Chapter 8

Tasks and language performance assessment

Peter Skehan

Developments over the last 15 years or so have suggested that pedagogy can fruitfully be organised by means of tasks that learners transact, and that tasks can be used as the basis for syllabus organisation as well as the unit for classroom activities. As Chalhoub-Deville (Chapter 10) points out, however, such developments constitute a source of difficulty for achievement testing. Conventional approaches to testing link with sampling frames which can be organised around some structural organisation for a syllabus. Tasks, in contrast, are centrally concerned with the learner achieving some purpose and outcome, and do not directly require the use of conformity-oriented language (Willis, 1991), of the sort that it would be convenient to engage if a syllabus is to be tested systematically.

A move towards tasks also poses problems for abilities-oriented *proficiency* testing. The most influential approaches of this type (Canale and Swain, 1980; Bachman, 1990; Bachman and Palmer, 1996) posit an underlying structure of the components of competence, and then propose mediating mechanisms by which such competences will impact upon performance. In principle, such an approach might be extremely rewarding but, in practice, the codifying nature of the underlying competence-oriented models has not interfaced easily with effective predictions to real-world performances (Harley et al., 1990; Skehan, 1998). At the most general level, the problem is that underlying and generalised competences do not easily predict across different performance conditions or across different contexts. Moving from underlying constructs to actual language use has proved problematic.

In response to these difficulties, a number of investigators have proposed alternative models of how spoken language might be conceptualised and measured. These models attempt to portray the assessment event in more comprehensive ways which (a) incorporate a larger number of performance elements directly, and (b) clarify how research studies might be organised and integrated more effectively to give an empirical basis for the claims that are made about spoken language assessment. The model shown as Figure 8.1 is based on work by Kenyon (1992), McNamara (1995) and Skehan (1998). It is useful to discuss the various components of this model briefly before turning to the factor which is the main focus of this chapter – the influence

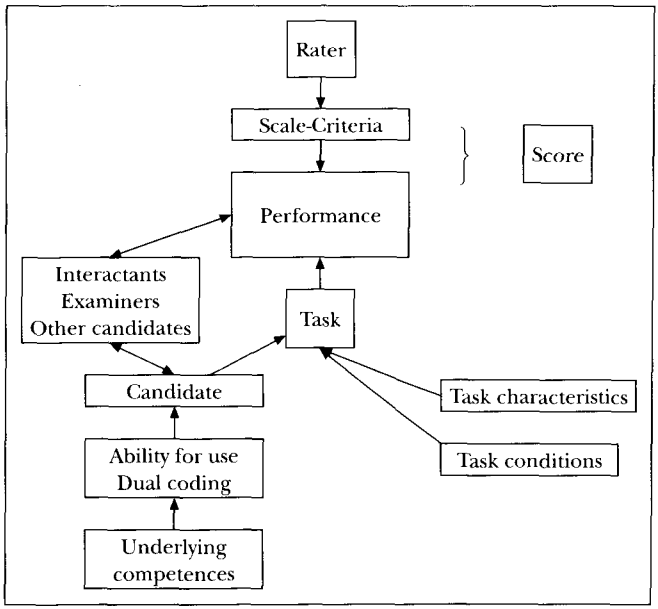

Figure 8.1 A model of oral test performance (Skehan, 1998)

of the task itself on assessment procedures. The model in general clarifies the potential fallibility of a test score as an indicator of underlying abilities. The section on tasks explores whether there are systematic influences on the nature of the performance which is elicited arising from task characteristics themselves.

Figure 8.1 shows that a test score is most immediately influenced by the rating procedures which have been used. The oral performance which has been elicited will have been judged by raters. In addition, the performance which is being rated will be filtered through a rating scale. Such scales vary in their origin, in their characteristics, and in their purposes (see Alderson, 1991; Fulcher, 1996a). As a result of these rater and scale factors, we have to consider the possibility that the score assigned to a candidate may not reflect candidate performance only, but may partly be based on biases and limitations arising from raters and scales.

Working systematically through the model, we can identify a number of additional influences on the score which is assigned. These fall into three major headings:

- the interactive conditions under which performance was elicited;
- the relevant abilities of the candidate;
- the task which was used to generate the performance, as well as the conditions under which the task was completed.

The interactive conditions under which performance is elicited have posed problems to oral language assessment that have been recognised for many

years. For example, in a conventional assessor–assessed arrangement, the power relations between the participants are manifestly unequal, and the asymmetry which results distorts the language subsequently used (Van Lier, 1989). It is also likely that there is important restriction in the functions of language which can be probed in any meaningful way. For these reasons, alternative organisational arrangements for oral testing have been tried in recent years, such as group-based encounters. At the cost of standardisation, they enable a wider range of language functions and roles to be engineered to provide a better basis for oral language sampling with less asymmetry between participants (Van Lier, 1989). At a more theoretical level, the group format enables us to portray the interaction in terms of co-construction, since participants will have some degree of mutual equality, and so the direction the discourse develops will not be pre-ordained and orchestrated by the assessor.

We turn next to the abilities of the candidate. Measuring these, one might say, is the major goal of the actual assessment procedure, so the first value of the dynamic represented in Figure 8.1 is to show how this ability may not have a dominant effect upon the score that is awarded because so many other factors intrude and in potentially unsystematic ways. Again, it is not the focus of this chapter to cover this area in detail, and so only a brief account will be given here. The model in Figure 8.1 suggests that we need to consider underlying competences *and* ability for use. The former is represented in models such as those proposed by Canale and Swain (1980) and Bachman (1990). Competence-oriented models describe different components of communicative competence and their interrelationships. They also propose some method by which such underlying competences might influence actual performance. The relevant section of Figure 8.1, however, takes this competence-to-performance linkage further, and proposes the construct of *ability for use* as a set of abilities which mediate between underlying competence and actual performance conditions in systematic ways (Skehan, 1995, 1998). It is then the goal of assessment techniques to devise methods of assessing this construct as well as the underlying competences.

Figure 8.1 describes what is largely a programmatic model. There has been significant research in the area of rater and rating scale influences (Lumley and McNamara, 1995; North, 1996). Further, proposals to describe underlying competences have received considerable theoretical and empirical attention in recent years, but the inclusion of interactive conditions and ability-for-use in the model is rather speculative at this point, and unconnected to any testing-oriented evidence. The same has been true of the influence of tasks until recently. The remainder of this chapter will be concerned with relevant research which tries to clarify how the task component of the model is increasingly susceptible to empirical investigation.

ASSESSING TASK DIFFICULTY: INTRODUCTORY ISSUES

This section will prepare the ground for a meta-analysis of a number of separate studies of task-based performance. Such an examination of a number

of different studies with common features is revealing about the way we might understand the impact of task characteristics on test performance. The section first discusses some measurement issues, and then describes the datasets used in the meta-analysis.

Measuring task performance

At the outset, one general issue needs to be clarified concerning the way task performance has been generally measured. Figure 8.1 has shown that it is typical, in assessing spoken performance, to use a rating scale approach. Such scales may be global scales, or they may be more analytic, with separate ratings for areas such as range, accuracy and fluency. In task-based research, in contrast, such rating scale measures are not typical (but see Wigglesworth, 1997 and this volume, for exceptions). Instead, reflecting the different psycholinguistic research tradition to which they belong, researchers into tasks have tended to use more precise operationalisations of underlying constructs.

In general, there is some consensus that measures are required in the three areas of complexity,[1] accuracy and fluency. These three areas are theorised to have important independent functioning in oral performance (Skehan, 1998).[2] In addition, they enter into competition with one another, with higher performance in one area seeming to detract from performance in others (Skehan and Foster, in press). So, for example, greater accuracy may well be achieved at the expense of greater complexity, and vice versa. Research is continuing to establish just how these three areas interrelate; however, a growing number of investigations into the task-based area are based on carefully computed indices in each of these three areas, and the competition between them will have an important impact in decisions that are made about task difficulty.

The datasets for the present research

The meta-analysis of task characteristics and their influence on task performance is based on six research studies conducted at Thames Valley University, with Pauline Foster as co-investigator. At the outset, it is essential to give a brief overview of the tasks that were used, and the purpose of the various studies (see Table 8.1).

ASSESSING TASK DIFFICULTY: TASK CHARACTERISTICS AND CONTRASTS

At the beginning of the series of research studies, the conceptualisation of task type was in terms of a contrast between personal, narrative and decision-making tasks. (These task types were chosen as maximally representative of tasks used in language-teaching coursebooks.) As the research programme has developed, however, it has become clear that this contrast between task

Table 8.1 Overview description of the six studies

Description	Tasks
Study 1: Comparison of the effects of planning on performance on personal, narrative and decision-making tasks (Foster and Skehan, 1996)	(a) *Personal*: Students* have to instruct their partner how to return to their homes, and then turn off an oven which has been left on. (Oven task) (b) *Narrative*: Students have to devise a story to a set of pictures: the pictures have common characters, but no obvious storyline. (Weave a story task) (c) *Decision*: Students are given a series of 'crimes', and have to agree on judicial sentences for these crimes, e.g. woman discovers husband in bed with another woman, and stabs and kills him. (Judge task)
Study 2: Comparison of the effects of planning on performance on personal, narrative and decision-making tasks, together with a comparison of the effects of a post-task and no post-task (Skehan and Foster, 1997)	(a) *Personal*: Students have to compare things that surprise them, pleasantly and unpleasantly, about life in Britain. (Surprise task) (b) *Narrative*: Students have to tell the story to a cartoon strip of pictures. The pictures have clear structure and an amusing ending. (Sempé task) (c) *Decision*: Students have to agree on advice to give letter writers to an Agony Aunt column. Each letter conveys a difficult problem, e.g. father is worried about his child living with mother and new partner in a drug-suspected communal house. How should he act? (Agony Aunt task)
Study 3: Comparison of the effects of planning on a decision-making task, together with comparison of the effects of introducing surprise information mid-task (Foster and Skehan ms)	*Decision*: Judge task, as in Study 1
Study 4: Comparison of four different conditions for video narrative retelling, with different processing loads. Three conditions require different versions of simultaneous tellings, i.e. telling the story while the video is playing, while the fourth is a delayed condition. Comparison of two tasks, one more structured and one less structured (Skehan and Foster, 1999)	(a) *Unstructured narrative*: Mr Bean video of Crazy Golf. In this video Mr Bean plays a round of golf, in which various mishaps occur. The events are an essentially disconnected series. (Golf task) (b) *Structured narrative*: Mr Bean video of restaurant meal. Mr Bean has a restaurant meal in which various amusing events occur, but against the background of a typical restaurant 'script'. (Restaurant task)

Table 8.1 *(cont'd)*

Description	Tasks
Study 5: Comparison of four different planning conditions: teacher-fronted, solitary, group-based and control (Foster and Skehan, 1999)	*Decision-making*: Balloon debate, with occupants to defend of actor, politician and EFL teacher. Students are assigned pre-task groups for teacher and group conditions where particular planning conditions are implemented. Then, students are assigned new groups and have to argue a position as to who should be thrown from the balloon. (Balloon task)
Study 6: Comparison of two different experimental post-task conditions (based on the need for participants to complete a transcription of their own recorded task-based performance) and a control group (Skehan and Foster, ms)	(a) *Decision-making*: Agony aunt task, as abo (b) *Narrative*: Picture cartoon strip. (Sempé task, as above)

* All studies were completed with students working in pairs, except for Study 4, where groups of four students were involved.

types, although useful, is not the whole story, by any means. It was originall thought that it would be possible to generate a number of (roughly equivalent personal, narrative and decision-making tasks. As the research results accu mulated, however, it became clear (somewhat unsurprisingly, in retrospect that not all exponents of each of these task types were indeed the same a regards the complexity, accuracy and fluency of the language produced. It became clear, in other words, that other, finer-grained features, operating at a more basic level, were influential. Where there were differences between the more global task types that had been the starting point for the research this was probably due to the combination of finer-grained factors that hap pened to be shared.

On the basis of the emerging results, Skehan (1998) proposed the following set of task characteristics which might impinge upon the nature of perform ance (in each case, the studies from the Thames Valley programme which bear upon the variable in question are indicated):

- *Familiarity of information*: Tasks vary as to whether they require information that is familiar to the participants because it is part of their personal experience, compared to tasks which require the assimilation of material presented by the experimenter. Tasks based on familiar information are Study 1(a), and Study 2(a), where, in each case, the retrieval of personally relevant information, which is well known to the participants, becomes the basis for completing the task. In Study 1(a) participants' contributions are based on their route home from college and house layout. In Study 2(a) participants

describe what surprises them about life in Britain, pleasantly and unpleasantly. Both these tasks (and all others) were piloted, and in the piloting performance gave no indications of effortful retrieval – such material as participants wanted to use seemed instantly available to them.

- *Dialogic vs monologic*: Some tasks require interaction, and a discourse style that leads participants to alternate in who holds the floor, compared to others where extended turns are required, with little need to interact other than listen and wait for one's turn. A sub-set of monologic tasks are narratives, where one participant tells a story. Clear dialogic tasks are the Judge task (Study 1(c) and Study 3), and the Agony Aunt task (Study 2(c) and Study 6(b)). Each of these is a decision-making task. There was one other such task, the Balloon debate, (Study 5), but in this case although the task was completed in groups of four, there was little dialogic performance when the learners 'took the floor' to defend their different characters – instead learners 'declaimed' at some length. The narratives in the research were Study 1(b) (Weave a story), Study 2(b) (Sempé cartoon), Study 4 (both tasks: video-based narratives) and Study 6(a) (Sempé cartoon). These were completed by pairs of students, with each taking it in turn to tell the narrative and be asked questions.
- *Degree of structure*: Some tasks contain a clear macrostructure, with the time sequence underlying the task fairly clearly identifiable. Other tasks do not have this clear over-arching structure. Examples of structured tasks are: Study 1(a) (personal), Study 2(b) (narrative), Study 4(b) (structured narrative) and Study 6(b) (the same task as Study 2(b)). In all these cases, the time line for the information underlying the task is clear and well organised, with the different stages in each case having a clear relationship with one another.
- *Complex outcomes*: Some tasks require only straightforward outcomes, in which a simple decision has to be made. Others require multi-faceted judgements, in which the case or position a learner argues during a task can only be effective if it anticipates other possible outcomes, and other learners' contributions. In the present research, this functions as a sub-category of dialogic tasks, in that the clearest comparison is between the Agony Aunt task from Studies 2 and 6, on the one hand, and the Judge task from Studies 1 and 3, together with the Balloon debate from Study 6, on the other. The Agony Aunt task is the only one which requires joint engagement with the ideas concerned, as opposed to superficial negotiation of appropriate custodial sentences (in the Judge task) or ejection from the Balloon.
- *Transformation*: Some tasks do not require participants to operate upon the information presented or retrieved, but instead simply to reproduce it. Others require some degree of on-line computation which changes the state or the relationship of the elements in the task. Most of the tasks in the six research studies do not require transformation of this sort, emphasising instead retrieval of information or judgements about material which is presented. An exception is the narrative task from Study 1, where participants had to use their imaginations to 'impose' a story structure upon an unrelated

series of pictures. In so doing, the shared characters within the given picture set had to be transformed in their relationships with one another. In this respect, it comes close to what Brown et al. (1984) term a dynamic task, except that here the dynamic qualities are derived from the mental operations of the participants themselves, rather than from the unfolding events of a *given* story.

We now have five task characteristics which can be investigated through the six studies in the Thames Valley research. These studies can be used to explore whether these different characteristics have systematic influences upon performance. It has to be admitted, however, that these comparisons were not planned at the outset of this research programme. They are nevertheless worth pursuing since the various studies do share sufficient common features to justify the comparisons which are made. In particular, the scoring of the dependent variables was approached in a fairly consistent manner, so that the scores which are quoted below can be validly related to one another. However, the numbers and nature of the tasks which enter into the comparisons are not so systematic. The generalisations which are offered below should therefore be treated as tentative, and the basis for future, more systematically organised research studies.

ASSESSING TASK DIFFICULTY: EMPIRICAL RESULTS

The following section will detail the results for each of the five task characteristics mentioned above. In each case, three sets of measures will be used to assess the various tasks concerned. These are complexity, accuracy and fluency. Complexity is measured through a subordination index. Data are coded into communication units and clauses (Foster et al., ms), and then an index is calculated representing number of clauses per *c*-unit. This has been shown in the research programme to be a sensitive measure of the degree of subordination in spoken language. It is also taken to be a surrogate measure of general language complexity. Accuracy is calculated as the proportion of clauses which are error-free. Finally, fluency is measured by the number of pauses greater than 1 second in duration per 5 minutes of performance. Better performance is therefore indicated by *higher* complexity and accuracy scores and *lower* numbers of pauses.

In the course of the research programme a number of other measures have been explored, such as the range of different syntactic structures that are drawn on; the type-token ratio of the lexis; or dysfluency indicators such as reformulation and repetition. The measures which are actually used in these analyses are those which have proved most sensitive to experimental differences, as well as the most clearly defined for operational purposes. It is not claimed that they are definitive measures (and indeed, reviews such as Wolfe-Quintero et al. (1998) are showing the range of measures that can be used in this regard). They are now serviceable and have been used by the Thames

Valley research team, and others (e.g. Wigglesworth, 1997), with encouraging results.

Familiarity of information

Skehan (1998) hypothesises that familiarity of information will lead to greater fluency and accuracy of performance, since the easy access to information should make only limited demands on attention, allowing material to be assembled for speech more easily, and with greater attention to form. He proposes that there will be no push towards greater complexity as a result of the greater familiarity, since speakers will be likely simply to draw upon well-established language to code familiar events.

The most direct test of this hypothesis is to examine the results from Studies 1 and 2, where, in each case, tasks that are based on familiar information (the two personal tasks) can be compared with tasks that are not. The relevant results are shown in Table 8.2.

Table 8.2 Familiar information, fluency and accuracy

		Personal	Narrative	Decision	Sig.
Accuracy:	Study 1	0.68	0.62	0.67	0.05
	Study 2	0.66	0.62	0.68	0.01
Fluency:	Study 1	14.8	22.2	27.1	0.001
	Study 2	23.1	17.8	21.7	0.04

In Studies 1 and 2, the personal task does indeed generate higher accuracy than the narrative, but so does the decision task, in each case. It appears, therefore, that while the results are not inconsistent with familiar information leading to greater accuracy, the supporting evidence is not strong, since there are alternative routes to achieving greater accuracy of performance. Turning to fluency, the results from Study 1 are supportive of the original hypothesis. The personal task in this case is associated with fewer pauses and greater fluency than in the other two tasks. In this case, it does appear that familiar information is associated with less interruption to the speech flow. However, these results are not particularly supported when we look at Study 2. In this case, the personal task produces the *least* fluent performance, with the narrative generating the fewest pauses, and the decision task leading to more fluent performance. Additional analyses were carried out with the planning variable, since it is possible that there might be an interaction with planning, such that when there is time to prepare, familiar information might be selectively associated with more fluent performance. The results, however, are not supportive of this: the same patterns occur under all planning conditions.

The evidence, therefore, is not strongly supportive of an effect for familiarity of information on either accuracy or fluency. The present results are based on a meta-analysis of studies not intended to make sustained systematic comparisons, and so it may be that other correlating variables are obscuring

potential relationships. The accuracy achieved from the dialogic decision-making tasks may be a case in point, since, as will be shown below, such tasks independently and consistently generate greater accuracy. The comparison made in Table 8.2 may not therefore be the best one to judge the effects of familiar information on task performance. What we can say, though, is that familiarity does not have such a strong effect on performance that higher accuracy is guaranteed. In other words, the effect seems weaker than was anticipated. Similarly, the effects upon fluency may depend on factors additional to the information itself. In the personal task in Study 2, for example, the need to retrieve information may introduce a strong processing element into performance, such that fluency is disrupted. This, however, brings us to the point made immediately above: familiar information does not guarantee more attention being available to achieve a higher level of performance.

Dialogic tasks

Skehan and Foster (in press) propose that interactive tasks are associated with greater accuracy and complexity, but lower fluency. They suggest that such effects are due to:

- *greater accuracy*
  - communication-driven push towards precision
  - 'creation' of more time to focus on form, as partner is speaking
  - recycling of partner's language, both with tendency to re-use correct language and to edit and correct it;
- *greater complexity*
  - collective reinterpretation of the task to make it more complex
  - scaffolded elaboration of partner's language;
- *lower fluency*
  - need to accommodate the unpredictability of partner's contributions, i.e. greater need to engage in on-line planning
  - uncertainty of turn-taking, and consequent disruption to fluency.

The descriptive statistics for the relevant comparisons are shown in Table 8.3. The comparisons involve:

- Studies 1 and 2, where dialogic (decision-making) tasks were contrasted with narrative and personal tasks;
- Study 6, where a dialogic (decision-making) task was contrasted with a narrative;
- Study 5 vs Studies 1, 2, 3 and 6, where the comparison was between *different* decision-making tasks, in that Study 5 used a Balloon debate, which was essentially monologic, compared to all other decision-making tasks which were much more interactive in nature.

In the first comparison, for Studies 1 and 2, a one-way within-subjects analysis of variance yields an F value of 5.64, and a significance level of $p < 0.001$. However, the significance is located in the contrast between the

Table 8.3 Accuracy, complexity and fluency on dialogic vs non-dialogic tasks

Study	Accuracy (% of error-free clauses)			Complexity (clauses per c-unit)			Fluency (No. pauses per 5 mins)		
	Decis. making	Nar.	Pers.	Decis. making	Nar.	Pers.	Decis. making	Nar.	Pers.
1 ($N = 32$)	0.67	0.61	0.68	1.32	1.35	1.16	27.1	22.3	14.8
2 ($N = 40$)	0.68	0.62	0.68	1.67	1.31	1.37	21.7	17.8	23.1
3 ($N = 60$)	0.68	–	–	1.41	–	–	22.8	–	–
5 ($N = 66$)	0.61	–	–	1.44	–	–	8.6	–	–
6 ($N = 42$)	0.68	0.56	–	1.47	1.35	–	12.8	10.8	–

narrative task accuracy level of 61% error-free clauses and the decision-making (67%) and personal (68%) tasks. In other words, while the decision-making (dialogic) task yields significantly more accuracy than the narrative task, it is not significantly different from the personal task. The other comparisons, are, however, clearer in their results. In the second comparison, the *t*-test between the decision-making task (68% accuracy) and narrative tasks (56% accuracy) in Study 6 generates a *t*-value of 4.14, translating to a significance level of $p < 0.001$. In the third comparison, the between-subjects *t*-test shows that the more monologic Study 5 decision-making task (61% accuracy) is significantly different from the other (dialogic) decision-making tasks (mean 68% accuracy), with a *t*-value of 3.56 and a significance value of $p < 0.001$. In a guarded fashion, therefore, and provided other relevant variables do not intrude, it can be claimed that dialogic tasks are associated with fewer errors.

A similar mixed picture emerges with the complexity results. The first comparison, Studies 1 and 2, does generate a significant effect for the dialogic decision-making task compared to the narrative and personal tasks ($F = 15.6$; $p < 0.001$), but this result should be modified in that the complexity mean for Study 1 narrative is actually higher than that for the decision-making task in that study (1.35 vs 1.32). The second comparison, for Study 6, does produce a clear result. The comparison yields a *t*-value of 1.84, which is significant at the $p < 0.05$ level (one-tailed test). However, the comparison between the more monologic Study 5 decision-making task and all the other (dialogic) decision-making tasks is not significant. This suggests that, as with the accuracy results, dialogic tasks *tend* to be associated with greater complexity, but this effect is mediated by other factors.

We turn finally to the fluency results. In the first comparison, for Studies 1 and 2, the dialogic task generates less fluency than the other two tasks ($F = 7.93$; $p < 0.001$). Once again, however, the results are not completely clear-cut, in that the personal task from Study 2 generates slightly less fluent performance than the dialogic decision-making task from this study. The trend, however, seems to be in the direction of lower fluency being associated with interaction. The second comparison, from Study 6, does not yield a

significant result ($t = 0.66$; $p > 0.05$). The third (between-subjects) comparison between the Study 5 decision-making task and all the other decision-making tasks ($t = 8.35$; $p < 0.001$) is highly significant, with the Study 5 more monologic decision-making task generating much more fluent language than that in the more interactive tasks.

Returning to the rationale for these effects proposed at the beginning of this section, it is clear that the next stage of the research is to return to the transcripts of the different performances to see whether the factors which are proposed to account for the task difference effects can be detected in the actual data. The quantitative results are mixed, and now need to be triangulated from another data source. It is encouraging, however, that the results described here complement those reported in Bygate (this volume), especially for complexity and fluency, in relation to the narrative and interactive tasks.

Degree of structure

Skehan (1998) proposes that this variable has an effect upon the fluency and accuracy of performance. He suggests that tasks which contain clear structure, especially sequential structure, facilitate task performance by clarifying the macrostructure of the speech event. As a result, the lack of need to engage in large-scale planning frees up attentional resources for on-line planning. This additional attention, he proposes, is directed towards the immediate goals of avoidance of error and breakdowns in communicational flow, i.e. accuracy and fluency.

Three tasks, out of the total set used, were identified as containing greater structure. These were the personal task from Study 1 (turn off the oven); the narrative task from Studies 2 and 6 (the Sempé story, i.e. the same task in each study), and the 'restaurant' task from Study 4. The first source of evidence here comes from within subject comparisons from Study 1, where the personal task results can be opposed to those for the other two tasks. Regarding fluency, the within-subjects one-way analysis of variance is significant ($F = 16.6$; $p < 0.001$), with the personal task generating significantly more fluency than the other two tasks. With respect to accuracy, the corresponding analysis also indicates significance ($F = 3.58$; $p < 0.05$), but the operative contrast is between the personal *and* decision-making tasks, which are significantly more accurate than the narrative. These results, therefore, provide only partial support.

The findings for the comparison from Study 2 show similar results for fluency ($F = 4.7$; $p < 0.04$), with the narrative task (the structured Sempé story) generating significantly fewer pauses than the other two tasks. The results for accuracy from Study 2 are, however, very different. Significance is achieved ($F = 6.2$; $p < 0.01$), but the significant contrasts are in the reverse direction to those predicted, with the Sempé task associated with *lower* accuracy at 62% error-free clauses. On this occasion, at least, the structured task did not produce greater accuracy. The same two tasks were used in Study 6,

where *t*-tests produced exactly the same results as for Study 2. Fluency was greater in the structured Sempé task, but this difference did not attain statistical significance ($t = 0.66$). Accuracy, however, was clearly greater in the decision-making Agony Aunt task ($t = 4.14$; $p < 0.001$). No different pattern emerges if these studies are analysed at a greater level of detail by examining the different planning conditions.

The trend towards clearer effects for fluency rather than accuracy is particularly evident in Study 4. Relevant results are shown in Table 8.4.

Table 8.4 Structured and unstructured narratives

Variable	Golf task ($N = 21$) (mean)	Restaurant task ($N = 24$) (mean)
Repetition	39.3	19.1
False starts	29.5	15.5
Reformulations	10.9	4.8
Replacements	8.2	5.2
Accuracy	47%	50%

All the fluency effects shown in Table 8.4 generate significant differences, all beyond the $p < 0.001$ level of significance. It is clear also that when fluency is operationalised in terms of repetition, etc.,[3] the structured task generated roughly *half* the amount of disfluency that the unstructured task generated. However, the accuracy effects, although showing a very slight superiority for the structured 'restaurant' task, do not remotely approach significance.

We can summarise the results in this section by saying that there is a fairly consistent pattern that tasks based on more structured information seem to be associated with greater fluency. There are some indications that accuracy might also be enhanced, but the evidence is, to say the least, mixed, and so it would be unwise at this stage to make any claims in this direction. If on-line planning and attentional availability are facilitated by structured tasks, these are directed towards fluency.

Complexity of outcome

This contrast is restricted to the decision-making tasks, and opposes the tasks which are susceptible to minimal interpretation for outcome, enabling low-level negotiation of consensus (Judge and Balloon tasks), and those which require engagement and careful examination of the different facets of a decision (Agony Aunt task). The results in this case are clear cut. Skehan (1998) predicts selectively for complexity here, and the differences found are indeed confined to this area. The relevant data are presented in Table 8.5 (overleaf).

Given that these comparisons are based on large groups (of 82 and 157 participants respectively), they represent powerful evidence that the complexity of task outcome is a major influence upon the complexity of the

Table 8.5 Complexity of outcome and task performance

Variable	Mean score complex outcome	Mean score simpler outcome	*t*-Value	Sig.
Accuracy	0.68	0.65	1.51	0.24
Complexity	1.59	1.41	5.01	0.001
Fluency	17.5	17.7	−0.127	0.90

language which is produced in a task. The less easily the consensus is achieved in a decision-making task, the more participants have to engage in subtler dialogue and the more extending is the language that is likely to be used.

Transformations of task material

The one task which required material to be transformed on-line was the narrative from Study 1. Skehan (1998) predicts that transformations will be associated with greater complexity, as learners have to wrestle with the need to bring the elements of the task into some sort of meaningful (and non-given) relationship with one another.

When one examines the results from this study, the within-subjects one-way analysis of variance for the complexity scores suggests that there are significant differences, but that the specific contrasts are between the personal task, on the one hand, and both the narrative and decision-making tasks on the other ($F = 11.3$; $p < 0.001$), with associated mean scores: 1.16 (personal), 1.32 (decision-making) and 1.35 (narrative). This only provides partial support for the hypothesis. However, a more supportive picture emerges if one examines the results when the mean scores for the different planning conditions are examined in more detail. These are shown in Table 8.6.

Table 8.6 Complexity measures for tasks requiring transformation and tasks not requiring transformation under different planning conditions

	Narrative	Personal	Decision-making
Unplanned	1.22	1.11	1.23
Undetailed planners	1.42	1.16	1.35
Detailed planners	1.68	1.26	1.52

Comparisons at each of the different levels of planning do not reach statistical significance, but the power of the comparisons is limited by the small sample sizes that result when such fine-grained comparisons are made. What is striking, however, is that the task requiring transformations always generates the highest level of complexity under planning conditions, and that *this advantage grows as the planning condition changes.* In other words, as the planning becomes more directed (Foster and Skehan, 1996, discuss this in terms of the task being interpreted as more challenging) there is an interaction with the complexity measure, such that the task requiring transformation

benefits most from this opportunity to plan. In other words, requiring learners to handle tasks requiring transformations immediately does not produce significantly greater complexity. When, however, planning time is given to enable them to respond to the *potential* complexity of the task, they are able to meet the challenge more effectively and the complexity of their language is greater.

Summary of the task results

It is easier now to try to summarise the results that have been obtained for each of the characteristics by representing the data in tabular form. The summary is shown in Table 8.7.

Table 8.7 Summary of the effects of task characteristics on complexity, accuracy and fluency

Task characteristic	Accuracy	Complexity	Fluency
Familiarity of information	No effect	No effect	Slightly greater
Dialogic vs. monologic tasks	Greater	Slightly greater	Lower
Degree of structure	No effect	No effect	Greater
Complexity of outcome	No effect	Greater	No effect
Transformations	No effect	Planned condition generates greater complexity	No effect

LIMITATIONS OF THE META-ANALYSIS

The existence of the six related datasets has enabled analyses to be performed which have the advantage of linking a range of different variables. The generalisations which are then possible can be more wide-ranging in their applicability. There are, however, serious limitations to this approach. The meta-analysis has had an inevitable opportunistic quality. The six datasets in question are related, since they derive from a common research framework, but they were not *designed* to ensure principled and systematic comparisons between the range of variables involved. Where it is possible to make broader-based but still clear comparisons (e.g. the large samples underlying the *complex outcomes* comparison), the conclusions made can have some force. On occasions, the comparisons have a rather tentative character. For example, the structured tasks are 'personal' and 'narrative', and these from different studies. These tasks then enter into contrasts with a whole range of 'non-structured' tasks. Clearly, the designation 'structured' was not by original design, but through post-hoc analysis. This must inevitably limit the force of the claims which are made. On many other occasions, the variables under investigation can only be partially disentangled. For example, one of the

predictions concerned accuracy. Structured tasks were predicted to generate greater accuracy. Hence the (structured) narrative in Study 2 would be predicted to be more accurately done. But it was also proposed that dialogic tasks (which narratives clearly are not) are also associated with greater accuracy. Hence, when the (structured) narrative in Study 2 was compared with the decision-making task from Study 2 (which may have been unstructured, but was dialogic) it was not possible to make an absolutely clear comparison because of the confound of variables. Other examples of this occur in the data, and clearly would suggest that the insights obtained so far should feed into the design of a more systematic study in the future. Still, the data that exist are all that can be analysed. Provided that the limitations of the dataset are understood, it is possible to draw the sort of tentative conclusions that have been proposed here, and extract some value from them.

IMPLICATIONS FOR TESTING

It is useful now to relate the findings shown in Table 8.7 to the model of oral language assessment presented in Figure 8.1. It was argued earlier that while model components from Figure 8.1, such as underlying competences and rater effects, have benefited from relevant empirical work, components such as interaction conditions, ability for use, and the role of tasks have not. It has been the goal of the present chapter to explore how this situation may be redressed by reviewing the contributions that can be made by a particular set of research studies into tasks.

Recalling that task fulfils an important mediating function which shapes the nature of the performance which will be the basis for the ultimate rating of the candidate score, we can see that the task itself is hardly a constant in this equation. The five task characteristics which have been explored show that systematic (and potentially uncontrolled and undesirable) effects are likely to follow from any task selection decision. In other words, there may be significant consequences when one task is chosen rather than another. Or to spell this out even more directly, if candidate performances are compared after having been elicited through the use of different tasks, the performances themselves may be very difficult to relate to one another. Different candidates, in other words, might be disadvantaged or advantaged by the particular task that they might have taken as part of their test, and so their performance may not be directly comparable to the other candidates.

Take, for example, the case of one candidate who was required to do a dialogic task compared to another candidate who had a narrative-based test. The above results suggest that the first candidate may well have been predisposed to achieve higher levels of accuracy and complexity than would otherwise have been the case, but lower fluency. The situation for the candidate taking a narrative-based test is exactly the reverse. The scores assigned these two candidates might then vary spuriously, even if the candidates were of a similar ability level. Public examination bodies are often attracted by narrative

formats to assess spoken language since they seem to contain useful standardisation potential: the present conclusions suggest that such an approach might inadvertently introduce another set of dangers.

In slight contrast, consider a situation where one candidate took a test containing clear sequential macrostructure, and another took a test in which transformations of input material were required. Assume further that in both cases there was some time for planning. The above research-based generalisations would lead us to expect an advantage in the first case for greater fluency and in the second, an advantage for complexity. If we relate these outcomes in performance to the rating scales which are used and/or the predispositions of the raters, we can see that there is even further scope for arbitrary score decisions. These may derive from the particular aspects of performance the rating scales and raters prioritise in importance, linked to the tasks which the candidates were required to do. The potential for inaccuracy is therefore magnified.

In short, to require spoken performances – which will be the basis for scoring – to be based on tasks which vary in the sort of language that they favour may well introduce error into spoken language assessment. Unless we are able to combat this through research-based studies which inform test design decisions, we are likely to treat candidates unfairly. There is therefore a strong need for research programmes which explore just how the range of factors which impact upon the scores assigned in spoken language tests operate in systematic ways. Unless this is done, incorrect decisions are likely to be made.

NOTES

1. The construct of complexity is close to what testers mean by range, in that both focus on a willingness to use a greater variety of syntactic forms.
2. They figure in other chapters in this volume, e.g. Bygate, Foster. In this section the focus is away from acquisition and towards measurement issues themselves.
3. The task required simultaneous retelling of a video-based narrative. For this reason, since the speed of the video tape influenced the performance, it was decided that measures of pausing-based fluency were inappropriate. Hence the use of alternative measures.

REFERENCES

Alderson, J.C. (1991) Bands and scores. In J.C. Alderson and B. North (eds) *Language Testing in the 1990s* (pp. 71–86). Modern English Publications and the British Council.

Bachman, L. (1990) *Fundamental Considerations in Language Testing.* Oxford: OUP.

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

Canale, M. and Swain, M. (1980) Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics*, 1 (1): 1–47.

Foster, P. and Skehan, P. (1996) The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18 (3): 299–324.
Foster, P. and Skehan, P. (1999) The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research*, 3, 3: 185–214.
Foster, P. and Skehan, P. (ms) Modifying the task: The effects of surprise, time and planning type on task based performance.
Foster, P., Tonkyn, A. and Wigglesworth, G. (in press) Measuring spoken language: A unit for all reasons. *Applied Linguistics.*
Fulcher, G. (1996a) Does thick description lead to smart tests? A data-based approach to rating scale construction. *Language Testing*, 13 (2): 208–40.
Harley, B., Allen, J.P.B., Cummins, J. and Swain, M. (1990) *The Development of Second Language Proficiency.* Cambridge: CUP.
Kenyon, D. (1992) Introductory remarks at symposium on development and use of rating scales in language testing. 14th Language Testing Research Colloquium, Vancouver, 27 February to 1 March.
Lumley, T. and McNamara, T. (1995) Rater characteristics and rater bias: Implications for training. *Language Testing*, 12: 55–71.
McNamara, T. (1995) Modelling performance: Opening Pandora's box. *Applied Linguistics*, 16 (2): 159–79.
North, B. (1996) *The development of a common framework scale of language proficiency based on a theory of measurement.* Unpublished PhD thesis, Thames Valley University.
Skehan, P. (1995) Analysability, accessibility, and ability for use. In G. Cook and B. Seidlhofer (eds) *Principle and Practice in Applied Linguistics.* Oxford: OUP.
Skehan, P. (1998) *A Cognitive Approach to Language Learning.* Oxford: OUP.
Skehan, P. and 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.
Skehan, P. and Foster, P. (1999) The influence of task structure and processing conditions on narrative retellings. *Language Learning*, 49 (1): 93–120.
Skehan, P. and Foster, P. (ms) *Using post task analytic activities to promote accuracy.* Manuscript, Thames Valley University.
Van Lier, L. (1989) Reeling, writhing, drawling, stretching, and fainting in coils: Oral proficiency interviews as conversation. *TESOL Quarterly*, 23: 489–508.
Wigglesworth, G. (1997) An investigation of planning time and proficiency level on oral test discourse. *Language Testing*, 14 (1): 85–106.
Willis, D. (1991) *The Lexical Syllabus.* London: Collins.
Wolfe-Quintero, K., Inagaki, S. and Kim, H.-K. (1998) *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity.* Technical Report No. 17, Second Language Teaching and Curriculum Center: University of Hawai'i.

Further reading

Bachman, L. and Palmer, A. (1996) *Language Testing in Practice.* Oxford: OUP.
Berry, V. (in preparation) *An investigation into how individual differences in personality affect the complexity of language test tasks.* PhD dissertation, Thames Valley University.
Candlin, C. (1987) Towards task based language learning. In C. Candlin and D. Murphy (eds) *Language Learning Tasks.* Englewood Cliffs, NJ: Prentice Hall.
Chalhoub-Deville, M. (in press) Task-based assessment: A link to second language instruction.

Fulcher, G. (1996b) Testing tasks: Issues in task design and the group oral. *Language Testing*, 13 (1): 23–52.

Long, M. and Crookes, G. (1991) Three approaches to task-based syllabus design. *TESOL Quarterly*, 26 (1): 27–55.

Lowe, P. Jr (1982) *ILR Handbook on Oral Interview Testing*, Washington, DC: DLI/LS Oral Interview Project.

Lumley, T. and Brown, A. (1996) Specific purpose language performance tests: task and interaction. In G. Wigglesworth and C. Elder (eds) *The Language Testing Cycle: from Inception to Washback* (pp. 105–36). *Australian Review of Applied Linguistics*, Series 8, No. 13.

Mosenthal, P.B. (1998) Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal*, 35 (2): 269–307.

Norris, J., Brown, J.D., Hudson, T. and Yoshioka, J. (1998) *Designing Second Language Performance Assessments*. Technical Report No. 18, Second Language Teaching and Curriculum Center: University of Hawaii.

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

Robinson, P. (1995) Task complexity and second language narrative discourse. *Language Learning*, 45 (1): 99–140.

Shohamy, E. (1994) The validity of direct vs semi-direct oral tests. *Language Testing*, 11: 99–124.

Skehan, P. (1996) A framework for the implementation of task based instruction. *Applied Linguistics*, 17 (1): 38–62.

Skehan, P. (ms-a) *Assessing ability for use*. Manuscript, Thames Valley University.

Skehan, P. (ms-b) *Task characteristics, fluency, and oral performance testing*. Manuscript, Thames Valley University.