Language Teaching (2024), **57**, 501–516
doi:10.1017/S0261444823000344

CAMBRIDGE UNIVERSITY PRESS

PLENARY SPEECH

Linking second language speaking task performance and language testing

Peter Skehan

University College London, London, UK
Email: peterskehan@gmail.com

(Received 10 July 2023; accepted 22 July 2023)

Abstract

This written version of a plenary to the Language Testing Research Colloquium relates research into second language speaking task performance to language testing. A brief review of models of communicative competence and of speaking is provided. Then, two major areas within testing are discussed: the concepts of difficulty and ability for use. The next section covers research into spoken task-based performance, covering effects from task conditions, and from task characteristics. In addition, the measurement of such performance is described and briefly compared with performance rating in testing. Then, the final section relates the task research findings to language testing. A framework for testing spoken performance is outlined, and the general claim made that effective sampling through tests, in order to generalise to real-world performance, can usefully draw on findings from second language task research, as well as the distinction between Conceptualiser and Formulator processes.

1. Introduction

It was a considerable surprise to me to receive the Messick award. The context for this award is language testing, and while there have been times when I have published in this area, they are (mostly) well in the past, and so I was uncertain what I could cover in the associated Messick Memorial Lecture (which is given by the awardee). As a solution to this problem, I decided to draw upon more recent work I have done, focussing on second language (L2) spoken task performance, and to relate this body of work to the testing of speaking. My justification for this is that speaking figures prominently in the general area of testing, but that the research I would draw on, task-based performance with a psycholinguistic approach, while hardly unknown to testers, is less prominent than other approaches to devising and calibrating tests. My assumption was that there is potential gain in making links between these different areas, not least because task researchers are not so test-format driven. They also have different theories for conceptualising tasks, highlight different influential variables, and use different methodologies for measuring task performance.

In this written version of the plenary, there are four sections. The first explores some major models of both communicative competence and speaking. Second, I discuss two general concepts: test-task difficulty, a central puzzle in language testing, and ability for use, the capacity to produce actual language, not simply knowledge about language. The third section tries to cover relevant research from the task literature, on task characteristics, on the conditions under which tasks are done, and finally how task performance is measured. Then, the final section tries to relate the task findings to the constructs of difficulty, and ability for use, and also to the field of language testing more generally.

Plenary to the 43rd Language Testing Research Colloquium, Messick Memorial Lecture, 11 March 2022, Tokyo (virtual conference).

2. Models of communicative competence and models of speaking

An early and very influential account of communicative competence (Canale, 1983; Canale & Swain, 1980) proposed that it subsumes linguistic competence, sociolinguistic competence, discourse competence, and strategic competence. The first three of these are underlying in nature, while strategic competence is drawn into actual performance when problems arise. Perhaps the next major development was Bachman's (1990) work, developed further in Bachman and Palmer (2010). Canale and Swain's underlying competences are repackaged slightly. An overall Language Knowledge subsumes Organisational Competence (itself subsuming grammatical competence and textual competence) and Pragmatic Competence (subsuming sociolinguistic competence and functional competence). A major change, though, is that Strategic Competence moves into a more central and independent role – not only implicated when problems occur, but also central to all communication – concerned to assess a situation, decide on what needs to be said, and then marshal resources for achieving the aims that are set – in other words, a concern for ability for use. More recent proposals have also had a role for the dynamic aspects of communication (Xi et al., 2021). Weir (2005), for example, in his Socio-Cognitive model discusses, within the cognitive component, the role of levels of processing and information sources. Purpura (2016), with a particular emphasis on how different types of meanings are conveyed, emphasises the role of knowledge sources.

More recently Hulstijn (2015) has proposed a somewhat different model, more psycholinguistic in nature, less explicitly testing-focussed, and embracing both second and first language (L1) proficiency. It is based on two distinctions. The first contrasts basic and higher (or extended) language cognition. Basic cognition, especially in the L1, concerns the automatic universal language system, possessed by all native speakers, while Higher Cognition is concerned with lower frequency language, often with literary connections. The second distinction, and the one that is more relevant to the L2 case, is between core and periphery. The core involves linguistic knowledge AND speed of access. The periphery includes interactional ability, strategic competence (especially for difficult conditions and also limited knowledge), metalinguistic knowledge, and knowledge of various types of oral and written discourse. Hulstijn, while trying to present a more unified version of L1 and L2 proficiency, introduces a new concern – speed of system operation – and then proposes a different view of how underlying knowledge is mobilised for actual language use.

So far, I have explored broader issues of communicative competence and proficiency. Next, I turn to speaking, specifically, the focus of this plenary, and the point of connection with research into task-based performance. There are several psycholinguistic models of L1 speaking. Arguably the one that has had the greatest impact on L2 speaking research is Levelt (1989) (and see De Bot (1992) and Kormos (2006) for extensions of the Levelt model to bilingual and L2 speakers). The model proposes three broad stages in speech production: Conceptualisation, emphasising ideas; Formulation, which takes the Conceptualiser pre-verbal message, and then clothes that message in language elements; then the final stage, Articulation, which builds on the phonetic plans generated by the Formulator to produce actual speech. The Conceptualiser draws upon important knowledge sources (situational knowledge, discourse models, encyclopaedic knowledge, and so on). The Formulator draws on the mental lexicon to handle pre-verbal message demands through processes driven by rich lemma retrieval. Importantly, in the L1 case, these demands are met, in real-time, below the level of consciousness, and enable parallel, modular processing – that is, each speaking stage is getting on with its job, its bit of a message, simultaneously.

In the L2 case, it is assumed the Conceptualiser functions perfectly well, most of the time, assuming no external problems. The major difficulties are likely to come at the Formulator stage because the SECOND language mental lexicon is likely to be less extensive, less elaborate, less organised, less connected and slower than the L1 mental lexicon. Two important implications follow from this. First, it is likely, when pre-verbal message demands are not met, or met with difficulty, and the speaker does not abandon the message, that the parallel approach to L1 speaking will become more serial in nature, as the Formulator uses up most of the attention that is available. Second, it is likely that

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

a declarative (rule-based) knowledge system, which probably derives from the instruction the speaker has received, will be needed and a more explicit mode of using language, based on rules, will be involved (De Bot, 1992; Kormos, 2006; Skehan, 2018). (Of course, as proficiency increases, these difficulties diminish, even if they do not disappear entirely.)

There are similarities and differences across these different models, and these offer potential connections between psycholinguistic views of speaking and the nature of communicative competence and L2 proficiency. Canale and Swain (1980), Bachman (1990), and Hulstijn (2015) all propose underlying knowledge sources, and then processes to draw upon such knowledge in actual communication. They vary in their account of the underlying competences, and even more in their proposals as to how competences are used. The implications for testing, though, are clear. If one wants to develop a test directed at speaking proficiency, then the models provide a framework for the underlying competences to be sampled, and a framework for the way competences are activated. The more extensive and systematic such sampling can be, presumably the more effective the test, with higher predictive validity. Models such as these have been hugely influential in test construction over the last 40 years or so.

When we turn to the Levelt model, a L1 model of speaking, there are interesting similarities, but also some important contrasts. A major similarity concerns Bachman's view of strategic competence and Levelt's views on Conceptualiser operations. Both are concerned with how speakers engage in idea generation, selection, and organisation, and also there are connections with the way a Leveltian preverbal message is translated into language. There are also links between Hulstijn's Linguistic Knowledge and Levelt's mental lexicon. Hulstijn also is explicit about the importance of speed of operation, which connects with how the Formulator can work in real-time, and support parallel processes.

It is the differences between Levelt and the L2 focussed models that are most significant here, since they have implications for how testing of speaking can be done, and how research into task-based performance may be relevant. The Levelt approach highlights:

- An empirically-grounded model of speaking.
- The normality of its stages working in harmony, in that the demands they place on each other are 'reasonable', and so parallel processing is feasible.
- An account of underlying knowledge sources, contrasting knowledge of the world (including sociolinguistic and pragmatic knowledge), and implicit knowledge of language.
- A clearer account of the lemma-based nature of communication and the role of the mental lexicon.

This analysis clarifies that it is misleading to over-emphasise competences, of different stripes, and that real-time communication is underpinned by the nature of mental lexicon (occupying slightly the place that strategic competence takes in testing models).

3. Implications for language testing

This analysis creates the foundation to consider two major insights, and associated implications. The first is based on the importance of Levelt's three major stages in speaking. I assume, following De Bot (1992) and Kormos (2006), that these same stages are also central in L2 speaking, even if there are operational differences. This leads to the second insight – that a major first–second language difference concerns the way the three stages interact with three major knowledge sources, two of which are language related (the L2 mental lexicon and the declarative L2 knowledge system). The implication, based on the first insight, is that when we consider test (task) difficulty, we need to separate out difficulty influences at the Conceptualiser stage from those relevant to the Formulator stage. The second insight gives the implication that we need to explore more deeply the concept of Ability-for-Use as a way to better understand any relationship between underlying competences and actual performance. We will consider each of these implications at a conceptual level next, and this will then pave the way for a section on how task-based research can make empirical contributions to this analysis.

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

3.1 Conceptualiser and Formulator difficulty are not the same

A little context is necessary. Language testing adores being able to talk about difficulty. Indeed, one of the major developments in the field in recent decades has been the routine use of Item Response Models to explore difficulty, with formidable statistical techniques, and gigantic datasets. But to my mind, these views of difficulty, mostly, though by no means exclusively, item-based, are a little light on theory of language and language use. The concern is that technical qualities may be prioritised over construct coverage. Such views also seem to me to make assumptions about the dimensionality of language performance, so that a minimal number of dimensions are often assumed. A contrasting approach, Levelt-based, seems to me refreshing as regards the concept of difficulty. As we have already seen, the sub-processes within Conceptualisation, focussed as they are on ideas, and context, and stance, are not primarily linguistic in nature. Their product, the pre-verbal message, is proposition-based, not linguistic-based. In contrast, Formulator operations are thoroughly linguistic. Drawing on the (L2) mental lexicon, lemmas are retrieved and 'plugged in' to the process of language construction. The information within the lemma is, potentially, very rich, and enables syntax-building, smooth phonological retrieval, discourse connections and much more. All this needs to be done at speed, with minimal attentional demands, so that other parts of the system are not derailed (and serial processing required).

The 'punchline' from this analysis should, by now, be obvious. What makes for difficulty in the first set of operations, within Conceptualisation, is quite distinct from the sorts of things that influence difficulty in the second, Formulation. In the first case, as we have seen, we are dealing with the ideas that underlie what is going to be said, and so difficulty is, broadly, going to connect with how ideas are handled. This means selecting relevant ideas, perhaps relating them to context to say things that are appropriate. It may mean developing a stance and even point of view. Drawing on previous encounters will be important and it is indeed likely there will be a tension between simply remembering as a basis for what is said, compared with needing to think actively, and even transform material. There is also the issue of scale, of how much can be addressed at the Conceptualisation stage to enable quite a bit of discourse, quite a few pre-verbal messages, to be produced at one time. These could provide insurance, as it were, for problems 'downstream', since they would give a general framework for the details of later communication. So, difficulty at this stage will be concerned with the nature of such demands and how people handle them. A small point here is that this analysis is relevant to monologic AND interactive language contexts. Although, it has to be admitted, the majority of task research (though not all) is based on monologic tasks, the issues raised are relevant to how Conceptualiser functions in both contexts. In some ways, it may be that Conceptualiser operations are slightly eased in interactive contexts, because the speaker-to-come may be able to listen to an interlocutor AND plan future utterances simultaneously.

Difficulty with Formulation may be very different. The focus is on execution, and much of this difficulty will derive from the capacity of the L2 mental lexicon to handle the demands driven by the pre-verbal message. The more difficult the pre-verbal message, the more difficult the lexical retrieval and associated Formulator processes that are needed. As we have seen, in the L1 case, 'routine' functioning is the norm. In the L2 case, difficulty is importantly influenced by the size of the lexicon, by its organisation, by the completeness of the information stored with any particular lemma, and most of all, by the speed needed for operation. More frequent language, one assumes, will make for greater ease, while less frequent and indeed more challenging language will make for difficulty. In addition, the conditions under which speaking is carried out become important. Communicative (time) pressure, including that from an interlocutor, may have an important role. So might familiarity of information, occasions when similar things have been done in the past, and also opportunity to plan. Such sources of difficulty are generally different to those which make for difficulty with Conceptualisation. As a result, we have to consider that, for any balanced assessment of communicative competence, sampling will need to cover both of these different sources of difficulty, and to do so systematically if there is to be a broad basis for generalisation.

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

The distinction can be illustrated through a matrix, given as Figure 1, contrasting two difficulty levels for Conceptualisation and two for Formulation. (This 2 × 2 arrangement is, obviously, for illustration only – in reality we are dealing with a cline.) Each significant cell in the matrix is identified by a capital letter. In each cell, Formulator influences are shown in normal font, and Conceptualiser influences in italics. The matrix functions as a sampling frame for potential test items or sub-sections, since more systematic sampling of the 'space' so defined would provide a more robust basis for generalisation to real-world performance.

For illustration, various tasks, both monologic and dialogic, can be placed within this framework. Cell A could contain the 'compare family trees' task (Willis & Willis, 1988), or describing a journey home from school (Foster & Skehan, 1996), or telling the story in a structured narrative (Wang & Skehan, 2014). Cell B could be illustrated by a narrative where different elements – for example, background and foreground information – have to be related to one another (Tavakoli & Skehan, 2005), or picking the right hike, given interesting stimulus information (Norris et al., 1998). Cell C might involve a narrative with unavoidably difficult lexis (Wang & Skehan, 2014), or the task of ordering coffee and dessert (Norris et al., 1998). Cell D would be exemplified by the Fire Chief task, rescuing people from a burning building, where complex criteria might be involved as well as pressured conditions (Gilabert et al., 2009).

While the organisation of the matrix derives from Levelt's model, the ideas it embodies are not exactly new to the testing community. Something of this sort was influential in the work of the Hawaiian group (Norris et al., 1998) in devising a large range of potential tasks that could be drawn on in an academic testing context. Similarly, Luo (2008), with secondary school children in China, and within the framework of a National Curriculum, used a broadly similar approach. A system was devised to generate test tasks appropriate to this age-level and context, and this was used to establish level of difficulty (and see Skehan & Luo, 2020).

3.2 The concept of ability for use is vital in understanding language use in tests

As indicated earlier, in the L2 case the mental lexicon is not so extensive, rich, or fast, so that gaps and slower operation mean that problems occur. As a result, guaranteed access to implicit knowledge is not available, and so a L2 learner's declarative knowledge system has to be used. This is slower, and attention-demanding, and problems with it may disrupt general communicative effectiveness. As a result, with L2s, the capacity to mobilise different knowledge resources, and integrate them within the stages of speaking becomes very important. Hence the need for a construct such as Ability for Use (Hymes, 1972). At the outset then, a quick overview of my proposals for Ability for Use provide a structure for what is to come, clarifying the underlying knowledge sources and then the different components. This is shown in Table 1.

In the rest of this section I will try to address each of these areas in more detail, although the section will not elaborate on knowledge sources, as these have been covered in the earlier section. The first

	Conceptualiser Easy	Conceptualiser Hard
Formulator Easy	Unpressured communication *Familiar, structured information only requiring recall.* **A**	Unpressured communication *Extending, planning, reasoning and transformation processes.* **B**
Formulator Hard	Pressured communication, heavy input, monologic, non-negotiable *Familiar structured information, emphasis on retrieval.* **C**	Pressured communication, heavy input, monologic, non-negotiable *Extending, planning, reasoning and transformation processes.* **D**

Figure 1. Conceptualiser and Formulator ease and difficulty

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

Table 1. Knowledge sources and ability for use

Knowledge sources
• General knowledge base, plus context and audience sensitivity • Second language mental lexicon: size, richness, organisation, speed • Declarative knowledge of the L2
Ability for use
• Conceptualiser and Formulator processes • Working memory • Metacognition o Management of Conceptualiser and Formulator operations o Awareness of knowledge sources, attentional demands, limitations o Foresight, trouble identification and avoidance o Synchronisation of resources • Compensation ability o Monitoring o Compensation o Resourcefulness

strand concerns Conceptualiser and Formulator processes themselves, and the central claim is that L2 speakers may vary in how effectively they handle such processes. At the Conceptualiser stage, it is important to retrieve, marshal, organise and manipulate ideas, to evaluate situations, including the contribution of other participants, and to decide what needs to be said. People vary in how effectively they might do this, with some faster than others, and more able to draw upon greater range of previous experience. Scale of Conceptualiser operations is also important. The 'classic' output is a pre-verbal message, but a more macroplanning approach at this stage might generate a set of inter-linked pre-verbal messages, easing subsequent Formulator work, and even protecting the Formulator as pre-verbal messages could be returned to more easily. Similarly, the capacity to mobilise and use memorised, ready-made ideas can ease both Conceptualiser and Formulator operations. Turning to the Formulator stage, speed of operation (and also perhaps the capacity to draw upon wider repertoires of formulaic language) confer considerable advantages. (Other aspects of Formulator operations will be dealt with below.) So, a capacity to communicate is partly dependent on the effectiveness with which the different stages of speaking are handled.

A second, perhaps relatively minor, aspect of Ability for Use is working memory, since there is a major role for 'buffers' to hold material during processing (Skehan, 2022). For speaking, we assume the existence of an assembly buffer that receives input from different knowledge sources within the stages of speaking, and that then outputs the actual message (and is then cleared). An implication of this is that the larger, faster, and more efficient working memory is, the greater the contribution to Ability for Use. Underlying knowledge sources can be accessed faster and more comprehensively, and operations to underpin Levelt's three stages can be more effective. So, it may be the case that those with better working memories are more effective communicators. There is a word of warning, though. The 'narrow window hypothesis' (Skehan, 2022) raises the possibility that, for at least some communication, the range of variation in working memory may not always have FUNCTIONAL significance – there may be differences, but given the speed and pressures of on-line communication, these differences may not impact upon performance. As we will see below, there is research that bears on this issue.

I turn next to the remaining components of Ability for Use, and I think I would argue that they are the most significant for L2 speakers. Metacognition is treated slightly separately here whereas a case could certainly be made to discuss it within Conceptualiser and Formulator operations. The motivation to look at it separately connects with the limitations of the second language mental lexicon, and, as a result, the need to integrate, where appropriate, a declarative knowledge store. Central to this is speaker insight into the speaking process, and how it can be managed, eased, and even improved. This depends upon awareness of knowledge sources and of attentional demands and

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

limitations. If one knows that modifications may need to be made because of such limitations, it could well be the case that greater anticipation of likely problems will push the speaker to modify the plan they are following. In addition, there may be problems in synchronising different resources, so that, for example, awareness of the advantage of developing a 'set' of pre-verbal messages may ease speaking processes.

Very often, though, despite anticipation and avoidance, problems will occur in L2 speech, and another aspect of Ability for Use then assumes importance – compensatory/recovery ability. Compensation focusses on dealing with a problem (syntactic, lexical, discoursal, sociolinguistic) when it occurs. This may, in turn, be related to effective monitoring, as problems are detected quickly, and thereby resolved more easily so that some degree of flow is maintained. Recovery is related to this but has the major difference that a problem may derail communication and force some degree of regrouping, thus presenting a challenge to ongoing flow. A first problem here is that repair is needed, but a second problem is that the thread of discourse needs to be rejoined if possible, and so an additional part of ability for use is to retain where one is, in speaking, and to be able, with repair available, to go back to that point, or to a new relaunching point.

Recalling the earlier communicative competence models, it is clear that there is a great deal of repackaging in the present account. Bachman's Strategic Competence, for example, relates to the discussion on Conceptualiser and Formulator processes. Canale and Swain's views on strategic competence relate more to the compensatory aspects of the discussion here. And Hulstijn's discussion of speed is psycholinguistic in nature and links well with the operation of the second language mental lexicon. Hulstijn (2015) also discusses Strategic Competence in very relevant ways. Weir's (2005) proposals on cognitive validity are clearly linked to Levelt's (1989) speaking stages and processes. All approaches, including this one, are wrestling with the knowledge/competence linkage in performance, so the degree of overlap across the different approaches is considerable.

But there are differences in the present account that are important. First, it links more with an adapted Leveltian perspective, with stages of speaking and the central role of the mental lexicon, and also the need to integrate additional knowledge sources such as declarative knowledge. Second, the database underpinning the discussion draws upon the L2 task performance literature. It is to this task-based literature that I now turn.

4. L2 task-based research and communicative competence

The literature on L2 task-based research has grown enormously over recent decades, with extensive theorising and a considerable database now available. It will be argued here that this literature also has considerable relevance to our understanding of communicative competence and to assessment. Three aspects of task-based research will be covered here:

- Task conditions, especially planning.
- Task characteristics.
- The measurement of performance.

Additional factors are also discussed, including metacognition, working memory, Conceptualiser–Formulator balance, and the relevance of proficiency level and speaker style.

4.1 Task conditions

One result of the growth in the task literature is that a range of generalisations are now available, based on multiple rather than single studies (Ellis, 2009; Skehan, 2018). For example, planning consistently raises language complexity and fluency, but less consistently accuracy and lexis. Other forms of readiness also have similar effects on performance, such as modelling, video-based preparation, and specialist knowledge (Bui, 2014). As more time is given, accuracy is first affected, then fluency, and

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

finally complexity (Mehnert, 1998). Planning also seems to have greater effects with more complex tasks and at higher proficiency levels (Bui et al., 2019).

In principle, such results should be relevant for language testing. Indeed, there is an argument (Skehan, 2001) that giving L2 speakers preparation time should have a 'levelling the playing field' effect, in that the speaker can relate the task to themselves and their own interests and opinions more, in a way that is closer to general language use – it is not clear how we can generalise from test tasks about sudden arbitrary (and even unfamiliar) topics to the use of language in more natural situations. But there is the problem that there have been studies investigating the planning variable, within an overt testing context – for example, O'Grady (2019) – which have not replicated the effects of task research. There may be something about the testing situation that changes approaches to performance, perhaps emphasising conservatism and accuracy, and this washes out the effects of planning. Alternatively, different measurement approaches may lead to different approaches to precision (greater detail in tasks, broader rating scale steps in testing). There is clearly scope for more research here to try to pin down why there are sometimes contradictory results from the two domains – task-based research and language testing research, where tasks and/or task conditions are central to the research. In any case, the discrepancy is a little two-edged: if consistent results from one domain, tasks, often in arguably more ecologically valid contexts, do not generalise, one can ask what that is saying about the usefulness (or not) of results from testing contexts as a basis for predicting real-world performance (Norris, 2018).

There is also interesting qualitative research with planning. Francine Pang and I (Pang & Skehan, 2014) carried out a study in which (a) we asked L2 speakers to tell us, retrospectively, what they did during earlier planning time, and (b) we related what they said about their planning activities to quality of performance. We discovered some surprising things. Higher CALF (complexity, accuracy, lexis, fluency) scorers on a narrative task reported that they were more likely, in planning, to emphasise ideas, rather than grammar and specific language; that they tended to plan small and specific rather than large and general; that they were more likely to be realistic about what they could remember and avoided being over-ambitious, tending to assess what they could manage and then not overdo things; that they sometimes tried to build structure into the way they did a task (and see the next section for this); and that they were more likely to think about how trouble might occur, and how they could deal with it.

These results are very interesting and show clearly that not all people use planning opportunities in the same way. The results also make connection with the earlier discussion on Ability for Use. Planning, generally, provides scope for Conceptualiser processes to have material to draw on, in a more organised way. It can also help Formulator operations (though more successful speakers tended to avoid specificity). But the qualitative research brings out that aspects of metacognition are very important: some participants clearly made decisions that connected with higher-level performance, and this seems to link to foresight, to management of complex knowledge sources, to performance and attentional limitations, and memory. Compensation was also relevant. So, we see that the mediating construct of Ability for Use had a clear connection with how well people did when speaking. The research database here may be a planning study, but I argue that this has simply enabled a clearer view of how L2 speakers approach tasks more generally.

There is more to planning than pre-task planning, though. Ellis and Yuan (2005) have researched online planning – that is, the sort of planning that is made possible when speaking occurs under unpressured time conditions. They propose that it is possible, in such circumstances, to handle ongoing Formulation-Articulation while simultaneously planning what will be said next. They report (Ellis & Yuan, 2005) that such planning is associated with greater accuracy. This basic insight has stimulated additional research, and this too is illuminating for language testing. Earlier it was proposed that working memory is an important part of ability for use. Yet, Wen (2009) showed that when one has pre-task planning, there is no correlation between working memory scores and task performance. However, when there is ON-LINE planning, working memory scores DO correlate with performance (Ahmadian, 2012). It appears that the benefits of working memory require more processing time

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

for their effects to become apparent – the combination of the greater working memory AND less time pressure. This chimes with the 'narrow window hypothesis' (Skehan, 2022), mentioned earlier, which suggests that more is needed than simply greater working memory – other supportive conditions need to be operative.

Another study that researched the way on-line planning interacts with other variables is by Wang (2014). She explored correlations between proficiency test scores and narrative task performance under three conditions: unplanned AND time pressured; pre-task planned BUT time pressured; unpressured, that is, on-line planning. She reports that in the first condition, there was no correlation between task performance and proficiency, that in the second, there was a moderate correlation. Importantly, in the unpressured on-line condition there was a strong correlation. In other words, greater proficiency does not seem to have an impact on performance when there is no planning support (pre- or during-); that it does help if there has been some pre-task planning; and that it makes its greatest contribution when there is little time pressure, that is, there is on-line planning. Wang et al. (2019) propose the Proficiency Mobilisation Hypothesis to capture this insight – that is, that the processing conditions need to be right for proficiency to have an impact. Clearly, there are important implications here for testing in that if there are underlying abilities (declarative knowledge), and these are a target for testing, it seems that little time pressure is helpful for such abilities to manifest themselves.

One final planning study is relevant to language testing. Wang (2014) also had a condition where one group of participants had the opportunity for pre-task planning AND ALSO did the actual narrative task under unpressured conditions. This produced the largest effect of all conditions in the study, raising complexity, accuracy, and fluency, and doing this with greater effect sizes than any other separate condition. This is consistent with the Conceptualiser–Formulator Balance principle (Wang et al., 2019): providing speakers with the opportunity to prepare ideas and organisation (Conceptualisation) AND the opportunity to produce language effectively (Formulation). In other words, something to say, and the means to say it.

It is clear, then, that the studies on task conditions, certainly planning, clarify how task performance can be influenced. This, in turn, has implications for the details of how testing takes place. Small changes may have an impact on performance, suggesting that standardisation of these influences may be important, or at least, that careful consideration is required if comparisons are made between different testing contexts. But equally importantly, the findings also clarify how Ability for Use is important in understanding test performance, as well as vital when one is designing a range of tests whose function is to sample behaviour as the basis for generalisation to real world performance.

4.2 Task characteristics

Research into the impact of task characteristics on L2 performance has been central to the task field since its beginning. As far as testing is concerned, there is the fundamental issue that if tasks are not neutral data elicitation devices, comparing test performances when different tasks are involved becomes significantly more difficult. On the basis of current task research, a range of generalisations regarding task characteristic effects are now available. The information type underlying a task is important: tasks based on concrete, familiar information tend to raise accuracy and fluency, whereas those based on unfamiliar and abstract information tend to raise complexity (Skehan & Foster, 2008). Tasks based on a tight structure (Skehan & Foster, 1999; Tavakoli & Foster, 2008) tend to raise accuracy, fluency, and, sometimes, complexity. Operations on the data within a task have also been shown to be important. Transformation of information, or integration of different aspects of information (e.g. need to combine background and foreground information in a narrative (Tavakoli & Skehan, 2005)) have been shown to increase complexity.

There have also been claims deriving from the Cognition Hypothesis (Robinson, 2015), which argues that resource-directing variables (reasoning demands, time perspective, number of elements and so on) raise accuracy AND complexity. The evidence, particularly as reported in meta-analyses

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

(Malicka & Sasayama, 2017) is mixed, certainly to support the claim that these resource-directing factors will JOINTLY raise complexity and accuracy. There is more basis for the claim that one area will be raised, such as reasoning demands raising complexity and sometimes lexis; with time perspective that there-and-then tasks raise complexity but lower fluency (Wang & Skehan, 2014). More broadly, other task features, not particularly theory-linked, tend to have consistent influences. Greater lexical demands can slightly reduce the level of fluency (Wang & Skehan, 2014). More negotiable tasks, where the speaker is not bound by particular input or task requirements, but can choose how to address a task, tend to lead to more complex language. Finally, dialogic tasks, if there is engagement between participants, can raise complexity and fluency, and sometimes accuracy (Skehan & Foster, 1997).

The amount of research activity with task characteristics has been considerable, but does contrast with the research on task conditions. There, we may not have as wide a range of results, but there is greater consistency and greater effect sizes within narrower areas (Skehan, 2016). With task characteristics, there is a greater range of results, but the level of consistency has not been so great, and the effect sizes have been smaller. Norris (personal communication) argues that this is likely connected with inconsistencies in the operationalisation of causal variables across studies, rendering interpretations of results very difficult if not impossible. Of course, it is the nature of tasks that they can be interpreted differently by different participants, and so this unpredictability may contribute to the relative lack of clear and consistent generalisations (Bachman, 2002).

There is obviously massive scope for additional research, and such research will contribute to a clearer picture with task effects. But two observations are worth making in relation to tasks and testing. The first concerns the potential non-neutrality of tasks as data elicitation devices. I would argue that, on the basis of task research, task conditions need to be taken very seriously when one is judging comparability of test task results. Regarding tasks themselves, however, it is possible that the impact of particular tasks being chosen may not be as influential as perhaps has been previously thought (by me, amongst others!). The second point is that while professional test developers may have not conducted much direct task research themselves, (though there are some major contributions, e.g. Galaczi, 2006; Nitta & Nakatsuhara, 2014), they are not unaware of the sorts of influences that have been discussed in this section. Salisbury (2010), for example, explored how expert test developers routinely consider the strengths and limitations of different test task types and of different test conditions. In many ways, the expertise that they have developed, by whatever means, seems to parallel (and anticipate) some of the findings from task research itself.

4.3 Task research and the measurement of performance

Task researchers obsess about complexity, accuracy, fluency, and lexis. Each of these performance areas are measured separately, carefully, and in detail. Testers do not use such detailed measurements, preferring to rely on carefully developed rating scales, used by trained raters, often with a distinction between general ratings, and more precise analytic ratings – for example, range, vocabulary, accuracy, and fluency. So, there is quite a degree of convergence between these two areas (as well as some reservations – Fulcher (2015), for example, questioning the strength of the relationship between isolated CALF measures and L2 language proficiency). O'Grady (2023) also draws attention, within a testing context, to the importance of halo effects, where ratings of one analytic dimension spill over in their influence to the other dimensions. The central point, however, is that the two areas can contribute to one another. Additional areas used in testing, such as task fulfilment and interaction, might be profitably used by task researchers. But worth discussing here is the way the detailed measures used by task researchers might generate insights into what descriptive elements might be included in rating scales. It will then be an empirical issue as to how these insights might contribute to greater effectiveness (or not!) in how judgements are made about speaker competence in actual tests.

Regarding complexity, task researchers typically rely on measures of subordination (such as clauses per AS-unit), or the number of words per clause. Interestingly, these two measures do not produce

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

strong inter-correlations (Inoue, 2016; Skehan, 2018). In view of this, task researchers have explored which of these measures of structural complexity are affected by which independent variables. Pre-task planning and narrative tasks raise both subordination and words-per-clause measures (Skehan, 2018). Structured tasks, and there-and-then time perspective raise subordination only. Interestingly, non-native speakerness (vs native), and lower proficiency (Skehan, 2018) raise words-per-clause! This connects with a suggestion (Pang & Skehan, 2021) that there may be a tension between what is termed a 'discourse' oriented style (higher subordination) and a clause-oriented style (higher words-per-clause), with the former also correlated with greater speed, less pausing and less repair and the latter correlated with lower speed and more pausing and repair. These results might have implications for the descriptors of the steps in a rating scale: there may not be only one dimension of complexity, in this regard.

Two other areas, accuracy and lexis, give slightly contrasting results based on task research. With accuracy, there are various potential measures, each with different emphases (errors per 100 words, error-free clauses, error gravity (Foster & Wigglesworth, 2016), error linked to length of clause). Overall, the particular choice of measure does not seem to matter. All measures correlate fairly highly and so it seems that whatever decision one makes, the overall assessment of accuracy is pretty much the same. The implication is that task measurement research does not have that much to offer language testing accuracy rating scale construction.

The situation is a little different with lexis. Different aspects of lexical use in speaking have been used in task work, with a major contrast between lexical diversity and lexical sophistication. Lexical diversity captures the extent to which speakers tend to recycle the same words during performance, or not. Lexical sophistication uses an external criterion, usually word frequency, to establish how many 'sophisticated' (usually low frequency) words are used. Interestingly, these two measures do not correlate particularly highly and seem to reflect different processes (Skehan, 2009, 2018). There is also evidence that lexical diversity shows a style effect, whereas lexical sophistication is less influenced by style and is more task dependent. It has also been proposed that lexical diversity is Formulator-linked, whereas lexical sophistication is Conceptualiser-linked (Skehan, 2018). In any case, these findings suggest that ratings of vocabulary use might attempt to distinguish between these two measures. Even so, there has to be a sense of realism as to the number of operations raters can make within the time constraints they have (O'Grady, 2023) – this is more of an implication for research, than immediate practicality.

The remaining component of CALF, fluency, is the most intriguing of all, and arguably the most complex. The measures used in task research have consistently suggested subdimensions of in this area, including speed, breakdown (silent pauses), retrievable disruption (filled pauses), and repair. Task research has also brought out the importance of place of disruption, with pausing at clause boundaries being more similar to native speaker dysfluency, and within-clause pausing being less native-like (bearing in mind that native speakers, too, are often dysfluent) (Skehan, 2009). So, there is immediate relevance to testing – the different sub-dimensions may each need some mention in scales that might be used in testing, though perhaps worth doing only in contexts where fluency is of central importance.

Two additional points are worth making. It could be argued that fluency is the area where already there has been most cross-fertilisation between task research and testing. Tavakoli et al. (2020), for example, has shown how task-based measures make important contributions, with such measures helping to distinguish between CEFR levels up to B2, but then proving less effective in separating B2 from C-level L2 speakers. The other point, touched on earlier, is the importance of style. Some aspects of fluency (filled and unfilled pauses, repair) show considerable cross-task consistency, while speed does not do so as much. These results suggest that there may be characteristic approaches to fluency, irrespective of task, with features of performance such as pausing more a characteristic of the person than the level of proficiency. There are limitations, in other words, on what tasks (and tests) can do to separate out test-takers in aspects of fluency.

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

5. Towards a framework for testing speaking

We can now recapitulate the earlier sections, before going on to propose a general framework for the testing of speaking.

- Tasks and task conditions do influence performance (more consistently with the latter than the former, perhaps), and so a test performance, and rating, might be partly the result of the specific tasks and conditions used in a particular assessment.
- Broadly the influences from tasks and task conditions may be different for Conceptualiser and Formulator stages of L2 speaking, and this may have relevance for test-task difficulty. This may be important for sampling, calibrating difficulty, and generalising test results to real-world performance.
- L2 speakers have to manage resources/knowledge sources/competences/lexicons where these are limited in nature. How they manage such limitations has a huge potential influence on performance. Ability for Use needs to be a major consideration in testing.
- Task measurement research can provide insights for testing, particularly relevant to analytic rating scale construction.

These points are the background to the development of a framework for the testing L2 speaking, as shown in Figure 2.

Starting at the right-hand side, and based on the Levelt model (Kormos, 2006), we have the three major knowledge sources that underpin performance: background knowledge, the L2 mental lexicon, and declarative knowledge of the L2. The only relevant finding from the task research section is the correlation between a conventional proficiency test and measured task performance (Wang et al., 2019). This correlation is higher when there is planning, whether pre-task or on-line, but particularly with on-line planning, that is, little time pressure. The finding suggests that this task condition can more directly tap particular knowledge sources, most likely L2 declarative knowledge, which is more accessible when more time is available. One wonders whether other test-task formats might

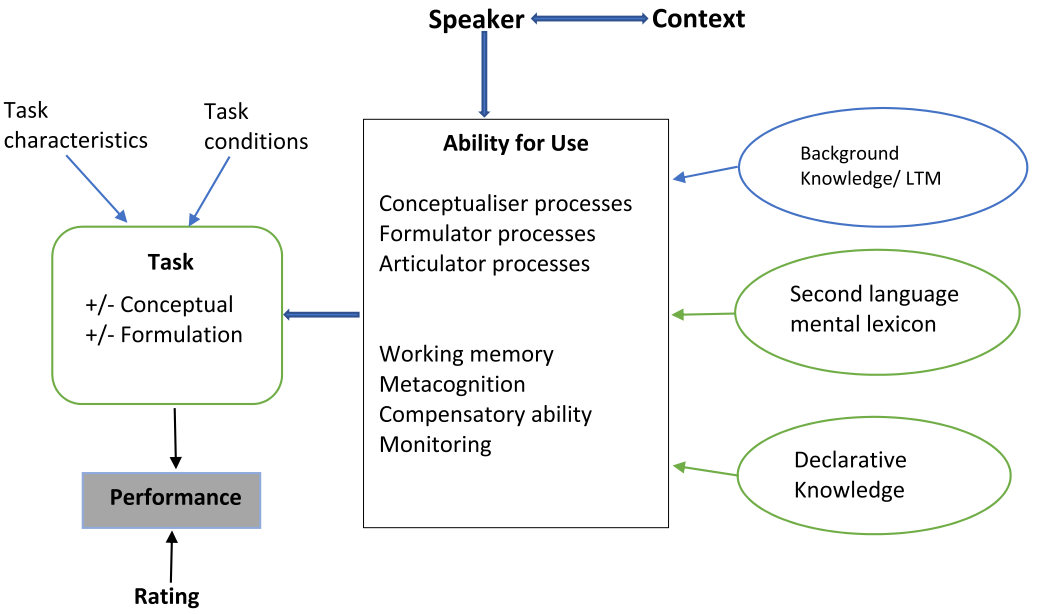

Figure 2. A general framework for testing speaking

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

catalyse the prominence of the other knowledge sources (e.g. specialist expertise and background knowledge). If we believe that particular underlying knowledge sources are important, then exploring how to identify their specific contribution may be very useful in a testing context for both general as well as specific-purpose testing.

In central place in the framework of Figure 2 is Ability for Use, and this is shown as comprising both the major stages of speaking, and also the other, speaker-linked attributes and strategies, especially metacognition. The first of these suggests that variation in how the stages of speaking are handled is important, and varies between people. The most relevant finding from previous discussion is Wang (2014) who reports that a combination of pre-task planning and on-line planning produces the greatest impact, across the performance dimensions (complexity, accuracy, lexis, and fluency). Conceptualiser processes (from pre-task planning) give content and ideas (complexity and lexis), and Formulator processes (from on-line planning) give effective means to produce actual language (particularly accuracy and fluency). Beyond this, pre-task planning, if associated with macro-planning, can give greater range to pre-verbal messages and this can sustain Formulator operations for several speaking 'turns'. There are also findings on the (limited) effects of working memory size – they exist, but may require supportive circumstances to manifest themselves. Working memory differences do not make much of a difference in the hurly-burly of normal speaking.

The remaining components of Ability for Use are supported by a qualitative study of planning (Pang & Skehan, 2014). Planning opportunities do not always occur, of course, but data from this type of study provides a window into the processes of L2 speaking. What emerges clearly from the retrospective reports is a picture not of 'passive' communicators', labouring away at transmitting messages, but rather speakers who can show considerable self-awareness, and management of their linguistic (and other) resources. These findings suggest that a significant influence on spoken language is the decisions that are made by speakers as they try to relate the resources they possess to the task they have to do. Strategies, that is, can often trump resources, or lack thereof. If one is looking for the basis for making generalisations from test information to real-world performance, then drawing on these abilities in a testing format is vital – Ability for Use is likely to transfer across contexts just as dependably as underlying knowledge.

Spoken performance also depends, obviously, on the speaking task that is involved (and see Weir (2005) on his analysis of 'task' within context validity). It is here, perhaps, that task research speaks most directly in its implications for testing. More theoretically, there is the point that task difficulty may vary independently for Conceptualiser and Formulator stages in speaking: what makes a task difficult at the first of these stages may be different to what makes tasks difficult for the second, and so an overall, one-dimensional idea of task difficulty may be difficult to defend. More empirically, we have seen a range of generalisations emerge from task research, regarding task characteristics and task conditions, and this too falls nicely into the frame provided by the Conceptualiser–Formulator distinction (effects of information type and operations, task conditions, and so on). Consequently, the performance that results may be based on these task factors, not simply the resources and abilities of the second language speaker. The findings suggest that tasks and the conditions in which they are completed are, as mentioned earlier, not neutral in their impact, and this needs to be considered in comparing test results, generalising from test results to real-world performances, and designing test batteries to obtain a wide-ranging sample of language.

Performance measures, the next stage in Figure 2, are important in assessing that performance, and here, too, task research has contributions to make. First, detailed task measurement may be suggestive about aspects of performance that could be relevant in test ratings. The two measures in each of structural and lexical complexity are examples of this, since while each pair focusses within the same area, they do not correlate with one another, and seem to tap different aspects of the dimension concerned. These findings are suggestive of useful cross-fertilisation of ideas between task research and language testing. Second, there are indications of style – for example, with things like fluency, especially pausing and repair – which may suggest that aspects of performance are not task-mediated but person-mediated.

https://doi.org/10.1017/S0261444823000344 Published online by Cambridge University Press

6. Implications and conclusions

In summing up, two qualifications are important. The first, already touched on, is the relevance of task research for testing. This point comes from the way some task-based results have not been replicated in some testing-oriented studies. This is important, and we have to await further research to see how strong this limitation might be. There is, though, as mentioned earlier, the point that if testing contexts themselves introduce new constraints on performance, this may be a problem for testing, just as much as for potential for task application. The second qualification, also touched on already, but worth restating, is that the discussion has been framed around task research findings, as though these findings are startlingly new. Yet, many of the claims that are being made are familiar to test developers. They may not be grounded in task research results, but they are the result of very considerable experience in thinking about good test-tasks.

The major point from this concluding section, though, is the issue of generalisation. Testing is necessarily done in controlled circumstances, and with a degree of time pressure that constrains breadth and extent of sampling. Yet, test information is required to overcome these limitations and to give us information about how people will perform beyond the test (Weir, 2005). So, a central testing challenge is to obtain information as efficiently as possible to enable such generalisations to be made, whether to more specific target situations, or to broader contexts of use. As Figure 2 indicates, in the area of testing speaking, there are 'a lot of moving parts' that complicate this challenge. Clearly, underlying knowledge sources are important (as in every other model), but so are Ability for Use, task characteristics and task conditions. The sampling implications are very important. All these stages are vital in providing the framework to elicit a systematic coverage of speaking abilities. The consequence of this is simply that, if generalisation to real world performance is the goal, it is not enough to sample underlying competences. The capacity to use these resources also becomes a vital area to sample, and the point of the present article is to offer some structure, based on task research, for the ways these other, performance-oriented factors, can play a role in language testing. Of course, there is much more to the testing of speaking than simply a task-research based perspective. It is clear, though, that task research can make an important contribution to such testing, and it is to be hoped that this plenary has contributed to this.

Acknowledgements. The author would like to thank Gavin Bui, John Norris, Stefan O'Grady, Graeme Porte, and three anonymous reviewers who have strengthened this written version of the plenary considerably. Thanks are also due to the Language Testing Research Colloquium organisation, whose invitation was the stimulus for this work.

References

Ahmadian, M. J. (2012). The relationship between working memory capacity and L2 oral performance under task-based careful online planning condition. *TESOL Quarterly*, *46*(1), 165–175. doi:10.1002/tesq.8

Bachman, L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

Bachman, L. F. (2002). Some reflections on task-based language performance assessment. *Language Testing*, *19*(4), 453–476. doi:10.1191/0265532202lt240oa.

Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice*. Oxford University Press.

Bui, G., Skehan, P., & Wang, Z. (2019). Task condition effects on advanced level foreign language performance. In P. Malovrh & A. Benati (Eds.), *Handbook of advanced proficiency in second language acquisition* (pp. 219–238). Wiley. doi:10.1002/9781119261650.ch12

Bui, H. Y. G. (2014). Task readiness: Theoretical framework and empirical evidence from topic familiarity, strategic planning, and proficiency levels. In P. Skehan (Ed.), *Processing perspectives on task performance*. (pp. 63–94). John Benjamins. doi:10.1075/tblt.5.03gav

Canale, M. (1983). *On* some dimensions of language proficiency. In J. Oller (Ed.), *Issues in language testing research* (pp. 333–342). Newbury House.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, *1*(1), 1–47. doi:10.1093/applin/i.1.1.

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics*, *13*(1), 1–24. doi:10.1093/applin/13.1.1.

Ellis, R. (2009). The differential effects of three types of task planning on fluency, complexity, and accuracy in L2 oral performance. *Applied Linguistics*, *30*(4), 474–509. doi:10.1093/applin/amp042

Ellis, R., & Yuan, F. (2005). The effect of careful within-task planning on oral and written task performance. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 167–192). John Benjamins. doi:10.1075/lllt.11.11ell

Foster, P., & Skehan, P. (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, *18*(3), 299–324. doi:10.1017/s0272263100015047

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case *for* a weighted clause ratio. *Annual Review of Applied Linguistics*, *36*(1), 98–116. doi:10.1017/s0267190515000082

Fulcher, G. (2015). *Re-examining language testing: A philosophical and social inquiry*. Routledge. doi:10.4324/9781315695518

Galaczi, E. B. (2006). Peer-peer interaction in a speaking test: The case of the First Certificate in English examination. *Language Assessment Quarterly*, *5*(2), 89–119. doi:10.1080/15434300801934702

Gilabert, R., Baron, J., & Llanes, M. (2009). Manipulating task complexity across task types and its influence on learners. *International Review of Applied Linguistics*, *47*(3), 367–395. doi:10.1515/iral.2009.016.

Hulstijn, J. (2015). *Language proficiency in native and non-native speakers: Theory and research*. John Benjamins. doi:10.1075/lllt.41

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin Books.

Inoue, C. (2016). A comparative study of the variables used to measure syntactic complexity and accuracy in task-based research. *The Language Learning Journal*, *1*(1), 1–18. doi:10.1080/09571736.2015.1130079.

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum. doi:10.4324/9780203763964

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

Luo, S. (2008). Re-examining factors that affect task difficulty in TBLA [Unpublished Ph.D. dissertation]. Chinese University of Hong Kong.

Malicka, A., & Sasayama, S. (April 17th-19th, 2017). *The importance of learning from the accumulated knowledge: Findings from a research synthesis on task complexity*. Paper presented at the 7th Biennial International Conference on Task-Based Language Teaching, Barcelona, Spain.

Mehnert, U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition*, *20*(1), 52–83. doi:10.1017/S0272263198001041.

Nitta, R., & Nakatsuhara, F. (2014). A multi-faceted approach to investigating pre-task planning effects on paired oral test performance. *Language Testing*, *31*(2), 147–175. doi:10.1177/0265532213514401

Norris, J. (2018). Task-based language assessment: Aligning designs with intended uses and consequences. *JLTA Journal*, *21*(1), 3–20. doi:10.20622/jltajournal.21.0_3

Norris, J. M., Brown, J. D., Hudson, T. D., & Yoshioka, J. K. (1998). *Designing second language performance assessments*. University of Hawai'i Press.

O'Grady, S. (2019). The impact of pre-task planning on speaking test performance for English medium university admission. *Language Testing*, *36*(4), 505–526. doi:10.1177/0265532219826604

O'Grady, S. (2023). Halo effects in rating data: Assessing speech fluency. *Research Methods in Applied Linguistics*, *2*(1). 10.1016/j.rmal.2023.100048.

Pang, F., & Skehan, P. (2014). Self-reported planning behaviour and second language performance in narrative retelling. In P. Skehan (Ed.), *Processing perspectives on task performance*. (pp. 95–128). John Benjamins. doi:10.1075/tblt.5.04pan

Pang, F., & Skehan, P. (2021). Performance profiles on second language speaking tasks. *Modern Language Journal*, *105*(1), 371–390. doi:10.1111/modl.12699

Purpura, J. (2016). Assessing Meaning. In E. Shohamy, & L. Or (Eds.), *Encyclopedia of language and education: Vol. 7. Language testing and assessment* (pp. 33–61). Springer International Publishing. doi:10.1007/978-3-319-02326-7_1-1

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. In M. Bygate (Ed.), *Domains and directions in the development of TBLT* (pp. 87–122). John Benjamins. doi:10.1075/tblt.8.04rob

Salisbury, K. (2010). The Edge of Expertise? Towards an understanding of listening test item writing as professional practice. [Unpublished Ph.D. dissertation]. King's College, London.

Skehan, P. (2001). Tasks and language performance. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Research pedagogic tasks: Second language learning, teaching, and testing* (pp. 167–185). Longman. doi:10.1017/9781108955638.035

Skehan, P. (2009). Lexical performance by native and non-native speakers on language-learning tasks. In B. Richards, H. Daller, D. D. Malvern, & P. Meara (Eds.), *Vocabulary studies in first and second language acquisition:The interface between theory and application* (pp. 107–124). Palgrave Macmillan. doi:10.1057/9780230242258_7

Skehan, P. (Ed.) (2014). *Processing perspectives on task performance*. John Benjamins. doi:10.1075/tblt.5

Skehan, P. (2016). Tasks vs. conditions: Two perspectives on task research and its implications for pedagogy. *Annual Review of Applied Linguistics*, *36*(1), 34–49. doi:10.1017/s0267190515000100

Skehan, P. (2018). *Second language task-based performance: Theory, research, and assessment*. Routledge. doi:10.4324/9781315629766

Skehan, P. (2022). Working memory and second language speaking tasks. In J. W. Schwieter, & Z. Wen (Eds.), *The Cambridge handbook of working memory and language* (pp. 635–655). Cambridge University Press. doi:10.1017/9781108955638.035

Skehan, P., & Foster, P. (1997). The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research*, *1*(3), 185–211. doi:10.1177/136216889700100302

Skehan, P., & Foster, P. (1999). The influence of task structure and processing conditions on narrative retellings. *Language Learning*, *49*(1), 93–120. doi:10.1111/1467-9922.00071

Skehan, P., & Foster, P. (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, & I. Vedder (Eds.), *Complexity, accuracy, and fluency in second language use, learning, and teaching* (pp. 207–226). University of Brussels Press.

Skehan, P., & Luo, S. (2020). Developing a task-based approach to assessment in an Asian context. *System*, *90*(1), 1–14. doi:10.1016/j.system.2020.102223.

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. doi:10.1111/j.1467-9922.2008.00446.x

Tavakoli, P., Nakatsuhara, F., & Hunter, A.-M. (2020). Aspects of fluency across assessed levels of speaking proficiency. *Modern Language Journal*, *104*(1), 169–191. doi:10.1111/modl.12620

Tavakoli, P., & Skehan, P. (2005). Planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–276). John Benjamins. doi:10.1075/lllt.11.15tav

Wang, Z. (2014). On-line time pressure manipulations: L2 speaking performance under five types of planning and repetition conditions. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 27–62). John Benjamins. doi:10.1075/tblt.5.02wan

Wang, Z., & Skehan, P. (2014). Task structure, time perspective and lexical demands during video-based narrative retellings. In P. Skehan (Ed.), *Processing perspectives on task performance* (pp. 155–186). John Benjamins. doi:10.1075/tblt.5.02wan

Wang, Z., Skehan, P., & Chen, G. (2019). The effects of hybrid on-line planning and L2 proficiency on video-based speaking task performance. *Journal of Instructed Second Language Acquisition*, *3*(1), 53–80. doi:10.1558/isla.37398.

Weir, C. J. (2005). *Language testing and validation: An evidence-based approach*. Palgrave Macmillan. doi:10.1057/9780230514577

Wen, Z. (2009). Effects of working memory capacity on L2-based speech planning and performance [Unpublished Ph.D. Dissertation]. Chinese University of Hong Kong.

Willis, J., & Willis, D. (1988). *The Collins COBUILD English course: Level 1*. Collins.

Xi, X., Norris, J. M., Ockey, G. J., Fulcher, G., & Purpura, J. (2021). Assessing academic speaking. In X. Xi, & J. M. Norris (Eds.), *Assessing academic English for higher education admissions* (pp. 152–199). Routledge. doi:10.4324/9781351142403.

Peter Skehan is an Honorary Research Fellow at the Institute of Education, University College London. He has taught at universities in the U.K., Hong Kong, and New Zealand. His main interests are second language acquisition, particularly task-based instruction, and also foreign language aptitude, as well, earlier in his career, in language testing. He has published *Individual differences in second language learning* (Arnold, 1989); *A cognitive approach to language learning* (OUP, 1998), and *Second language task-based performance: Theory, research, and assessment* (Routledge, 2018), as well as edited collections such as, most recently, *Language aptitude: Theory and practice* (with Edward Wen and Richard Sparks: CUP, 2023). He has also published research articles on second language task-based performance, exploring issues such as the effects of pre-task planning, task characteristics, such as task structure, and post-task conditions. More theoretically, he has argued for the relevance of a Limited Capacity Approach to second language task-based performance.

Cite this article: Skehan, P. (2024). Linking second language speaking task performance and language testing. *Language Teaching*, *57*(4), 501–516. https://doi.org/10.1017/S0261444823000344