-clause and those that occur at the ends of clauses, especially at Analysis of Speech (AS) boundary points (Foster *et al*. 2000). From a database of first and second language learners doing identical tasks, this issue was investigated to explore how the two groups contrast (Skehan and Foster 2008). The relevant data are shown in Table 1.

The personal task here (and all tasks involved pairs of students) was to explain to one's partner how to get to one's home to turn off an oven that had been left on. The narrative task was to make up a story that linked a series of pictures which had no obvious storyline, but did have common characters. The decision-making task was to agree on (as if a judge) the appropriate sentence for a series of crimes. Those without planning opportunities had to complete the task once they had understood instructions, while those with planning opportunities were given 10 minutes to prepare, during which time they could take notes, but these notes were taken away before the actual performance.

Table 1: Comparison of native and non-native speaker pausing on three tasks

	Personal task		Narrative task		Decision-making task	
	−Plan	+Plan	−Plan	+Plan	−Plan	+Plan
Native speakers						
AS pauses	*2.8*	*1.4*	*4.2*	*2.1*	*3.6*	*0.8*
Mid-clause pauses	1.1	1.3	*1.3*	*2.6*	*1.6*	*0.4*
Ratio: AS-to-mid	(2.5)	(1.1)	(3.2)	(0.8)	(2.3)	(2.0)
Non-native speakers						
AS pauses	1.6	1.5	*3.9*	*1.4*	*3.7*	*1.6*
Mid-clause pauses	2.6	2.1	*3.8*	*4.8*	*4.8*	*2.7*
Ratio: AS-to-mid	(0.6)	(0.7)	(1.0)	(0.3)	(0.8)	(0.6)

Significances are shown through italicization.

Pre-task planning is similar in its effects with native and non-native speakers for end-of-clause boundaries—planners generally pause less (with univariate tests used, following Multivariate Analyses of Variance (MANOVAs)). Interestingly, native speakers, in two out of three cases, pause *more* than non-natives at this point under the non-planning condition. Native speakers, in other words, seem to regard AS boundaries as a natural place for online planning. The figures for mid-clause pausing, though, are very different. Here the native speakers pause very clearly less mid-clause, in fact, hardly at all, whereas for non-native speakers, this is the more frequent pause location. The difference between the two groups is shown even more clearly by the AS-to-mid-clause ratio, shown in parentheses in Table 1. Five of the six native speaker ratios are above unity (with one marginally below), where five of the six non-native speaker ratios are below. If native speakers are taken to indicate a baseline level of performance, the natural place to pause is an AS boundary, and that sufficient online regrouping can be achieved with such pausing (one also assumes that pauses located at this point are advantageous for interlocutors). The non-native speakers, in contrast, seem to have pauses thrust upon them, in the middle of what become turns lacking in smoothness, probably as they are required to handle unforeseen lexical choices. These findings suggest that, to make sense of CAF, the characterization of fluency needs to become subtler and deeper.

The last few paragraphs have proposed measurement improvements for an established performance area. But there is another area, lexis, which has been strikingly absent in task research. This is a serious omission. The lexis–syntax connection is vital in performance models (Skehan 2009a) such as Levelt's, and lexis represents a form of complexity that has to be assessed in second language speech performance if any sort of complete picture is to be achieved.

Generally, research into lexical measures (Daller *et al*. 2003) distinguishes between text-internal measures (so called because the text itself is sufficient for their calculation) and text-external measures (which require some sort of general reference material, usually based on word frequency). The former is typically measured through some sort of type–token ratio (TTR), but with the qualification that since TTRs are strongly related to text length, there has to be a correction made. A generally acceptable measure is *D*, as calculated by the VOCD subprogram within Computerised Language Analysis (CLAN) (MacWhinney 2000), as an index of lexical variety (Malvern and Richards 2002, Richards and Malvern 2007).

In contrast are measures of what is called lexical sophistication (Read 2000). These take frequency lists from corpus analysis and then compute how many words defined as difficult (on the basis of lower frequencies) are used in a text. The most well-known measure of this sort is Laufer and Nation's Lexical Frequency Profile (1999). At the time the research reported on here was being done, the Laufer and Nation approach lacked spoken language corpus-based frequency lists. Accordingly, another approach was used,

following Meara (Meara and Bell 2001; Bell 2003). This uses a Poisson distribution (a distribution appropriate for events which have low frequency levels). A text, for example the transcribed performance of someone doing a task, is divided into 10-word chunks, and then, for each 10-word chunk, the number of 'difficult' words is calculated. (Difficult is usually defined in terms of a threshold frequency; in the present case, this was 150 occurrences or fewer per million words.)[1] The number of 10-word chunks which contain no difficult words, or one difficult word, or two difficult words, and so on, is calculated. Then a statistic, lambda, is calculated, which represents the best fit of the distribution of numbers of difficult words. In other words, it is assumed that the higher the level of lambda, the greater the ability of the speaker to mobilize a wider vocabulary, drawing on less frequent words (see Skehan 2009a for further details). In pilot studies with this measure, it was established that *D*, the measure of lexical variety, and lambda, as a measure of lexical sophistication, did not correlate (Skehan 2009a). In addition, the value of *D* was not particularly revealing regarding task effects. Accordingly, we focus here on the findings related to lambda. The relevant values are shown in Table 2.

Planning opportunities make a big difference to the speaker's capacity to mobilize less frequent words. With the exceptions of the non-native speaker narrative and the native speaker personal task comparisons, the lambda scores in each case are significantly higher for planners than non-planners (and are in the predicted direction for the non-significant results). All four significant results generate effect sizes which are large (judged through Cohen's *d*). Planning time enables speakers, whether native or non-native, to make different lexical selections (the TTR, *D*, did not generate similar significant differences for planning). In addition, there are major task differences, with the narrative task generating the highest values for lambda (and the largest effects for planning), and the decision-making tasks generating the lowest.

Table 2: Lexical sophistication (lambda) and task type for native and non-native speakers

	Personal task		Narrative task		Decision-making task	
	−Plan	+Plan	−Plan	+Plan	−Plan	+Plan
Native speakers	1.27	1.47	1.46	1.95	0.76	0.94
	(0.27)	(0.27)	(0.41)	(0.52)	(0.13)	(0.25)
	(*N* = 13)	(*N* = 14)	(*N* = 16)	(*N* = 13)	(*N* = 14)	(*N* = 18)
Non-native speakers	0.94	1.15	1.00	1.18	0.54	0.79
	(0.28)	(0.23)	(0.25)	(0.36)	(0.25)	(0.24)
	(*N* = 16)	(*N* = 11)	(*N* = 12)	(*N* = 11)	(*N* = 14)	(*N* = 15)

Standard deviations are given in the first line of parentheses in each cell, cell sizes in the second.

This pattern is repeated with different narrative and decision-making tasks, as reported in Skehan and Foster (2008). It is interesting that the narrative task is the most input-driven, and the least negotiable. It seems to require the use of difficult-to-avoid lexical items. The decision-making task, in contrast, is the most negotiable, and also interactive. Speakers do not pack their contributions here with as many less frequent words.

One final aspect of these performances is interesting. With the native speakers, there is a positive correlation between lambda and structural complexity, indexed by subordination. The correlations are 0.43 with the personal task ($p < 0.05$), 0.57 for the narrative ($p < 0.001$) and 0.21 for the decision-making (not significant). With the non-native speakers, this correlation is negative (as, for them, is the correlation between lambda and accuracy). The contrast is intriguing. For native speakers, making more demanding and less obvious lexical choices does not cause problems. Consistent with the Levelt model, it even seems to drive syntax to some degree, as these lexical choices are more likely to require more complex syntactic clothing. But for non-native speakers, it seems more likely that making more complex lexical choices derails syntax in two ways—making it less complex and also less accurate—as problems occur through the additional difficulties that arise from lexical retrieval. The attention consumed here poses problems for the way more complex syntax is used, and also for avoidance of error. All these measures relate to Levelt's Formulator stage, implicating lemma retrieval and the assembly of the current utterance. They also enable us to explore the role of lexis as a driving force for syntax, as we will see below.

WIDER ANALYSES OF TASK INFLUENCES

It is also possible to propose additional generalizations to those offered earlier. Three will be covered here. The generalizations are interesting in their own right, but they also illustrate the usefulness and relevance of the Levelt model. They also bear upon rival claims of the Trade-off and Cognition Hypotheses.

The first generalization concerns information manipulation and organization. In earlier research, it was argued that the need to manipulate and reorganize information leads to a pressure to produce more complex language. For example, Foster and Skehan (1996) used a narrative task in which learners were required to devise a story to link a series of pictures which contain common characters but no obvious storyline. This task contrasts with most narratives (Ortega 1999) which simply provide a series of pictures (often in cartoon format) or a video (Skehan and Foster 1999) and simply require retelling. The task requiring the development of a storyline (Skehan and Foster 1997) requires the speaker to create the storyline itself which links the pictures, and so necessarily requires manipulation and arrangement. The task in question generated higher complexity scores than other, 'pre-organized' narratives. This result raised the possibility that other forms of information manipulation might also lead to higher complexity scores.

One such could be the need, within a narrative task, to integrate foreground and background information if the story is to be told effectively. Consequently, Tavakoli and Skehan (2005), using a conventional narrative telling based on a series of organized pictures, also included a picture series which required such integration. This picture series generated significantly higher complexity scores. This finding has been replicated by Foster and Tavakoli (in press) who show that the need to make connections between picture elements is a dependable way of increasing language complexity. Information manipulation and integration seems to require more extensive Conceptualizer use, which is reflected in a more complex pre-verbal message, and this, in turn, leads to the need to formulate more complex language. Task design, in other words, can influence the level of language complexity appropriate for a particular task.[2]

Next, we consider a generalization based on the impact of lexis on other performance areas. It concerns the way that other dimensions of performance are affected when a task makes it difficult to avoid using less frequent lexical elements. Relevant data from a series of studies is shown in Table 3. The table covers three studies and reports results for narrative and decision-making tasks, in two cases, and a narrative only in one. Skehan and Foster (1999) used a video-based narrative, but the other two studies used picture series for the narratives. The decision-making tasks were a Judge task (agree on the appropriate sentences for a series of crimes), in Foster and Skehan (1996), while Foster and Skehan (MS) used an Agony Aunt task (agree on the advice to give to the writer of a letter to an Agony Aunt).

A narrative is necessarily input-driven, and unforgiving in what needs to be covered. More interactive, decision-making tasks allow greater latitude. Unavoidability, or at least, infrequency of lexis is taken to be reflected here in the lambda figures. These show a clear difference between narrative tasks on the one hand, and decision-making tasks, on the other. Narratives seem to push second language speakers (and first language speakers, Skehan 2009a) into using less frequent lexis, presumably as they are responding to the events within the narrative. Decision-making tasks, in contrast, although

Table 3: Unavoidable lexis and task-based performance

Study	Narrative			Decision-making		
	Lambda	Mid-clause pauses	EFC	Lambda	Mid-clause pauses	EFC
Foster and Skehan (1996)	1.14	4.52	0.56	0.65	3.82	0.60
Skehan and Foster (1999)	1.66	7.32	0.45	–	–	–
Foster and Skehan (MS)	1.45	4.23	0.54	0.48	3.49	0.68

EFC = Error Free Clauses.

purposeful, for example agreeing on advice to Agony Aunt letters, enable different methods of expression as speakers take different routes to expressing their positions. The differences in lambda scores are dramatic and give clear indications of the consequences of task requirements.

More interesting is the potential impact of the need to mobilize less frequent lexis on other aspects of performance. We will examine the findings first, and then relate them to psycholinguistic functioning. Using the data from Skehan and Foster (1997) the lambda scores correlated with accuracy at −0.48 (personal task), −0.23 (narrative), and −0.23 (decision-making). Although the second and third correlations are not significant, the results taken together suggest that there is a weak negative relationship. Regarding lambda and fluency (mid-clause pausing), the correlations are 0.57 (personal), 0.36 (narrative), and 0.20 (decision-making). The first two correlations are significant, and suggest that the higher the lambda scores, the more mid-clause pausing there is. In other words, the need to retrieve rarer lexical items seems to have a cost in terms of how error can be avoided and a smooth flow of speech maintained—such retrieval creates processing demands and consumes attentional resources. Achieving accuracy is thereby compromised, an interesting finding in that one would imagine that more complex tasks (cf. Robinson *et al*. this issue) would be associated, other things being equal, with the use of more sophisticated lexis (Read 2000). Interestingly, the highest value of lambda, 1.66, (and the lowest accuracy value, 0.45) is with the video-based narrative (Skehan and Foster 1999). Hence, there may be important differences within narratives also which are dependent on performance conditions.

These results, too, fit within the Levelt model. In the Levelt model, the Conceptualizer delivers the pre-verbal message to the Formulator, which then has to undertake the process of lemma retrieval that can subsequently drive syntactic encoding (since the syntactic information is stored in the lemma). With native speakers, this lemma retrieval proceeds smoothly, enabling parallel processing (as the Formulator deals with previous Conceptualizer cycles, while the Conceptualizer simultaneously attends to the current cycle). Native speaker mental lexicons are extensive, and well-organized. Accordingly, the demands the pre-verbal message makes are met without undue difficulty. In contrast, the pre-verbal message developed by the non-native speaker arrives at a Formulator equipped with access to a smaller mental lexicon and with significantly less organization and elaboration. The result is that the pre-conditions for smooth parallel processing are not met, and the Formulator stage is more effortful, often requiring repair and replacement. Pre-verbal messages which require more difficult lexical items make greater demands on this mental lexicon and the operation of the Formulator as an automatic process is particularly disrupted. If we assume that accuracy is the consequence of attention being available for Formulator operations, then needing to access more difficult lexical items will be particularly hard to handle.

The third area for cross-study generalization concerns task organization. Organization will be defined here in relation to macrostructure of component elements, often involving schematic knowledge, or simply knowledge of text structure. This can be operationalized in a number of different, but related ways. This could simply be a story with a clear beginning, middle, and end. It could also be a schematic structure, or well known 'script', such as visiting the doctor. Or it could be a discourse structure, such as Labov's analysis of narratives. All of these have in common that the details of what is being done at any one moment fit in easily to the wider macrostructure, so that less processing is required during speaking. Early examples of research studies exploring this are Foster and Skehan (1996), with the need to instruct someone on a familiar journey one makes, as well as Skehan and Foster (1997) with cartoon picture series with a clear beginning, middle, and end. Subsequently, Skehan and Foster (1999) compared video-based retellings (of Mr Bean stories) contrasting a structured narrative (following the well-known restaurant script) with a less structured story (a chaotic round of Crazy Golf).

More recently, Tavakoli and Skehan (2005) attempted to operationalize structure more predictively. Picture series were analyzed into four levels of structure, operationalized in terms of the number of individual pictures (other than first and last) which could be exchanged without impairing the storyline. The more structured picture series followed a problem–solution pattern (Hoey 1983). The levels did not generate an ordered four-step scale, but there was still a clear distinction between unstructured and structured performances. Tavakoli and Foster (2008) used a similar research design and reported the same structured–unstructured contrast. Both studies deliver the same general performance. The degree of structure in a task does not tend to influence subordination-based measures of complexity (with an exception to be covered below), but it does consistently have a positive impact on fluency and accuracy. It appears that the support of having a wider macrostructure enables the speaker to operate within helpfully limiting parameters. Conceptualizer-outcome, that is the pre-verbal message, is the basis for communication, but the ideas that need to be expressed fit into the wider structure fairly clearly, and so do not take much attentional capacity. The consequence is that more attention is available for ongoing performance, and so the performance dimensions which are more affected by this immediate, Formulator-linked attention, benefit. These, it seems, are more likely to be fluency and accuracy. We can recap these three generalizations as follows:

- Need to organize and integrate more demanding information
  - Pressure on Conceptualizer and push for greater language complexity
- Need to retrieve and make available more difficult lexis
  - Pressure from pre-verbal message, difficulties for lemma retrieval, and pressure on Formulator
- Capacity to draw on clear macrostructure
  - Easing of Conceptualizer and release of attention for Formulator operations

In the introductory section, generalizations from earlier task research were presented, and these were portrayed as (i) consistent with a limited capacity view of attention, and (ii) suggesting a default contrast between certain performance areas. These generalizations lack theory and predictive power. The last two sections, though, on new measures and new generalizations, have shown the centrality of lexis, the need for subtler measures of fluency, and have started to make contact with the Levelt model of speaking. The Conceptualizer stage seems to be influenced by the ideas that need to be expressed. Any influences which push for the need to formulate more complex ideas raise the importance of this stage, and lead to increases in the structural complexity of the language which is used. The integration of foreground and background information exemplified this. Lexical Formulator influences will impact upon the way the pre-verbal message interfaces with mental lexicon access and lemma retrieval. In this case, involvement of a mental lexicon that can handle the demands of the pre-verbal message will be associated with the capacity to produce fluent and accurate language through parallel processing. More demanding lexical pressures with second language speakers will disrupt this process and impair harmonious real-time parallel speech production, with delays, a need for repair, and heavy attentional demands. The difficulties of narrative tasks, and the lexical demands they make, illustrate this. Both these processes involve pressures on the resources a speaker has available for second language performance. But the third area, structure, reflects a more supportive influence. Here greater structuring in a task can ease attentional demands, and enable the speaker to channel attention to immediate performance. The result is that more attention is available for ongoing Formulation, and consequently the areas the Formulator is responsible for, accuracy and fluency, are more effectively handled. So we see that within the constraints of limited attentional capacities, difficulties manifest themselves in non-random manner—the stages of speech production give us some insight into which CAF areas are affected by which influences.

COMPLEXITY–ACCURACY RELATIONSHIPS: TRADE-OFF AND COGNITION

The generalizations from previous sections provide new ways of explicating trade-off effects while linking performance to a model of speech production. But the sharpest challenge for a trade-off account, and where it differs most clearly from the Cognition Hypothesis, is where and why it predicts a correlation between accuracy and complexity in performance. In addition to this Trade-off vs. Cognition Hypotheses contrast, this specific relationship is revealing with respect to CAF relationships more generally; how these dimensions can be accounted for theoretically; and what sorts of influences can be manipulated to raise or lower performance in each of the two areas.

We have seen that a trade-off interpretation proposes that the 'natural' tension when resources are limited is between accuracy and complexity. This contrasts with the Cognition Hypothesis, which proposes that task complexity will be associated with increases in complexity *and* accuracy. A key issue, therefore, is what the evidence is regarding situations where complexity and accuracy are simultaneously advantaged. As it happens, researchers within the Cognition Hypothesis framework have very rarely provided evidence of this sort, with many studies reporting raised accuracy, but no increase in complexity (Gilabert 2007a; Kuiken and Vedder 2007a,b; Michel *et al*. 2007). There is even the possibility that the accuracy effects themselves are due to lexical, rather than syntactic–morphological differences (Kuiken and Vedder 2008), and they seem to have greater importance in narrative, but not interactive tasks (Michel *et al*. 2007). Ishikawa (2007) does report significant effects for specific accuracy measures (although general accuracy measures do not reach significance), and for complexity. This, then, is an important study but it involves written, not oral performance, and is effectively also a study of planning, since five minutes preparation time was given, that is we do not know if these results would apply to a non-planning condition. Gilabert (2007b), too, presents data supportive of the Cognition Hypothesis. Three tasks were used, and although no complexity scores are given, he does report greater accuracy for narrative and instruction giving tasks, where, interestingly, task complexity was operationalized in different ways. The lack of complexity scores, though, means that this study can only provide limited evidence in favour of the Cognition Hypothesis.

In this section, we will discuss three studies which do actually report joint accuracy–complexity effects. The first of these studies investigated different types of planning (Foster and Skehan 1999). The planning was either solitary (as in most previous studies), or group-based, teacher-led (and focused on either content or language). General planning effects were found comparable with other studies (and see Ellis this issue). But the teacher-led planning was particularly interesting, because in this condition both complexity and accuracy were raised, unlike the other planning conditions. This result suggests that planning has more than one function (Ortega 2005). Ellis (2005), for example, discusses distinctions between planning-as-rehearsal and planning-as-complexification (as well as online planning). What seems to be happening is that teacher-led planning was able to combine these different aspects of planning, and so both complexification and rehearsal were handled more effectively. The result was that planning-as-complexification led to greater levels of language complexity, while planning-as-rehearsal was associated with increases in accuracy. The two, frequently competing, areas can go together therefore, but they are mediated by planning opportunities of a particular kind, in this case orchestrated by the teacher, with the result that attentional limitations during the subsequent task performance are eased. Carefully organized planning was used, in other words, to overcome normal performance constraints.

Another study which reports a joint accuracy–complexity effect is Tavakoli and Skehan (2005). Several variables were manipulated in this study, but the focus here is the effect of task structure, in the context of a narrative retelling. Task structure yielded results following predictions, which is greater fluency and accuracy. But the joint raising of accuracy and complexity in one of the tasks (structured, and with a need to integrate foreground and background information) was particularly interesting. A tentative *post hoc* interpretation was that there was a need to incorporate not simply the 'driving' storyline embodied in foreground information, but also to bring into play background information from earlier pictures in order to make sense of the events in later pictures in the story. This might account for the way both accuracy and complexity were raised. Tavakoli and Foster (2008) and Foster and Tavakoli (in press) built on this *post hoc* interpretation to design studies specifically aimed at manipulating these two variables. They confirmed the findings: both accuracy and complexity were raised when there was a structured task which also required the integration of foreground and background information.

Hence, two performance areas can be jointly affected, but their raised levels are the result of separate influences. Structure tends to lead to a greater accuracy. Information manipulation leads to greater subordination-based measures of syntactic complexity. Each influence is a task design feature, but they seem to be able to operate additively. When this happens with a picture series that is both structured *and* requires information integration, complexity and accuracy are both raised. This set of studies suggests that under supportive conditions the constraints of limited attentional capacity can be overcome to some extent. The key issue regarding the Cognition Hypothesis, though, is that a task which is simultaneously structured and information integrating cannot be regarded as a more complex task (as the Cognition Hypothesis requires). In other words, it is not task complexity *per se* which drives the joint increases in these performance areas: it is the separate but additive impact that each of the influencing variables has.

The final study concerns the effects of a post-task condition. Skehan and Foster (1997) explored whether learners' knowledge while doing a task that they *might* be required to re-do the task publicly later would lead them to prioritize accuracy. The results provided limited support for this hypothesis. In a subsequent study (Foster and Skehan, MS), they used a different post-task operationalization, the need to transcribe one minute of one's own performance. This was a more personally applicable and more language-related implementation of a post-task activity. In addition, accuracy was measured slightly differently focusing on the length of clause that could be accurately handled rather than the proportion of error-free clauses (see Skehan and Foster 2005 for discussion). In this subsequent study, a significant accuracy effect was found for both decision-making and narrative tasks. This suggests that learners can prioritize attention to particular areas. The post-task

condition induced changes in attention management on the part of second language speakers.

But the second of these post-task studies (Foster and Skehan MS) proved even more interesting. For the decision-making task, not only did the post-task condition produce significantly higher accuracy, it also produced significantly higher language complexity. This was unpredicted, but very interesting. At first sight, this finding is supportive of one of the central predictions of the Cognition Hypothesis—a simultaneous effect on accuracy and complexity. It should be recalled, though, that the Cognition Hypothesis proposes both that (i) raised accuracy and complexity are brought about by greater task complexity, and that (ii) this joint increase in performance is not constrained by attentional limitations. As a first point, we have to consider whether doing a task with a post-task activity such as transcription can be regarded as making the task more complex. This is difficult to see. It is the same task that is being done, after all, and the change in task performance seems to be the result of an interpretation by the task participants of what they should do.[3] The participants seem to be the ones, in other words, who shift their goals and prioritize certain performance areas. It also seems to be the case that they need to prioritize attention in this way—the baseline performance from control groups suggests that this is not a natural thing to do with such task. This is, in other words, a manipulated effect.

This result poses problems for both the Cognition Hypothesis and the Trade-off Hypothesis, and what we need to do is explore how serious these problems are. For the Cognition Hypothesis we have already explored the dissociation between task complexity, on the one hand, and accuracy–complexity relationships, on the other. The results therefore appear to support one aspect of the Cognition Hypothesis but without involving its central tenet—that it is task complexity which simultaneously drives structural complexity and accuracy. Turning to a trade-off account, the results suggest that the (empirically founded) generalization that accuracy and complexity rarely go together, and that this reflects the consequence of limited attention, may not always apply. But there is a complicating issue to be addressed in this regard. Research so far which bears upon the Cognition Hypothesis has sought to explore whether significant effects can be found simultaneously for accuracy and complexity. If such effects are found then support is inferred for the Cognition Hypothesis.

But one could apply a stronger test here than simply the demonstration of joint experimental effects. This is because one cannot assume from such effects that group effects are mirrored at the individual level. It is also possible that one can obtain significant effects for both performance areas *where each effect is accounted for by different participants*. In other words, to support the Cognition Hypothesis, not only does one need to establish experimental effects—one also needs to demonstrate that there is a correlation between performances in the two areas. If we examine correlations between accuracy and complexity in Foster and Skehan (MS), the results are that complexity and accuracy correlate at 0.26 for the decision-making task (not significant, but one has to take

account of the small sample size of only 16), and negatively, at −0.13 for the narrative task. The post-task condition did seem to work in orienting participants towards form, but different participants seemed to gravitate to different aspects of form. Some emphasized structural complexity and some emphasized accuracy, but in general there was no suggestion that both areas were advantaged together. This sits comfortably with a trade-off account because it appears that at the individual level, prioritizing accuracy *or* complexity is the norm.

APPLYING THE LEVELT MODEL TO SECOND LANGUAGE TASK PERFORMANCE

The findings which have accumulated over the last 20 years or so about second language task performance are extensive and provide a foundation for theorizing about the influences on such performance especially through connections with the Levelt (1989, 1999) model of first language speaking. It is important, though, to try to systematize the discussion of this potentially relevant model, and to try to identify patterns in the influences on *second* language performance. In that respect a first approach is represented in Table 4, which organizes a range of task-based findings in terms of the model.

In the centre of the table, the influences are organized following three major stages from the model: Conceptualization, Formulation: Lemma Retrieval, and Formulation: Syntactic Encoding (all shown in bold). These three stages provide a good structure for locating the range of influences that are available and which have been described in previous sections. Next, though, we need to distinguish the influences shown on the left-hand side from those on the right-hand side. On the left are influences which do one of the two things (but both of which add to the difficulty of a task). First, some influences may complexify the performance that results. All the influences shown at the Conceptualizer stage fall into this category. They lead to the speaker developing a message which is more demanding of cognitive resources, and which requires more active working memory use during speech production. But second, we have the influences which are relevant at the Formulator stage. These generally have the effect of pressuring performance. At the lemma access stage, it is proposed that two related influences are important: the need to use less frequent lexis and the non-negotiability of the task. Each of these creates pressure because the learner is forced to do something difficult during online speech production and wrestle with the problems that come from more effortful and slower access to the information stored in lemmas. In other words, both influences are concerned with the way that an easy route cannot be taken because task demands do not allow flexibility. As indicated earlier, the Levelt model, applied to the first language case, presumes a well-organized and elaborated lexicon. When second language speakers need to access more difficult lemmas, or when they

Table 4: The Levelt model linked to influences on second language performance

Complexifying/Pressuring		Easing/Focusing
	Conceptualizer	
• Planning: extending • More complex cognitive operations • Abstract, dynamic information • Greater quantity of information		• Concrete, static information • Less information • Less complex cognitive operations
	Formulator: Lemma Retrieval	
• Need for less frequent lexis • Non-negotiability of task		• Planning: organizing ideas • Dialogic
	Formulator: Syntactic Encoding	
• Time pressure • Heavy input presence • Monologic		• Planning: rehearsing • Structured tasks • Dialogic • Post-task condition

cannot be flexible to ease their processing problems, ongoing Formulation is disrupted, and it becomes very difficult to sustain parallel processing (Kormos 2006, Skehan 2009b).

A different set of pressuring influences are associated with the syntactic stage of Formulation (although in fact they would also have an impact on lemma retrieval, or at least, the time available for lemma retrieval and information extraction from the lemma). They concern the online pressures that the speaker has to deal with, and include particularly the time pressure under which speaking has to take place, as well as the amount of input that is received by the speaker preparatory to (or even during) speaking. The former, which connects with Ellis' (2005) concept of online planning, is an index of the amount of time the speaker has to access material, to build syntactic frames, and to regroup as necessary. Influencing this amount of time is the need to deal with input, if relevant. Monologic tasks do not present any different or additional difficulties—they simply represent the task type which combines the different pressuring conditions, since they are likely to contain significant quantities of input, especially if a narrative is involved, and especially a video-based narrative retelling, and they are also likely to have to be done under pressuring conditions, since there may simply be a time limit or a quantity of language which has to be achieved, or there may be the need to keep up with a video. It is interesting, in passing, that this analysis of pressuring conditions would be associated with Robinson's (2007) 'here-and-now' condition, which he claims to be the less complex. It may be that while there is a difference between 'there-and-then' and 'here-and-now' regarding complexity, since the 'there-and-then' condition does require memory, in contrast, it is less input-dominated and more negotiable as a result. In other words, there is something of a genre difference between the two conditions which significantly complicates their comparison. Conceptualizer operations may be more difficult with the 'there-and-then' condition, but Formulator conditions may be eased.

We turn next to the right-hand side of Table 4. Here we have influences which either ease the task or alternatively focus attention in a particular area. Regarding Conceptualization, the issues mainly affect the information contained in the task, particularly the quantity and the nature of the information. Concrete, static information is seen as being easier to deal with, and retrieval less demanding than manipulation and transformation. So tasks using such information types, and requiring such information manipulation processes are seen as less demanding for Conceptualizer operations. Essentially these are the reverse of some of the features shown on the left-hand side which complexify Conceptualizer operations.

There are also beneficial, easing influences on Formulator stages, although since many of these function by making more attention available, it is more difficult to separate the lemma retrieval and syntactic building stages. We saw earlier that planning can function to complexify ideas. It can also work, rather differently, to organize ideas. In this case, pre-task work can help to identify

ideas and their inter-relationships and thereby prime lexical elements so that they can be retrieved more effectively when the Formulator stage arrives. The pre-verbal message, in other words, has made certain lemmas more salient, and then their access is smoothed. We have also seen that dialogic tasks have a beneficial influence on performance. Here it is assumed that interaction confers two benefits on lemma retrieval. First, more globally, the fact that interaction is occurring means that while one's interlocutor has the floor, one has more time to regroup, to replan, and to prepare the ground for the message one will utter very soon. In other words, the process of lemma retrieval is given more time to work. But secondly, there is the issue that one's interlocutor can provide useful scaffolding and priming opportunities so that one's own task can be eased because one's interlocutor has done some of the hard work (Skehan 2009b).

There are comparable favourable influences on the syntactic stage of the Formulator. First, again, we have planning, where on this occasion pre-task planning time is directed at syntactic operations. Learners may prepare syntactic frames, sentence fragments, or even complete sentences, that is, they may rehearse (Ortega 2005). These may or may not be retrieved at relevant points during performance, but they are generally helpful preparation for what happens during speech and help the speaker avoid error. There is also the benefit for syntactic encoding of having more time available, and here the evidence on structured tasks is relevant. If speakers do not have to plan so actively for the wider macrostructure content of what they are saying, they can focus more on the detail and have attention available for syntax building (Bygate 2001).

There is another possible influence with interactive tasks. It may be that the presence of an interlocutor makes more salient the need to be precise and to avoid error. In other words, the immediate presence and reality of the person who is being spoken to, coupled with tasks which require precise information, may lead to error avoidance.[4] This is similar to the effect of a post-task activity, but for completely different reasons. In this second case, it is awareness of one's own performance, and its greater salience in terms of formal structure, which leads to an attentional prioritizing that pays more attention to error. But with both of these influences, we have a Formulator (and possibly monitoring) activity that is concerned with the surface of the language which is produced, and which leads the speaker to try to avoid making syntactic mistakes.

The discussion has been around complexification, pressuring, easing, and focusing as influences on the different stages of the Levelt model. These four categories of influence have been instantiated through existing task research findings. But it is also important that the Levelt model connects with the performance areas that are the focus of this special issue. The proposals made here are that complexification links mainly to the Conceptualization stage and then structural and lexical complexity. Pressuring, easing, and focusing are more relevant for the Formulator, and then, accuracy and

fluency. It is to be hoped that future research can extend the categories of influence by identifying additional salient variables that extend this picture, and fine-tune the functioning of the influences covered here.

CONCLUSIONS

There are three broad conclusions to draw from the material presented in this article—that complexity, accuracy, and fluency are important dimensions of second language performance but that they need to be supplemented by measures of lexical performance; that trade-off accounts of findings can account for the results that have been obtained even where accuracy and complexity are simultaneously raised; and that the Levelt model is a useful framework for theorizing this performance.

Regarding the first, it is clear that CAF provide an important touchstone for considering and comparing different models of second language performance. These dimensions are revealing as to the ways task characteristics and task conditions impact upon performance, and seem to capture how second language speakers adapt to different speaking demands when faced with different communication problems. But it is clear that one also needs to consider lexical performance, and so the range of measures needs to be widened to cover this additional area. Of course, one can debate whether it is better to consider lexis as a separate area, or whether it is sufficient to include it within complexity, so that structural complexity and lexical complexity would be considered to be different aspects of this same performance area. The contrast in correlation between lexical sophistication and structural complexity for non-native and native speakers is intriguing in this regard. For native speakers, indeed, complexity may be more unidimensional in that lexical complexity and structural complexity go hand in hand, but for non-native speakers, the two areas do not seem to be integrated so well. Research at other proficiency levels between the ubiquitously studied low intermediates and native speakers is vital in this regard.

But the research reviewed makes it clear that measures of CAF bear upon hypotheses regarding second language performance, and how we can understand the Trade-off and Cognition Hypotheses. Evidence has been reviewed, which shows clearly that trade-off accounts are only part of the picture, and that in some ways, understanding how attentional limitations constrain second language performance is only the starting point. We have seen arguments that the Cognition Hypothesis is not automatically needed to account for cases where complexity and accuracy come together. Alternative accounts, totally compatible with attentional limitations being overcome through judicious task manipulation and task conditions, are satisfactory in accounting for the results reported. Identifying more contexts and conditions that generate raised performance in structural complexity, lexis, accuracy, and fluency, and accounting for such simultaneous influences will

provide a more stringent test of the claim that attentional limitations are fundamental.

That brings us to the utility of the Levelt model. There are several advantages to using this model in relation to second language performance. Even so, and somewhat paradoxically, it should be made clear that it may also be limited in its direct appropriateness. The model portrays a modular, parallel process for speech production. What we know is that second language speakers, depending on conditions and proficiency level, often operate in this way with difficulty. A major reason for this difficulty is the difference between first and second language lexicons, in terms of size, elaborateness, and organization. These differences cause difficulties for second language speakers regarding modular parallel processing (Kormos 2006), leading them to 'shift down' to more serial processes on occasions. Despite this, however, the model is still extremely useful. This applies because it can be linked to what happens as second language proficiency increases (even though there is a dearth of research examining what happens as proficiency grows in relation to the performance areas of complexity, accuracy, lexis, and fluency). We can use the model, and comparisons of native and non-native speaker performance, to give us a handle on how second language speakers change as their proficiency grows, and the ways in which they come to approximate first language speakers. But there is another point of relevance to the model. It separates speech production into more conceptual areas and into more linguistic areas. When studying the second language speaker, it would seem useful to retain this distinction, and then to explore how it is relevant for a differently equipped second language speaker in terms of mental lexicon use. Essentially the framework allows us to distinguish between factors that address the complexity of tasks, since these relate more to the Conceptualizer stage, and factors which affect the way expressions are actually built, since these will impact more on the Formulator stage. In other words, one can propose that certain influences, for example type of information, operations upon information, connect with the pre-verbal message which is developed, while other factors, for example task structure, presence of a post-task activity, are concerned with the amount of attention available, and its focus, when a message is actually expressed. In this way, one can relate limitations in working memory capacity to how the two stages may or may not function smoothly and effortlessly. This may also be relevant to establishing task difficulty for assessment purposes (Skehan 2009b). One needs to distinguish the complexity of the pre-verbal message from the difficulties that arise from the expression of that message subsequently. It is vital in this regard to have measures of complexity, lexis, accuracy, and fluency, because they capture the different facets of performance that have to be rated. There is no absolute standard of task difficulty, because task difficulty is going to be affected differently by these two major stages in speech production.

ACKNOWLEDGEMENTS

The author would like to thank the editors of this special issue of *Applied Linguistics* as well as three anonymous reviewers for comments on an earlier version which have strengthened the contribution considerably. The author would also like to thank the Research Grants Commission, Hong Kong SAR, for the support, through Grant No. 450307, that made the preparation of this article possible.

NOTES

1 An anonymous reviewer pointed out, entirely fairly, that frequency is a useful but dubious criterion for difficulty. In defence of this approach, one can only point to the workability of this criterion as a surrogate measure in lieu of an acceptable, more valid alternative.

2 An anonymous reviewer is unconvinced by the strength of this claim, and suggests that the connection between task design features, Conceptualizer operations, and subsequent language complexity may be more complicated than is proposed here. Further research on this entirely valid point would be interesting.

3 Or, as an anonymous reviewer points out, of the criteria for performance.

4 An anonymous reviewer pointed out, though, that this must depend a lot on the interlocutor. It can also be argued that with tasks where precision of expression and outcome are not so prominent, interlocutors can rely on one another to solve problems, not only of meaning, but also of form.

REFERENCES

Baddeley, A. 2007. *Working Memory, Thought, and Action*. Oxford University Press.

Bell, H. 2003. 'Using frequency lists to assess L2 texts,' Unpublished PhD thesis, University of Swansea.

Bygate, M. 2001. 'Effects of task repetition on the structure and control of oral language' in M. Bygate, P. Skehan, and M. Swain (eds): *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Longman.

Crookes, G. 1989. 'Planning and interlanguage variation,' *Studies in Second Language Acquisition* 11: 367–83.

Daller, H., R. van Hout, and **J. Treffers-Daller.** 2003. 'Lexical richness in the spontaneous speech of bilinguals,' *Applied Linguistics* 24/2: 197–222.

Davies, A. 2003. *The Native Speaker: Myth and Reality*. 2nd edn. Multilingual Matters.

Ellis, R. (ed.). 2005. *Planning and Task Performance in a Second Language*. John Benjamins.

Ellis, R. 2009 (this issue). 'The Differential Effects of Three Types of Task Planning on the Fluency, Complexity and Accuracy in L2 Oral Production,' *Applied Linguistics* 30/4, doi:10.1093/applin/amp042.

Foster, P. and **P. Skehan.** 1996. 'The influence of planning on performance in task-based learning,' *Studies in Second Language Acquisition* 18/3: 299–324.

Foster, P. and **P. Skehan.** 1999. 'The effect of source of planning and focus on planning on task-based performance,' *Language Teaching Research* 3/3: 185–215.

Foster, P. and **P. Skehan.** Submitted. 'The effects of post-task activities on the accuracy and complexity of language during task performance.'

Foster, P. and **P. Tavakoli.** 2009. 'Native speakers and task performance: Comparing effects on complexity, fluency and lexical diversity,' *Language Learning* 59/4.

Foster, P., A. Tonkyn, and **G. Wigglesworth.** 2000. 'Measuring spoken language,' *Applied Linguistics* 21/3: 354–75.

Gilabert, R. 2007a. 'The simultaneous manipulation of task complexity along planning time and +/- here-and-now: Effects on L2 oral production' in M. P. Garcia Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Gilabert, R. 2007b. 'Effects of manipulating task complexity on self-repairs during L2 oral production,' *International Review of Applied Linguistic* 45: 215–40.

Hoey, M. 1983. *On the Surface of Discourse*. George Allen and Unwin.

Ishikawa, T. 2007. 'The effect of manipulating task complexity along the +/- here-and-now dimension on L2 written narrative discourse' in M. P. Garcia Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Kormos, J. 2006. *Speech Production and Second Language Acquisition*. Lawrence Erlbaum.

Kuiken, F. and **I. Vedder.** 2007a. 'Cognitive task complexity and linguistic performance in French L2 writing' in M. P. Garcia Mayo (ed.): *Investigating Tasks in Formal Language Learning*. Multilingual Matters.

Kuiken, F. and **I. Vedder.** 2007b. 'Task complexity and measures of linguistic performance in L2 writing,' *International Review of Applied Linguistics* 45: 261–84.

Kuiken, F. and **I. Vedder.** 2008. 'Task complexity and linguistic performance in L2 writing and speaking: The effect of mode.' Paper presented at the AAAL Conference, Washington, DC.

Laufer, B. and **P. Nation.** 1999. 'A vocabulary-size test of controlled productive ability,' *Language Testing* 16: 33–51.

Levelt, W. J. 1989. *Speaking: From Intention to Articulation*. MIT Press.

Levelt, W. 1999. 'Language production: A blueprint for the speaker' in C. Brown and P. Hagoort (eds): *Neurocognition of Language*. Oxford University Press.

MacWhinney, B. 2000. *The CHILDES Project: Tools for Analysing Talk: Volume 1: Transcription Format and Programs*. 3rd edn. Lawrence Erlbaum.

Malvern, D. and **B. Richards.** 2002. 'Investigating accommodation in language proficiency interviews using a new measure of lexical diversity,' *Language Testing* 19/1: 85–104.

Meara, P. and **H. Bell.** 2001. 'P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts,' *Prospect* 16/3: 5–19.

Michel, M. C., F. Kuiken, and **I. Vedder.** 2007. 'The influence of complexity in monologic versus dialogic tasks in Dutch L2,' *International Review of Applied Linguistics* 45: 241–59.

Ortega, L. 1999. 'Planning and focus on form in L2 oral performance,' *Studies in Second Language Acquisition* 21: 109–48.

Ortega, L. 2005. 'What do learners plan? Learner-driven attention to form during pre-task planning' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Read, J. 2000. *Assessing Vocabulary*. Cambridge University Press.

Richards, B. and **D. Malvern.** 2007. 'Validity and threats to the validity of vocabulary measurement' in H. Daller, J. Milton, and J. Treffers-Daller (eds): *Modelling and Assessing Vocabulary Knowledge*. Cambridge University Press.

Robinson, P. 2001. '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*. Cambridge University Press.

Robinson, P. 2007. 'Re-thinking-for-speaking and L2 task demands: The Cognition Hypothesis, task classification, and sequencing.' Plenary address at the Second International Conference on Task-based Language Teaching, Hawaii.

Robinson, P., T. Cadierno, and **Y. Shirai.** 2009 (this issue). 'Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production,' *Applied Linguistics* 30/4, doi:10.1093/applin/amp046.

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

Skehan, P. 2001. 'Tasks and language performance' in M. Bygate, P. Skehan, and M. Swain (eds): *Research Pedagogic Tasks: Second Language Learning, Teaching, and Testing*. Longman.

Skehan, P. 2009a. 'Lexical performance by native and non-native speakers on language learning tasks' in B. Richards, H. M. Daller, D. Malvern, P. Meara, J. Milton, and J. Treffers-Daller (eds): *Vocabulary Studies in First and Second Language Acquisition: The Interface between Theory and Application*. Palgrave Macmillan.

Skehan, P. 2009b. 'Models of speaking and the assessment of second language proficiency' in A. Benati (ed.): *Issues in Second Language Proficiency*. Continuum.

Skehan, P. and **P. Foster.** 1997. 'The influence of planning and post-task activities on accuracy and complexity in task based learning,' *Language Teaching Research* 1/3: 185–211.

Skehan, P. and **P. Foster.** 1999. 'The influence of task structure and processing conditions on narrative retellings,' *Language Learning* 49/1: 93–120.

Skehan, P. and **P. Foster.** 2005. 'Strategic and on-line planning: The influence of surprise information and task time on second language performance' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Skehan, P. and **P. Foster.** 2008. 'Complexity, accuracy, fluency and lexis in task-based performance: a meta-analysis of the Ealing research' in S. Van Daele, A. Housen, F. Kuiken, M. Pierrard, and I. Vedder (eds): *Complexity, Accuracy, and Fluency in Second Language Use, Learning, and Teaching*. University of Brussels Press.

Tavakoli, P. and **P. Foster.** 2008. 'Task design and second language performance: The effect of narrative type on learner output,' *Language Learning* 58/2: 439–73.

Tavakoli, P. and **P. Skehan.** 2005. 'Planning, task structure, and performance testing' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. John Benjamins.

Towell, R. and **J.-M. Dewaele.** 2005. 'The role of psycholinguistic factors in the development of fluency amongst advanced learners of French' in J.-M. Dewaele (ed.): *Focus on French as a Foreign Language*. Multilingual Matters.

Towell, R., R. Hawkins, and **N. Bazergui.** 1996. 'The development of fluency in advanced learners of French,' *Applied Linguistics* 17/1: 84–115.