JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.1 (41-114)

Chapter 9

Strategic planning, task structure, and performance testing*

Parvaneh Tavakoli and Peter Skehan
Kings College, London / Chinese University of Hong Kong

Introduction

Pollitt (1990) makes a distinction between approaches to language testing which *count*, e.g. a multiple choice reading test, and those which *judge*, e.g. with rating scales. A "judging" approach to assessment is a form of performance testing. Within this area, there has been significant interest in recent years in the use of tasks to underpin the assessment decisions that are made. There are several reasons for this. First, task-based research has been an active area within second language acquisition, and as a result, there are a range of research findings which could, potentially, be related to language testing contexts. Second, there are the beginnings of models of task-based performance which could be the basis for applications to assessment. Above all, though, researchers in this neighbouring area have attempted to explore how tasks can be described and analysed, and their properties linked to the nature of the performance that is elicited. To put this another way, researchers generally argue that tasks are not neutral devices which elicit performance in a straightforward manner – they influence the nature of the performance which results. As a result, insights from this closely related area may clarify how performance-testing procedures may need to take account of the systematic influences, as well as level of difficulty, that tasks provide. For these reasons, we will next review some of the recent task research.

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.2 (114-161)

Table 1. Task influence on performance dimensions

Task characteristic	Accuracy	Complexity	Fluency
familiarity of information	greater	no effect	slightly greater
dialogic vs. monologic tasks	greater	greater	lower
degree of structure	greater	no effect	greater
complexity of outcome	no effect	greater	no effect
transformations	no effect	planned condition generates greater complexity	no effect

Modelling and researching task-based performance

Two approaches will be outlined here, following work by Skehan and Robinson. Both approaches are of a generally cognitive orientation, although they differ in the claims they make about the systematic influences of tasks upon performance. Following extensive work within cognitive psychology (see, e.g., Miyake & Shah 1999), Skehan (1996a, 1998b) makes the assumption that human beings operate with limited capacity attentional systems, and that to pay attention to one area of performance may well be to reduce the attention available elsewhere. In other words, if performance is multi-dimensional, improving performance in one area may well cause achievement in other areas to be lowered. In a series of studies, Skehan and Foster and their collaborators have demonstrated that a number of task characteristics have systematic influences upon performance. First of all, regarding performance itself, they have shown (Skehan & Foster 1997) that it is useful to explore the complexity of language, its accuracy, and its fluency, and that these three areas enter into competition with one another for scarce attentional resources. Second, the range of evidence from several studies is consistent with different influences upon each of these performance areas. This is shown in the following table (taken from Skehan (2001)).

The claim here is a that a set of task characteristics have predictable influences upon performance. For example, if a task draws upon familiar information, then, other things being equal, it is likely to yield a performance which is more accurate and more fluent, but without any particular impact upon the complexity of the language which is used. If one takes into account that raters

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.3 (161-219)

of performance will be attending to areas such as accuracy and fluency, then the use of familiar information tasks is likely to give such raters reasons to grade more positively. If all tasks were based on such information, this would not be an issue, but if tasks varied, without control, such that different test takers were rated on tasks which varied in the familiarity of information they were based on, the differences in the ratings given could be partly artifactual. The same issue applies to the other task characteristics that influence performance.

A more problematic area within this research has been to explore task characteristics which impact upon the general construct of task difficulty. Skehan (1998a), for example, proposes that a number of features impact upon difficulty, including:

– number of elements in the task (more means more difficult);
– type of information (concrete (less difficult) vs. abstract) (Brown et al. 1984).

But there are difficulties with difficulty, and these difficulties become more salient when task-based performance testing is being considered. A concept such as difficulty implies an overarching dimension into which other features fit. But we have seen that task performance is multidimensional, so that devoting attention to one area may be at the expense of others. In this case, the problem is to decide on an *overall* level of difficulty for a task. Would it be the general complexity of language achieved? Or the accuracy? These difficulties currently seem to have no obvious solution for testers, and so using tasks to assess the capacity to speak contrasts with the degree of success that has been achieved in predicting difficulty in other areas, e.g. reading (Kirsch & Mosenthaler 1990).

Finally, in this review of task research, it is clear that the conditions under which a task is done can exert a strong influence upon performance. The bulk of the research here has been into the effects of pre-task planning (strategic or rehearsal). It is now clear that giving learners (or test-takers) such planning time leads to significant improvements in performance. The clearest generalisation is that pre-task planning is associated with greater complexity and fluency, in almost all studies. There is a less robust relationship with accuracy, with some studies supportive of an accuracy effect, and others much less so. Further work is needed to uncover which particular conditions in planning lead to greater accuracy.

As a result of this range of studies, Skehan (1998a, 2001) proposes a model of task-based performance in relation to language testing (see Figure 1).

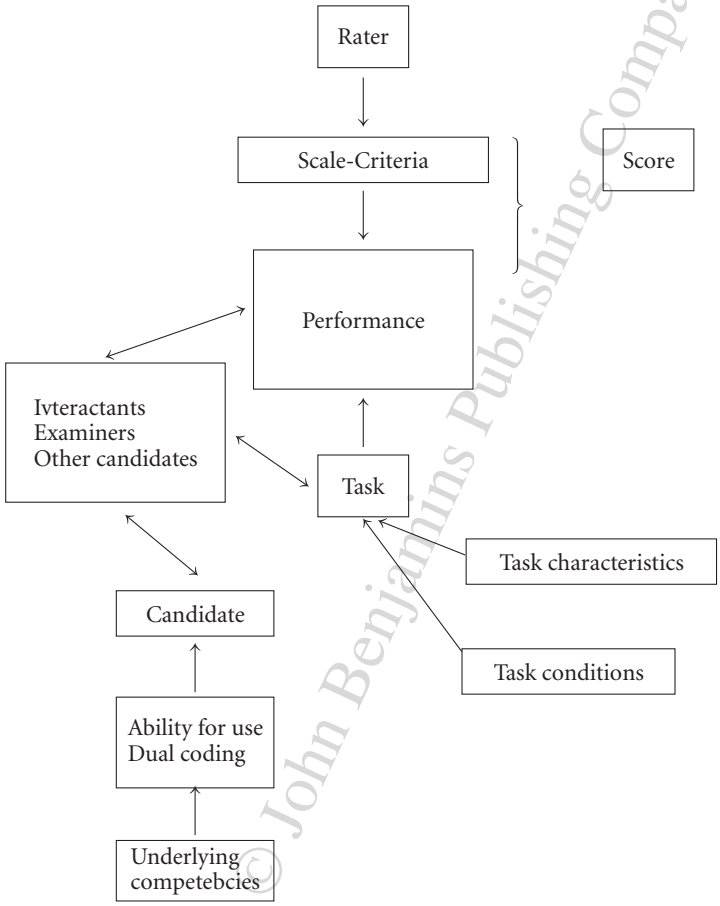

Figure 1. Task based performance and language testing

The broadest purpose of this model is to make clear that the rating assigned someone on the basis of their performance on a task is the consequence of a whole range of factors, only one of which can possibly be their underlying competence. In addition, we have to consider:

– the method by which the rating is done, with the potential this has to introduce error;
– the context for the performance, including the nature of the interactants involved, and their relationship to one another;
– the extent to which the testee can engage strategies of performance, and general processing skills, handling rule-based and memory-based language;

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.5 (233-289)

– the task that is involved and the conditions under which it is done.

Essentially, these different factors, besides helping us to understand how test scores may be the result of a multiplicity of influences, also provide an agenda for research. With regard to the study to be reported below, this means that we need to advance our understanding of the influence of task characteristics on performance, as well as what impact the conditions under which tasks are done might have on that same performance.

There is, though, an alternative perspective on the use of tasks, and of their potential for assessment, which is represented in the work of Peter Robinson, (see Chapter 1). Three main differences are apparent. First, Robinson (2001c) does not assume that attention is limited in capacity, preferring instead to view attention as an expandable resource, particularly in relation to memory in different, non-competing modality areas. Second, Robinson sees tasks as driving performance in a slightly different way. For him, language complexity follows functional complexity (Givon 1985). In other words, he predicts that more difficult tasks will push learners to engage more complex language *and* push them to achieve accuracy so that communication is more effective. In other words, he is proposing that learners will "rise to the challenge" of completing more difficult tasks, and perform better. Third, Robinson distinguishes between what he terms resource directing and resource depleting factors. The former are the influences which push learners to engage in more difficult language and are therefore what task designers need to attend to if they are to influence learner (and test-taker) performance. The latter, resource depleting factors, such as the opportunity for pre-task planning time, are those which influence the overall difficulty of a task.

The relevance of task-based studies within an SLA performance framework has certainly been noticed within language testing. Iwashita et al. (2001) used task features identified by Skehan (1998b, 2001) and Robinson (2001c) and explored the relevance of these features for test-task performance. They focussed on a narrative task, on the basis that this task-type is the most typical of testing formats. This is because it is non-interactive, and so potentially open to greater control, and therefore standardisation, on the part of the test producer. In particular, they explored the effects of:

– *perspective*: whether a narrative story is told from the teller's point of view, or from someone else's point of view;
– *immediacy*: whether the test taker told the story with the pictures in front of them, or without the pictures present;

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.6 (289-336)

– *adequacy*: whether test takers told the story with access to all the pictures in a picture series, or with two of the pictures (out of six) missing;
– *strategic planning time*: whether there was three minutes planning time or not.

The results obtained did not follow the predictions arising from the task literature. Only one of the twenty-four comparisons yielded a significant result. This was for the immediacy condition, with accuracy scores. All other comparisons were non-significant. Worse, the one significant result was *against* the predicted direction, with the more difficult non-immediate conditions producing greater accuracy.[1] Clearly, these results are disappointing from a task research perspective, and lead to three alternative interpretations. First, it may be that the narrowness of the range of tasks used (all picture-prompt narratives), and the constraints of testing conditions lead to the failure to find any significant differences. Second, it is possible that the variables which were manipulated were not the most salient ones from the task literature. Third, even if the variables manipulated were well chosen, it is possible that their operationalisations were not ideal. But these three interpretations are all post-hoc attempts at rationalising what one would have expected to have been a series of significant results. Obviously, there is a fourth interpretation – that task manipulations are relevant to pedagogy and certain types of performance, but not to the context of testing. The study by Iwashita et al. (2001) presents a challenge to those who think that the task-based literature can make contributions to assessment.

Task structure and task performance

Further work is clearly required to help clarify the precise ways in which task research may be relevant for assessment and the ways it may not. In Table 1, five task characteristics were proposed which may impact upon performance: information familiarity, dialogic vs. monologic tasks; degree of structure; complexity of outcome; and number of transformations. The first of these was included by Iwashita et al. (2001) and shown to have only a marginal influence on performance. The second could be regarded as difficult to implement in testing, to the extent that there is a need to have standardised conditions, of a sort that would be compromised with multiple participants in a task. It could also be argued that Iwashita et al. (2001) indirectly explored the feature, 'transformations', when they examined 'adequacy', in that missing pictures in

the task they used required test takers to transform the input material. The remaining two features would seem worth investigating. In the present study, the focus will be on the first of these, degree of structure, principally because this task feature is more readily manipulable with narrative tasks.

In two early studies, Foster and Skehan (1996) and Skehan and Foster (1997) explored three tasks: a personal information exchange task, a narrative, and a decision-making task. Predictions were made about the effect of task type on task performance, and these predictions were partially borne out. However, there were aspects of performance which did not follow the predictions. A *post-hoc* interpretation of these studies suggested that the tasks which contained a clear macrostructure seemed to advantage fluency and accuracy, but left complexity unaffected.

Given the *post-hoc* nature of these interpretations, a subsequent study was designed with the intention of comparing tasks containing structure with those which did not. Participants were required to retell narratives based on video prompts. The prompts were, in fact, Mr. Bean episodes. One of these "Crazy Golf", involved Mr. Bean playing a round of crazy golf. Over-interpreting the golf attendant's instruction that he should, on no account, touch the ball while playing a hole, Mr. Bean ended up knocking the golf ball outside the golf range which led to a series of unlinked events. There was no structure to these events, and the way one event moved to the next was fairly unpredictable. The second video prompt was "Mr. Bean goes to the restaurant", reflecting Mr. Bean treating himself to a meal on his birthday. Although a number of things happened which do not normally happen in restaurants, e.g. Mr. Bean hiding a disgusting meal in the pockets of fellow diners, the general structure of the narrative simply followed the conventional restaurant "script": i.e. greeting by maitre d'hotel, being seated, being shown the menu etc. Results were supportive of the predictions. The structured narrative (Restaurant) produced by the participants was considerably more fluent than the unstructured narrative, and, when there was strategic planning, it was also more accurate.

The results of this study are encouraging for our understanding of the role of task structure on performance. However, they do raise some additional questions. Clearest of these is that it is important to make progress in understanding what task structure is, and how it can be characterised. Foster and Skehan (1996) and Skehan and Foster (1997) represented task structure as consisting of a clear time line, so that "structure" was introduced into the task as the clarity of the macrostructure involved when a series of events unfolded in time. It was assumed that the second language speakers clearly apprehended an overall macrostructure and so were aware of this macrostructure in general while con-

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.8 (391-440)

centrating on telling particular parts of the story. To portray this in relation to Levelt's model of speech production (Levelt 1989), it is hypothesised that the Conceptualiser component of the storytelling is relatively unpressured, with the result that attention can be more easily allocated to the Formulator, and this will impact upon fluency and accuracy. In the Mr. Bean restaurant task, similarly, it was assumed that general knowledge of a restaurant "script" would lead learners to have attention available to achieve greater accuracy and fluency with the details of the story to be told.

The difficulty here is that while one may have a general notion of a macrostructure in the speaker's mind, how that macrostructure is justified or described is not so clear. Already we have:

- a clear time line
- a script
- a story with a conventional beginning, middle and end
- an appeal to what is familiar and organised in the speaker's mind

with all of these functioning to create a macrostructure which will then impact upon performance. Essentially, this raises the possibility that macrostructures can be created from a number of sources, and that we therefore need to make progress in understanding what the possible sources are.

In fact, the four characteristics just mentioned do not exhaust the different ways in which a task can be regarded as structured. Other work in testing may be instructive. Kobayashi (1995, 2002), following work by Winter (1976) and Hoey (1983), demonstrated the relevance of a problem-solution structure, i.e. a narrative sequence whose centre is a problem which is resolved. Optionally this structure can have the more extended sequence *Situation* (which sets the scene and introduces relevant information) – *Problem-Solution* and *Evaluation*, (which functions as a sort of commentary on the satisfactoriness (or otherwise) of the proposed solution). Kobayashi (1995, 2002) showed that reading comprehension texts based on a problem-solution structure produced different results to those which were not structured in this way. The structured texts distinguished more clearly amongst *more* proficient students, especially when comprehension was measured by summary and open-ended questions, rather than a cloze test format.

It follows that a clear macrostructure can be achieved by ensuring logical relations between the elements of the story instead of relying on a clear and simple story line or on some sequence of events familiar to the learners. In a sense therefore there may be a greater degree of universality to the structure involved, since it will be based on causal connections. Such a set of links

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.9 (440-487)

might be regarded as having a "tighter" organising frame than other sorts of macrostructure.

This analysis suggests an interesting possibility: that within the general concept of macrostructure, there are different varieties of macrostructure available, and that these different varieties may be arrangeable in the form of a cline, reflecting the *degree* of structure. This would lead to the prediction that if task structure leads to greater fluency and accuracy, then greater task structure would lead to particularly high fluency and accuracy. This would be interesting from a pedagogic or a testing perspective. It could also be the basis for research, since it would generate measurable comparisons between tasks at different levels of structure.

Another area which impacts upon task and test performance is that of strategic planning. A number of studies have shown clear effects in this area, effects which are reviewed in the chapters of this book. In general, the results have shown that giving learners planning time has a favourable effect upon performance, with consistent and appreciable increases in complexity and fluency (Skehan & Foster 2001) and less consistent, and smaller increases for accuracy. The fluency effects do seem the clearest and most consistent. With complexity (which connects with the concept of *range* in the testing literature) most studies report gains, but there are some contradictions. Wigglesworth (2001), for example, working in a testing context, does not find the usual effects here, raising the possibility that the prominence of an assessment framework may cause differences in the results which are found. Regarding accuracy, a series of studies have shown significant effects, such as Foster and Skehan (1996, 1999), Mehnert (1998) and Skehan and Foster (1997, 1999). But other studies have not reported similar outcomes (Crookes 1989; Ortega 1999; Wigglesworth 2001). In this latter case, there were slightly different patterns of results at two different proficiency levels, suggesting that the effects of the independent variables may interact with proficiency level. It is also possible that there is an interaction between strategic planning, performance area and task type, since the tasks used by these different investigators have varied. There may be a tendency, in fact, for narrative tasks, the very tasks most favoured in testing contexts, to be least likely to deliver a significant accuracy effect. Although there are studies using narrative tasks in which accuracy effects are found (Ellis 1987; Yuan & Ellis 2003), these only use narrative tasks, and so a comparative task dimension is lacking. Where studies use more than one task type, e.g. Foster and Skehan (1996), Skehan and Foster (1997), the narrative task is associated with smaller accuracy effects than, for example, decision-making and personal information

exchange tasks. In any case, it is clear that there is further scope to investigate the effects of strategic planning in this particular performance domain.

Drawing on this review of the literature, with tasks and with planning, a number of hypotheses can be formulated:

Hypothesis One: degree of structure in a narrative task will influence the fluency and accuracy of performance on the task. Further, progressively greater structure, in the sequence no structure>schematic structure>weak causation structure>problem solution structure will lead to greater increases in fluency and accuracy.

Hypothesis Two: there will be no influence of task structure on complexity.

Hypothesis Three: strategic planning will influence complexity, fluency and accuracy.

Hypothesis Four: the effects of these variables will not vary as a function of proficiency level.

Method

Design

A $2 \times 2 \times 4$ factorial design was used in the current study with pre-task planning condition, proficiency level, and task structure as the independent variables. Planning condition and language proficiency were between-participants variables and each had two levels with the participants belonging to either of the two conditions and levels. Task structure, which was operationalised through 4 different picture series, had four levels representing a scale in the degree of structure of a task. Task structure was a within-participant variable, i.e. all participants performed all four tasks.

Task

As discussed earlier, narrative tasks are frequently employed in the context of assessing language performance (Elder et al. 2002; Iwashita et al. 2001; Robinson 2001c). Narrative tasks are also routinely used as a single type of stimulus by some international testing organisations (e.g.: Test of Spoken English). Narrative tasks in this sense refer to stories based on a sequenced set of picture prompts, which are given to participants in order to elicit language perfor-

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.11 (539-596)

mance. The rationale for using narratives is justified in terms of construct validity, reliability and authenticity of the test. However, the prime reason for selecting narratives in the present study is to have conformity with the literature from which the theoretical assumptions of the study are drawn.

In order to find appropriate picture series, two main sources were consulted:

1) EFL sources including course books and supplementary materials for teaching English and other modern languages; and 2) non-EFL sources including a wide range of different materials such as cartoon books, newspapers and pictorial stories. The aim was to find picture series which were clear, had worthwhile stories to be told, were of a length suitable for the study, were culturally familiar to the participants, were neither linguistically cued nor linguistically demanding, and looked interesting. The picture series identified in this way were then carefully examined and categorised into structured and unstructured. Structured tasks were defined as having either a problem-solution structure or a schematic sequential structure. The unstructured picture series were further distinguished between completely unstructured series and those containing minimal structure. Lack of structure was operationalised in terms of the number of pictures, other than the first and the last, that could be interchanged with one another without the story being compromised.

To achieve the purpose of the study regarding degrees of structure, two structured tasks and two unstructured tasks were selected. From among the structured-tasks category, one picture series was selected to represent the problem-solution structure and one picture series was selected to represent the a weak causation structure. Based on the theoretical assumptions of structure discussed earlier, the two structured tasks differed from one another in terms of the type of structure they exposed and the degree of structure they presented. In fact, following Kobayashi (1995, 2002), the problem-solution structure was assumed to have a stronger type of structure than the weak causation structure because the solution involved has a greater unifying or "resolving" effect on the overall story macrostructure. The task selected in the problem-solution category, i.e. the Football task, was a picture series with a transparent problem-solution structure and a well-presented sequential organisation. The second structured task, the Picnic task, on the other hand, was based on a clear organisation and contained an implicitly stated problem, which was only revealed in its last frame. However, this task did not propose a clear transparent problem-solution structure, which made it less structured than the Football task. Therefore, both tasks were structured but differed in the degree of structure they demonstrated.

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.12 (596-671)

Similarly, two tasks with varying degrees of structure were selected from the unstructured group. The lack of a problem-solution relationship on the one hand and lack of causative elements on the other hand suggested that both tasks were unstructured. However, they differed from one another with regard to the amount of sequential organisation they contained. Task four, the Walkman task, did not contain any sequential organisation and, therefore, was less structured than task three, the Unlucky Man, which had a loosely presented sequential organisation. In effect, events in task four were loosely related to one another and the sequence of organisation of events hardly followed a timeline. The four picture sets can be found in Appendix 1.

All the picture sets consisted of six pictures, except the Unlucky Man task, which due to the assumptions of the study, had a set of moveable pictures in the middle and, therefore, had ten pictures. Figure 2 indicates how the four tasks can be located on a continuum representing a scale of the degree of structure hypothesized in the study reported here:

+Structured **–Structured**

1. Football 2. Picnic 3. Unlucky Man 4. Walkman

Figure 2. Degrees of structure in the four tasks

In order to avoid a practice effect, a counterbalanced design was used. In effect, participants performed all four tasks but in four different sequences. Table 2 demonstrates the four sequences of tasks in the study.

Table 2. Counterbalanced sequence of the tasks

Sequence 1	Football	Picnic	Unlucky Man	Walkman
Sequence 2	Picnic	Unlucky Man	Walkman	Football
Sequence 3	Unlucky Man	Walkman	Football	Picnic
Sequence 4	Walkman	Football	Picnic	Unlucky Man

Strategic Planning Conditions: As planning was a between-participant variable, half of the participants performed the tasks under planned and half under the unplanned conditions. Planning was operationalised in terms of the amount of strategic planning time provided to the participants. As discussed earlier, the amount of time given to the participants was determined on the

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.13 (671-726)

basis of the findings of previous studies (Elder et al. 2002; Mehnert 1998; Wigglesworth 2001). The unplanned group was given 30 seconds to look at each of the picture series before they started telling the stories. Participants in the planned groups were given 5 minutes to look at each of the picture series and were advised to plan for what they were going to say. Moreover, under the planned conditions each participant was given a sheet of paper to take notes. However, they were informed that they would not be allowed to use their notes while they were telling the story. The instructions given to both groups were identical in all other respects.

Language proficiency level: The participants were drawn from two levels of language proficiency, i.e. elementary and intermediate. Prior to the study the candidates were placed in to their levels on the basis of an institutional placement test. Nevertheless, to confirm the homogeneity of the groups and also the distinction between the two proficiency levels, their language proficiency was tested by the "Oxford Placement Test 2" (Allan 1992). It should be noted that due to the practical limitations, only the grammar part of the test was run. Participants' responses were checked and scored on a scale of 100 points. The elementary group had a range of scores between 17 to 44 and the intermediate group scored between 45 and 75. These grammar-test results confirm the decisions made on the basis of the wider-ranging institutional placement test.

Perceptions of task difficulty: In order to explore participant perceptions of task difficulty, separate (but related) questionnaires were developed for each of the planning conditions. All participants were asked to complete the appropriate questionnaire as soon as they performed the four tasks. Both questionnaires contained questions about participant perceptions of task difficulty as well as an open-ended question in which participants could give suggestions and comments about the tasks. The planned-group questionnaire differed from the unplanned-group since it included an extra question about the usefulness of the strategic planning time for each of the four tasks. Regarding task difficulty, answers were given on a four-point scale with 1 representing "very easy tasks" and 4 "very difficult tasks". The answers to the extra question for the planned group were also given on a four-point scale with 1 indicating that the strategic planning time "helped very much" and 4 showing that the strategic planning time "did not help at all". To avoid any potential confusion or misinterpretations resulting from the participants' reading ability, the questionnaires were translated into the participants' first language.

Pilot study: In order to see whether the selected tasks were functioning in line with the theoretical assumptions of the study, the four tasks were piloted twice. In the first pilot study, three Farsi-speakers, one at elementary and two intermediate, performed the four tasks and completed the questionnaires, with this data revealing one of the tasks to be confusing to participants. This task was then replaced with another task from within the same category of structure. The new set of four tasks and the two planning conditions were then piloted on 14 elementary and intermediate learners in a college in London. The participants were aged between 18 and 24, and were from 3 different language backgrounds (Farsi, Chinese and Arabic). They were assigned to either a planned or an unplanned group and completed the tasks in a one-on-one setting with the researcher. They performed the four tasks in a counterbalanced design and completed the questionnaires afterwards. The results of the pilot study suggested that the selection of tasks and the amount of pre-task planning time were appropriate and practical.

Participants in the main study: Participants in the main study were 80 language learners studying English at an educational association in Tehran, Iran. They were all adult females aged between 18 and 45. They were studying English as a foreign language at an elementary or intermediate level and had been studying English at the same language school for at least 18 months. The participants were all Farsi speakers and had a similar language learning history both in the state school system and at the above-mentioned language school. But they differed regarding the period of time they had been studying English in the past, the contact they had with English outside classroom, and the purposes for which they were studying English.

As they had already taken part in similar testing situations in their language school and performed similar tasks, they were all familiar with both the testing conditions and the test format, i.e. narratives. One participant was withdrawn from the study and replaced as she expressed unwillingness during the test. The participants were randomly assigned to a planned or unplanned condition and one of the four sequences of the counterbalanced design, as demonstrated in Table 3.

Setting: All participants were tested in a one-to-one setting by the first author who met them individually and explained the purpose of the test to them. After each participant was randomly assigned to either the unplanned or planned conditions and to one of the four counterbalanced sequences of the four tasks, the instructions were given to them. The participants were given each of the

Table 3. Design of the study

Planning condition	Proficiency level	No. of participants in sequence 1	No. of participants in sequence 2	No. of participants in sequence 3	No. of participants in sequence 4
Planned	Low-proficiency	5	5	5	5
	High-proficiency	5	5	5	5
Unplanned	Low-proficiency	5	5	5	5
	High-proficiency	5	5	5	5

picture series, in turn, and asked to tell the story to the researcher in a way that she could understand what was happening in each story. Under the unplanned condition, they were told that they had just 30 seconds to look at the pictures before they started telling the story. For each task they were given 3 to 4 minutes to tell the story. After the initial 30 seconds, the participants had the picture series in hand, looked at them and told the story to the researcher who tape-recorded the participant's performance on the first task. Then, the same process was repeated for the second, third and fourth tasks, one after the other. Under the planned condition, the participants were told that they had 5 minutes to look at each picture series and plan what to say, and that they would eventually have 3 to 4 minutes to tell the story. Each of these time intervals was chosen to ensure comparability with task-based studies in the literature with an assessment focus (e.g. Wigglesworth 2001; Iwashita et al. 2001). They were also given some paper to take notes if they wished. They were reminded that they would not be allowed to use their notes while they were telling the stories. After the five minutes, the participants told the story to the researcher who tape-recorded the participant's performance on the first task. Then, the same process was repeated for the second, third and fourth tasks one after the other.

Based on the type of the planning condition, the participants were asked to complete the appropriate questionnaire. They were also encouraged to comment about the test and the tasks in general in the last question of the questionnaire. All the introductory talk and instructions to the participants were given in Farsi.

Analytic measures

The recorded performances of all 80 participants of the study were transcribed, word-processed, and then digitised. The following sections will provide a de-

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.16 (825-889)

tailed description of the dependent variables as well as how the speech samples were coded and analysed.

Fluency: Koponen and Riggenbach (2000) have discussed different aspects and representations of fluency in detail and argue that fluency may refer to smoothness of speech in terms of temporal, phonetic, and acoustic features; it may represent proficiency at a macro or micro level; it may mean the automaticity of psychological processes; or it may be expressed as a notion contrasting with the concept of accuracy. Freed (2000) proposes that fluency spans a continuum that ranges from studies of its psychological manifestations and reflections of underlying speech-planning and thinking processes to studies of speech production, hesitation phenomena, and the temporal dimensions of speech.

Based on this multifaceted nature of fluency, different researchers have adopted various measures to assess fluency. These measures, however, can be categorised into some sub-dimensions. The first sub-dimension of fluency is silence, or as Skehan (2003) puts it, breakdown fluency. Length and number of unfilled pauses, filled pauses and total amount of silence are some of the measures researchers have used to assess this aspect of fluency. There is, though, some disagreement regarding the minimum length for a pause to be counted as a pause, with proposals as low as .25 of a second (Kormos & Denes 2004). Freed (2000), in a study aimed at exploring the construct of fluency in the speech of L2 learners of French, investigated fluency in terms of 7 measures including unfilled pauses. Regarding the unfilled pauses, she measured the disfluent-sounding silences occurred at places other than predictable juncture boundaries which tended to be of .4 a second or longer in duration. She argues that:

> Since silent pauses of shorter duration, frequently termed micropauses and measured in milliseconds, are characteristic of native speech and accurately measured by computerized acoustic analysis, we chose to identify and measure only those unfilled pauses [.4 a second or larger] that were heard as dysfluent. (Freed 2000: 248)

We will follow Freed's proposals for minimum length of pause in the present research. A second sub-dimension of fluency deals with the speed with which language is produced. Measures of speech rate, articulation rate, amount of speech, time ratio and mean length of run are typical here. Speech rate and length of run are the two most commonly used measures in SLA studies. Mehnert (1998), Towell et al. (1996) and Freed (2000) have used mean length of run to measure fluency of the speech production. Mean length of run in Towell et

al. (1996) is calculated as the mean number of syllables produced in utterances between pauses of .28 seconds and above. Mehnert (1998) found mean length of run by calculating the mean number of the syllables between pauses of 1 second. Freed (2000) defines length of run as continuous streams of running speech (measured in words) not interrupted by disfluent pauses or hesitations. Therefore, mean length of run is a manifestation of how lengthy the language produced between two pause boundaries is. Speech rate, i.e. number of syllables or words on average per minute, is another measure frequently used by researchers as an index of fluency (Yuan & Ellis 2003; Mehnert 1998; Raupach 1980; Robinson 2001c). Freed (2000) measured speech rate on the basis of the number of "nonrepeated" words or semantic units per minute. Towell et al. (1996) have calculated speech rate by dividing the total number of syllables produced in a given speech sample by the amount of total time including the pauses. It can be concluded thus that speech rate refers to how fast and dense the produced language is in terms of the time units.

The third sub-dimension of fluency is what is known as repair fluency (Skehan 2003). Repair fluency includes reformulation, replacement, false starts and repetition of words or phrases. Wigglesworth (1997) measured the percentages of clauses containing self-repairs and reported that planned performance is significantly more fluent than unplanned. Skehan and Foster (1999) used repetitions, false starts, reformulations and replacements to measure fluency in their study of the effect of structure on narrative task performance. Freed (2000) operationalised repair fluency in terms of repetition of exact words, syllables or phrases, reformulations, false starts, corrections and partial repeats in the learner speech.

These various conceptualisations of the nature of fluency have rarely been investigated together in task-based studies. Hence, in order to have a more detailed and precise exploration of the nature of fluency and to know what effects different task characteristics would have on various aspects of fluency in task-based contexts, a wide range of different measures is used, i.e. the number of false starts, reformulations, replacements, repetitions, length of run, speech rate, number of pauses, mean length of pauses and total amount of silence.

Accuracy: With measures of accuracy, there is greater consensus among researchers. In a few studies accuracy has been measured by specific measures, such as past tense morphemes (Ellis 1987) and plural -s (Crookes 1989; Wigglesworth 1997; Ortega 1999). Some of these studies did not reveal any significant differences between different planning or task conditions, e.g. Crookes and Ortega. On the basis of these results, Skehan and Foster (1999) argued

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.18 (944-979)

that such specific measures are less sensitive to detecting differences between experimental conditions. As a result, they have used general measures of accuracy, such as the number of error free clauses divided by the total number of clauses (Foster and Skehan (1996) and Skehan and Foster (1999)). In both of these studies accuracy effects were detected when pre-task planning time was provided to learners and when task structure was present. Furthermore, interactions were found between task structure and pre-task planning time. In contrast, Ortega (1999) measured accuracy by means of targetlike use of analysis of two grammatical areas: morphology agreement of a noun and its modifiers (including possessives, adjectives and quantifiers), and use of the Spanish article system. She argues that general measures have the disadvantage of being too broad to capture small changes in targetlike use since they combine multiple error types and obscure errors that might be important at a given level of development. Interestingly, and with a degree of compromise, Mehnert (1998) used general measures (percentage of error-free clauses and the number of errors per 100 words) as well as more specific measures (word order and lexical choice). The results of her study showed that both of the general measures of accuracy generated significance, but that neither of the specific accuracy measures did. This result is consistent with the view that while general measures are more blunt instruments, they do capture more variance in performance, and as a result, are more sensitive to the detection of significant effects.

In the current study, accuracy was measured by an index of error-free clauses. Error-free clauses were defined as clauses in which no error was seen with regard to syntax, morphology, native-like lexical choice or word order. However, errors in stress, intonation patterns or pronunciation of the words and utterances were not included. The native-like use of the language, in terms of the grammar and lexis, was generally considered as a criterion in determining whether the clauses were error-free. All error-free clauses were then identified and coded in the transcribed data, and the ratio of the error-free clauses to the total number of clauses was calculated.

Complexity: Foster, Tonkyn and Wigglesworth (2000) have discussed the analysis of spoken data in detail and emphasised that such analysis requires a principled way of dividing the transcribed data into units in order to assess features such as accuracy and complexity. Identifying the shortcomings of measures like T-units and C-units, they have introduced the AS-Unit, (Analysis of Speech Unit). They provide a number of reasons to show that the AS-unit is more appropriate than units used by previous researchers. Foster et al. (2000) define the AS-Unit as "a single speaker's utterance consisting of *an independent*

clause, or sub-clausal unit, together with any *subordinate clause(s)* associated with either" (p. 365). In this definition, an independent clause will be minimally a clause including a finite verb. An independent sub-clausal unit will consist of either one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation. The definition of AS-unit also considers minor utterances which are one class of "Irregular sentences" or "Nonsentences" identified by Quirk, Greenbaum, Leech, and Svartvik (1985). Furthermore, Foster et al. (2000) explain that "a subordinate clause will consist minimally of a finite or non-finite verb element plus at least one other clause element (Subject, Object, Complement or Adverbial)" (p. 366).

Following Foster et al. (2000), the transcribed data was coded into AS units that contained independent clauses, subordinate clauses and sub-clausal units. The intonation and pausing patterns of speech had a direct influence on determining whether a clause was an independent clause or a dependent one. As a result, the complexity of the performance was measured through an index of subordination by dividing the number of clauses by the number of AS units.

Coding the data and inter-rater reliability. Once the data were coded, 10% of the data was coded by an independent expert against which the data coded by the first author was tested. The inter-rater reliablity coefficients were all above 0.90 for the codings of the complexity measures, i.e. the AS units and the dependent clauses, as well as repetitions and replacements. However, the reliability coefficient for measures of accuracy, false starts and reformulations were initially lower. As a result, there was a reassessment of the measures of accuracy, false starts and reformulations until inter-rater correlations of above .94 were achieved.

Results

In order to test the hypotheses of the study, a number of different statistical analyses were carried out. Task structure, pre-task planning time and proficiency level were the independent variables, while 12 measures of fluency, accuracy and complexity were the independent variables. Factor analyses were run to investigate whether the dependent variables truly represented distinct factors. Based on the results of the factor analysis, a repeated measures MANOVA was performed to test the overall effect of the independent variables on the dependent variables. Finally a series of ANOVAs and t-tests were run to exam-

ine the differences among tasks, between the planning conditions and language proficiency levels.

Underlying factors in language performance

Separate factor analyses were run for each of the four tasks with the 12 measures of fluency, accuracy and complexity. Prior to the analysis, the suitability of the data for factor analysis was investigated. Inspection of the correlation matrixes revealed the presence of many coefficients of .4 and above. The Kaiser-Meyer-Oklin values were above .68 for all the tasks, exceeding the recommended value of .60 (Kaiser 1974). Barlett's Test of Sphericity reached statistical significance, supporting the factorability of each correlation matrix. The factor structure obtained was remarkably similar for all four tasks, and so only the results for the Football task only are shown here, as Table 4, to save space.

As Table 4 shows, factor 1 is made up of six high loadings on measures of length of run, speech rate, total amount of silence, total time spent speaking, number of pauses and length of pauses. These measures refer to different temporal aspects of fluency and suggest one relatively unified dimension. This hypothetically means that the more fluent participants would be expected to use a significantly greater length of run, a faster speech rate, less silence, fewer pauses, shorter pauses as well as more time spent speaking in their performance. The result of the factor analysis in Mehnert's study (1998) supports the same loadings for speech rate, length of run and total amount of silence.

The second factor loading is based on reformulations, false starts, replacements and repetitions, measures associated with another aspect of fluency, i.e. repair fluency (Freed 2000; Skehan 2001, 2003). The loadings for reformulation and false starts define the factor most clearly and replacement and repetition follow with lower, yet significant loadings. (This pattern was identical across all four tasks.)

The results of the factor analyses indicate that accuracy and complexity, together with length of run, load highly on the third factor suggesting that more accurate language was also more complex. Furthermore, the fact that these measures are associated with each other indicates that they are reflecting the same underlying constructs. This confirms the assumption that accuracy and complexity are both aspects of form and are in contrast with fluency as an aspect of meaning. Length of run loaded on this factor for the Football and Picnic tasks, but not in the factor analysis for the two less structured tasks.

Table 4. Factor analysis for the football task

Measures	Factor 1	Factor 2	Factor 3	Communality
Reformulations		.88		.880
False starts		.94		.892
Replacements		.41		.276
Repetitions		.62		.490
Accuracy			.65	.662
Complexity			.87	.716
Length of run	–.66	–.44	.43	.767
Speech rate	–.84			.793
Total silence	.95			.912
Time spent speaking	–.94			.902
No. of pauses	.80			.736
Mean length of pause	.87			.844

The effects of task structure, planning condition and proficiency level on language performance

To investigate these effects a repeated measures MANOVA was carried out. Based on the results of the factor analyses, four measures were selected from the total number of 12 measures to represent the dependent variables: number of false starts, number of pauses, accuracy and complexity of the performance. The criterion for selecting one measure from the temporal and one from the repair fluency in each factor group was the consistency in loadings of these measures across all the tasks. As regards language form, since complexity and accuracy have shown themselves in previous research to be influenced by different independent variables, they were both included in the analysis. The independent variables of the analysis were planning and proficiency level, each with two levels and task structure with four levels. The results from the repeated measures MANOVA are presented in Table 5.

With regard to the between-participants effect, the analysis revealed a significant difference between the planners and non-planners (Pillais = .179, $F = 4.00$, $P = .006$) as well as low and high proficiency levels (Pillais = .374, $F = 10.89$, $P = .001$). A significant difference was further seen across the four tasks as the within-participants variable (Pillais = .754, $F = 16.78$, $P = .001$) with differences being concentrated in the number of pauses (Walkman: M = 26, Unlucky Man: M = 22, Picnic: M = 19, Football: M = 18); in complexity (Walkman: M = 1.36, Unlucky Man: M = 1.32, Picnic: M = 1.60, Football: M = 1.43);. in false starts (Walkman: M = 5.2, Unlucky Man: M = 4.41, Picnic: M = 4.59, Football: M = 3.97); and in accuracy (Walkman: M = .30, Unlucky Man:

Table 5. Results of repeated measures MANOVA

Effects	Pillai's Value	F	BGdf	WGdf	Sig.
Between-participants effect					
Planning	.179	4.00	4	73	.006*
Proficiency	.374	10.89	4	73	.001*
Planning × Proficiency	.103	2.09	4	73	.09
Within-participants effects					
Task	.754	16.78	12	65	.001*
Task × Planning	.263	1.93	12	65	.16
Task × Proficiency	.278	2.09	12	65	.12
Task × Planning x proficiency	.288	2.19	12	65	.09

*Significant difference is reached

Table 6. Univariate test of within-participant effect

Source	Measure	Sum of squares	df	Mean square	F	Sig.
Task	No. of pauses	3047.55	3	1015.85	20.21	.001*
	Complexity	3.53	3	1.18	25.65	.001*
	False start	63.00	3	21.00	3.95	.009*
	Accuracy	1.22	3	.407	29.80	.001*
Task × Planning	No. of pauses	64.55	3	21.52	.42	.73
	Complexity	.431	3	.144	3.126	.02
	False start	24.85	3	8.28	1.55	.2
	Accuracy	.082	3	.027	2.015	.11
Task × proficiency	No. of pauses	6.83	3	2.27	.045	.98
	Complexity	.465	3	.155	3.37	.01*
	False start	21.60	3	7.20	1.35	.25
	Accuracy	.123	3	.041	3.00	.03
Task × Pl. × Prof.	No. of pauses	291.9	3	97.3	1.93	.124
	Complexity	.184	3	.06	1.33	.26
	False start	29.18	3	9.72	1.83	.14
	Accuracy	.034	3	.011	.83	.36

M = .30, Picnic: M = .43, Football: M = .42). When the results for the dependent variables were considered separately through a Univariate F test, using a Bonferroni adjusted alpha level (recommended by Tabachnic & Fidell 1996), significance was reached for all the four measures as a result of the task effect. However, the only significant result in the interaction effects between task and

proficiency level was seen for complexity. Results of the Univariate F test are provided in Table 6.

The results indicate that there is a statistically significant difference across the tasks. Further comparisons of all the four measures showed that the structured tasks were not different from one another but they were significantly different from the unstructured tasks in terms of the number of pauses and accuracy. Regarding complexity and false starts, the unstructured tasks were not significantly different from one another but were different from one of the structured tasks.

The effects of task structure

A one-way ANOVA was carried out on each of the 12 independent variables to determine which measures yielded significant differences. Where significance was reached the Scheffé test was run to establish where the differences were located. In cases of non-significant results, pairwise comparisons between tasks were run to explore the differences between pairs of the tasks to gain an understanding of trends within the data. Results of the ANOVAs for all the tasks are given in Table 7 with the F-values, significance levels, means for the four tasks, standard deviations, and an indication of where differences reached significance.

The results show that differences across the four tasks were significant on the measures of total amount of silence, length of run, speaking time, number of pauses and false start (See Table 7). For all these measures the differences reached significance with performance on one or both of the structured tasks being more fluent than performance on one or both of the unstructured tasks. For number of pauses and speaking time the two structured tasks, i.e. Football and Picnic, were significantly more fluent than the two unstructured tasks, i.e. Unlucky Man and Walkman. For length of run Football was significantly more fluent than Unlucky Man. For total amount of silence Football and Picnic are different from Walkman and Unlucky Man, and for false starts Football was significantly more fluent than Walkman. The results of the ANOVAs reveal that significant differences were reached between the structured tasks and unstructured tasks with regard to the accuracy measure ($F = 9.79,\ P < .001$). The results of the Scheffé test showed that the two structured tasks generated significantly more accurate language than the two unstructured tasks. Although it was hypothesized that there would be no significant difference between the complexity of the language generated by the structured tasks and that of the un-

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.24 (1395-1395)

Table 7. Results from the ANOVA on all measures for the four tasks

Measures	*F*	*P*	Task		Structure		Location of Sig. differences	Sig. Pairwise Comparison
			Walkman	Unlucky Man	Picnic	Football		
Total silence	3.80	.04*	29.62 (SD = 23.87)	27.39 (SD = 25.09)	20.57 (SD = 22.72)	19.47 (SD = 19.56)	F vs. W	F P vs. U W
Length of run	4.99	.008*	3.59 (SD = 1.07)	3.29 (SD = 1.09)	3.85 (SD = 1.41)	4.05 (SD = 1.57)	F P vs. U	F vs. U W
Pause length	2.72	.16	1.02 (SD = .54)	1.09 (SD = .61)	.9 (SD = .38)	.9 (SD = .4)		F vs. U
No of pauses	6.90	.001*	26.6 (SD = 12.71)	22.51 (SD = 11.46)	19.92 (SD = 12.41)	18.47 (SD = 11.89)	F P vs. W	F P vs. U W
Speaking time	6.45	.001*	.71 (SD = .15)	.69 (SD = .15)	.76 (SD = .13)	.79 (SD = .13)	F P vs. U	F P vs. U W
Speech rate	1.57	.19	94.66 (SD = 29.42)	87.76 (SD = 29.95)	99.27 (SD = 41.77)	94.85 (SD = 33.09)		
False start	1.87	.13	5.2 (SD = 3.88)	4.41 (SD = 3.35)	4.58 (SD = 3.05)	3.96 (SD = 3.02)		F vs. W
Reformulation		2.79	3.06 (SD = 2.4)	2.28 (SD = 1.96)	2.79 (SD = 2.09)	2.21 (SD = 2.17)		F vs. W U vs. W
Replacement	1.29	.27	.61 (SD = .83)	.71 (SD = 1.41)	.43 (SD = .69)	.48 (SD = .79)		
Repetition	.53	.65	4.28 (SD = 4.58)	3.88 (SD = 4.69)	3.46 (SD = 3.45)	3.78 (SD = 3.70)		
Accuracy	9.79	.001*	.30 (SD = .20)	.30 (SD = .18)	.43 (SD = .19)	.42 (SD = .22)	F P vs. U W	
Complexity	15.19	.001*	1.36 (SD = .28)	1.31 (SD = .20)	1.59 (SD = .33)	1.43 (SD = .28)	P vs. F U W	P vs. U W

* Significant differences are reached across tasks; F = Football, P = Picnic, U = Unlucky Man, W = Walkman.

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.25 (1395-1553)

structured tasks, the results indicate that the Picnic task (M = 1.59) generated significantly greater complexity of language compared with the other tasks.

The effects of strategic planning and proficiency

Hypothesis Three predicted that language performance under planned conditions would be more fluent, more accurate and more complex than that produced under unplanned conditions. A series of t-tests were carried out on each dependent variable to locate the effect of strategic planning time on different measures of fluency, accuracy and complexity. Furthermore, to compare the effect of planning with the effect of language proficiency on the dependent variables, a number of t-tests were carried out on all measures of fluency, accuracy and complexity for the two levels of proficiency. The results of the t-tests for planning conditions and proficiency levels are presented in Table 8.

The results of the t-tests show that the effect of strategic planning reached statistical significance for measures of total silence ($t = 4.16$, $P = .001$), length of run ($t = 4.16$, $P = .001$), pause length ($t = 5.93$, $P = .001$), speaking time ($t = 5.80$, $P = .001$) and speech rate ($t = 3.14, P = .008$). The mean scores for each measure show that performances were significantly more fluent under planned conditions. Although the measures of number of pauses and reformulations do not reach statistical significance, the reductions in these measures clearly show that performance under planned conditions tends to be more fluent than unplanned performance.

All measures of temporal fluency are significantly higher in the intermediate language proficiency group indicating that the language performance of high proficiency participants is more fluent than the performance of low proficiency participants. Interestingly, the effect of strategic planning on the total amount of silence, pause length and speaking time is greater than the effect of language proficiency, i.e. it appears to be better to be a low proficiency planner than an intermediate proficiency non-planner.

It was further hypothesized that language performance would be more accurate under planned conditions. Results of the *t*-tests show that accuracy significantly improved under the planned condition ($t = 5.52$, $P = .001$). Language performed by high proficiency participants is also significantly more accurate than low proficiency participants' language ($t = 7.34$, $P = .001$). However, the effect of proficiency level on accuracy is greater than the effect of pre-task planning.

Language produced under planned conditions was predicted to be more complex under planned conditions. As can be seen in Table 8, a statistically

Table 8. Results of T-test for planning conditions and proficiency levels

Measures	*T*	*P*	Unplanned	Planned	*T*	*P*	Elementary	Intermediate
Total silence	4.16	.001*	29.53 (SD = 27.42)	19 (SD = 16.55)	3.07	.004*	28.22 (SD = 22.94)	20.32 (SD = 22.88)
Length of run	4.16	.001*	3.39 (SD = 1.18)	4.00 (SD = 1.4)	6.12	.001*	3.26 (SD = 1.10)	4.12 (SD = 1.46)
Pause length	5.93	.001*	1.14 (SD = .6)	.81 (SD = .3)	3.72	.001*	1.08 (SD = .56)	.87 (SD = .45)
No. of pauses	1.68	.18	23.05 (SD = 12.91)	20.70 (SD = 11.92)	2.00	.08*	23.26 (SD = 11.28)	20.48 (SD = 13.42)
Speaking time	5.80	.001*	.69 (SD = .15)	.78 (SD = .13)	2.65	.01*	.71 (SD = .15)	.76 (SD = .14)
Speech rate	3.14	.008*	88.23 (SD = 38.04)	100.04 (SD = 28.36)	6.72	.001*	82.15 (SD = 25.82)	106.12 (SD = 36.94)
False start	.21	.82	4.5 (SD = 3.17)	4.58 (SD = 3.55)	2.87	.008*	5.07 (SD = 3.63)	4.00 (SD = 3.00)
Reformulation	1.12	.26	2.45 (SD = 2.03)	2.72 (SD = 2.33)	2.46	.03*	2.88 (SD = 2.53)	2.28 (SD = 1.74)
Replacement	.57	.56	.53 (SD = 1.12)	.59 (SD = .81)	1.95	.1	.66 (SD = 1.18)	.45 (SD = .70)
Repetition	.78	.43	3.67 (SD = 3.57)	4.00 (SD = 4.63)	1.35	.34	4.16 (SD = 3.82)	3.54 (SD = 4.41)
Accuracy	5.52	.001*	.30 (SD = .19)	.42 (SD = .21)	7.43	.001*	.28 (SD = .19)	.44 (SD = .20)
Complexity	2.23	.04*	1.38 (SD = .29)	1.46 (SD = .29)	6.62	.001*	1.32 (SD = .23)	1.53 (SD = .31)

* Significant differences are reached.

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.27 (1553-1639)

significant difference is reached for the complexity of performance between the two planning conditions ($t = 2.32$, $P = .04$) with the planned group achieving higher degrees of complexity in their performance. The results of the t-tests also reveal that the effect of proficiency level on complexity seems to be greater than the effect of pre-task planning ($t = 6.62$, $P = .001$).

Perceptions of task difficulty

A three-way ANOVA, using responses to the task difficulty questionnaire items as the dependent variable and task structure, pre-task planning and proficiency levels as the independent variables, was carried out. Considering the Bonferroni adjusted alpha level (Tabachnic & Fidell 1996), the results of the three-way ANOVA show a significant difference for task structure ($F = 32.63$, $P = .001$) and also a significant difference for the planning conditions ($F = 6.11$, $P = .02$). However, no significance was reached for proficiency level or the interaction between the dependent variables. Table 9 shows the results of the three-way ANOVA on participant perceptions of task difficulty.

Mean scores of the perceptions of difficulty across tasks under the two pre-task planning conditions are shown in Table 10. The comparison shows that the two unstructured tasks, i.e. the Unlucky Man and Walkman tasks, were rated as more difficult than the two structured tasks under both the planning conditions.

Table 9. Three-way ANOVA on perceptions of task difficulty

Source	Type III Sum of Squares	df	Mean Square	*F*	*P*	Eta. Square
Task	42.10	3	14.03	32.63	.001	.244
Planning	2.62	1	2.62	6.11	.02	.02
Proficiency Level	1.12	1	1.12	2.62	.106	.009
Planning × Task	.58	3	.19	.45	.71	.004
Task × Prof.	.93	3	.311	.72	.53	.007
Prof. × Planning	.37	1	.37	.87	.34	.003
Plan × Prof. × Task	.33	3	.11	.25	.85	.003

Table 10. Mean scores of perceptions on task difficulty

Tasks	Football	Picnic	Unlucky Man	Walkman
Unplanned	1.90	1.95	2.67	2.55
Planned	1.80	1.62	2.52	2.40

Table 11. Three-way ANOVA on usefulness of planning time

Source	Type III Sum of Squares	df	Mean Square	*F*	*P*	Eta. Square
Task	1.53	3	.51	.81	.48	.016
Proficiency	.94	1	.94	1.50	.22	.010
Task × Prof. Level	6.25	3	2.08	3.33	.06	.063

A Scheffé test was then carried out to explore where the significant differences were located across the tasks. The multiple comparisons showed that the two structured tasks, i.e. the Football and Picnic tasks were not statistically different from one another but were statistically different from the two unstructured tasks.

Questionnaires of the planned group included a section on the usefulness of the strategic planning time for each of the tasks. A two-way ANOVA was carried out to investigate whether the participants from the two proficiency levels found pre-task planning time more useful for one or some of the tasks. Results of the analysis did not reveal any significant differences across the tasks or the proficiency levels. Table 11 shows the results of the analysis.

Discussion

At the outset, the results of the factor analyses are worth brief comment. First, the consistency of the results across the four tasks is noteworthy, and provides some confidence in the robustness of the structures that are involved. Second, the results regarding fluency are striking, in that there is a separation between repair fluency, on the one hand, and breakdown fluency/rate of speech/unit size, on the other. The repair fluency group loads together consistently, and separately from other aspects of fluency, suggesting that a concern to modify utterances on-line is somewhat distinct from a capacity to organise speech in real time. In addition, the grouping together of the three non-repair aspects of fluency (avoidance of pausing; speed of production; and size of unit that is produced) suggests that while these areas may be distinguished conceptually, in actual performance they draw upon one general capacity to orchestrate speech effectively. Third, it is interesting that the two form-focussed areas of accuracy and complexity load together, consistently as the third factor, suggesting that a concern for language form is less prominent in the data matrix, and does not differentiate so clearly between these two different aspects of performance as

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.29 (1703-1760)

is the case in other task-based studies, e.g. Skehan and Foster (1997). We will return to the relationship between these two below.

The findings regarding task are particularly interesting. The figures for silence and pausing are clear, since they each reduce (i.e. indicate greater fluency) as a function of hypothesised task structure. For example, number of pauses reduces from 26.6 (Walkman) to 22.51 (Unlucky Man) to 19.92 (Picnic) to 18.47 (Football). False starts and reformulations pattern similarly, and it could be argued that the trends are similar for replacement and repetition. It is interesting that the other aspects of the bundle of measures which defined the first factor do not pattern so clearly, and it appears that the putatively more structured Picnic and Football tasks lead to somewhat higher performances than do the less structured tasks, especially Unlucky Man, for length of run, speaking time, and speech rate. It appears that fluency is a complex construct and that explaining these patterns will be an interesting challenge for future research and theorising.

The results for accuracy suggest that the two least structured tasks, (Walkman and Unlucky Man) contrast with the two more structured tasks (Picnic and Football) but that within each pair of tasks, there is little difference. In a sense, therefore, this does provide support for the claim that task structure influences accuracy, but there is not the same general progression over all four tasks as there is with breakdown fluency. In other words, there is not evidence in support of a scale of structure, with gradations. The evidence is more consistent with a threshold of structure, such that above this threshold, accuracy is supported, but below it, accuracy is reduced. If the Football task contains additional structure to the Picnic task, this does not manifest itself in more accurate performance. It may be necessary therefore to rethink how structure impacts on the accuracy of performance in greater detail.

Regarding complexity, there are further complications in the results. The prediction was that there would be no effects for complexity. In the event, there is a significant finding, in that the Picnic task has generated greater complexity than the other three tasks. None of the other tasks differ from one another. This result presents a puzzle, and will have to await further research. However, it could be argued that a post-hoc analysis of the different picture series does suggest that an additional feature distinguishes the Picnic task from the other three. This is that the point of this story sequence requires subjects to make connections between foreground and background elements in each of the various pictures in the sequence, and that it is this which underlies the greater complexity of the language used. Future research will be needed to explicate this finding.

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.30 (1760-1800)

The findings on planning are fairly consistent and clear. Fluency is significantly advantaged by strategic planning. Regarding breakdown fluency, this applies to total silence, and pause length, but not to number of pauses. It applies consistently to rate of speech and size of unit indices (speaking time, speech rate, and length of run). None of the repair fluency measures produces significance, and interestingly each is higher for the *planned* condition, i.e. the planners seem to be more likely to engage in modification of speech on-line. But taken broadly, these results indicate strong support for the claim that pre-task planning leads to a significant increase in the main dimensions of fluency.

Turning to the form-linked measures, the results for complexity are significant, though somewhat surprisingly only at the 0.05 level, with a difference of 1.46 to 1.38, measured in AS-units. This does suggest that pre-task planning produces greater complexity, but the size of the effect is not especially great. On the other hand, the results for accuracy reach a much more demanding significance level, and indicate that pre-task planning time has much more of an impact on accuracy. Bearing in mind that there is an assessment context for these results, it may be the case the learners taking tests shift priorities somewhat, and value accuracy more than complexity, whereas learners doing tasks as part of classroom behaviour may be more inclined to focus attention on doing a task to its potential, and allocate attention to complexity.

The findings for proficiency level are mostly straightforward, and provide interesting validity confirmation that this data elicitation format produces results consistent with conventional test results. The repair fluency measures present their usual mixed picture, in this case with false starts and reformulations being significantly less frequent in the higher proficiency group, and with no significances for replacement and repetition. Otherwise, all measures except number of pauses produce significant results in favour of the higher proficiency group, with all these significances being at least at the 0.01 level. The effect sizes for accuracy and complexity are particularly noteworthy.

There is one particular interesting finding when one looks at the relationship between pre-task planning and proficiency level. As indicated earlier, there are occasions where there is higher performance by the Elementary proficiency planners compared to the Intermediate non-planners. This is interesting because it suggests that higher performance can be achieved if task/assessment conditions allow for planning compared to simply having a higher proficiency level. If we relate this to the model of spoken language assessment shown in Figure One, it makes it clear that scores assigned may not reflect simply profi-

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.31 (1800-1854)

ciency level, but also the conditions under which a task is done. We will return to this below.

Finally, the results for perceptions of task difficulty are worth discussing. Broadly, the unstructured tasks (Unlucky Man and Walkman) were rated as more difficult than the structured tasks (Football and Picnic), with there being little difference between the pairs of tasks in each case. Further, the non-planners rated the task as more difficult than did the planners, although this difference was not so great in scale. In other words, participants rated as *less* difficult the tasks and task conditions when they did better. If we assume that they did better because they were more able to do justice to what they perceived as their "true" proficiency, then it would appear that the participants themselves were aware, at some level, that they were performing differently, and presumably, in a more satisfactory manner. It would appear, in other words, that having either a clear general structure within which to frame the narrative, or having pre-task planning opportunity to enable resources to be mobilised (and possibly achieve the same end) led to learners who felt they were more in control of the situation (cf. awareness of Formulator operation). It is particularly interesting that the lowest difficulty rating, i.e. the "easiest" task, was the Picnic story under the planned condition. This is interesting, and worthy of future research, because this was the experimental condition which led not simply to higher fluency and accuracy, but also complexity. Despite interpreting the task in such a way that the most advanced language was produced, participants nonetheless regarded the task as easier.

Conclusions

Above all, the present study has contributed results which clarify the functioning of the model shown in Figure 1, and help to take the model beyond schematic value and towards a sounder empirical base. The study clarifies that:

- task structure is an important influence on performance;
- strategic planning has an effective and predictable influence, and generally improves the level of performance which results. This applies to all aspects of performance, including accuracy;
- different aspects of performance are affected slightly differently by structure and pre-task strategic planning.

These results suggest that it is fruitful to chart potential task-based influences on actual candidate performance. The wider issue, clearly, is that tasks and

task conditions vary, with the result that a particular testing encounter could use a combination of circumstances which inadvertently impact upon performance, and subsequent assessments. Without knowledge of these effects, the danger is that test scores which are assigned are partly artifactual, and difficult to compare with results obtained under different conditions.

The results also indicate how more experimental studies can contribute to language test validation. It may be the case that some variables, e.g. adequacy, perspective, immediacy, as defined by Iwashita et al. (2001) do not have significant effects. The present study has shown that there are, though, other relevant variables (task structure) which do impact upon performance. Clearly, further research to uncover other potential influences is warranted. But the present results also question the findings regarding the non-influence of strategic planning on accuracy which have occasionally been reported in the literature (e.g. Crookes 1989; Ortega 1999). In the present case, strategic planning generated consistent significances, even, on occasion, leading to stronger influence on performance than proficiency. Future research will have to explore why strategic planning works in some circumstances and not in others.

The present results are also compatible with the revised system proposed by Brown et al. (2002) for assessing task difficulty. Amongst the factors that they propose that can influence the difficulty of tasks they include the cognitive operations that are required to carry out a task, and within this area, they have sub-headings of input/output organisation and input/output availability. The first concerns the degree of transformation of the elements of the task that are required, and the second focuses on the information that is the basis for the task and the ease of accessibility of this information. The findings on structure are relevant to each of these. More clearly structured tasks contain a clearer organisational framework, removing the need for attention to be directed to re-organising material. As a consequence, attention becomes available to access information which will enable the task to be carried out more easily. In other words, greater structure does appear to ease the task faced by the test-taker, and permits the allocation of attention to formulation (Levelt 1989), and as a result, fluency and accuracy. The findings, in other words, suggest slight modification of these two sub-areas of cognitive operations in the system provided by Brown et al. (2002), so that within their sub-category of input/output organisation, information structure is highlighted a little more than it is at present.

Two additional problems, though, are worth mentioning. First, there is the issue of task difficulty itself. Brown et al. (2002) regard the problem of identifying task difficulty to be a central aim for research. They regard difficulty to be a joint function of ability requirements and task characteristics. In other words

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.33 (1919-1952)

they want to ask a question like: "What difficulty level in a task can a candidate of a certain proficiency level complete?" In this way, in a testing situation, one would want to give candidates of different proficiency levels tasks of appropriate difficulty level to transact. The aim of testing would then be to most efficiently identify the maximum level a particular test candidate could cope with. But the present research has once again confirmed that performance is multi-dimensional, (cf. the degree of independence between performance areas such fluency, accuracy, and complexity) and compensatory and that it is difficult to propose that there is a central "gold standard" criterion one can use to identify difficulty level. Task characteristics may vary and this variation may connect systematically with different aspects of performance, but the problem is that these different dimensions do not function in unison: increasing performance in one area may not be associated with increased performance in another. The problem then becomes how to handle this inconsistency at the level of the ratings and measurements that are at play. Perhaps this is what Bachman (2002) means when he argues that difficulty is not a separate factor, but resides in the interactions of all the features that are involved in an assessment situation.

The second general problem may be a version of the Observer's Paradox. It is clear that there are inconsistencies in the findings between different studies in the literature and that one of the goals of future research will be to account for such discrepancies. In that respect, one can distinguish between two sorts of research:

– studies which are pedagogy or acquisition oriented;
– studies which replicate testing conditions.

The inconsistency between these two areas in terms of the likelihood of finding significant effects for experimental variables is worth pursuing. Potentially, four sources of difference may be relevant:

that testing-linked task research leads to a different type of language use, because of the prominence of the assessment context, and possibly a greater focus on accuracy (since this is what tests are traditionally perceived to be about). This might mean that there is less scope for experimental manipulations to produce additional accuracy effects, since attention is already being directed to that area of performance. Even so, this would not explain why studies such as Iwashita et al. (2001) have failed to find complexity or fluency effects either. However, a broader Observer's Paradox interpretation would presumably be arguing for

JB[v.20020404] Prn:20/10/2004; 12:48 F: LLLT1109.tex / p.34 (1952-2004)

inconsistency going beyond accuracy, and incorporating other domains as well. That the act of testing distorts performance remains a worrying possibility.

that different types of experimental variables are researched in the two contexts, and that it is the variables associated with acquisition or pedagogy (A/P) which are more likely to produce significant effects. Certainly the A/P studies have used a greater range of tasks, including decision-making, narrative, personal information exchange, picture description and so on, compared to the more narrative-task focus in assessment. This has permitted the use of variables which emphasise the interactive nature of the wider range of tasks. Conversely, assessment oriented researchers have tended to choose variables which can be manipulated to generate potential differences in difficulty level, and act as a blueprint for generating multiple versions of tasks of equivalent difficulty level in a straightforward manner. Hence the use of variables such as adequacy and perspective. But these differences in findings do not account for the discrepancies in the effects of planning in the two areas, or in variables such as immediacy of information, which have yielded significant findings for Foster and Skehan (1996) and Robinson (2001c) but not for Iwashita et al. (2001). There seem to be other, unexplored variables at play which need to be probed further.

that common variables have been operationalised differently in the two contexts, with the result that inappropriate comparisons are being made. There is certainly some evidence that this is the case. Wigglesworth (2001), for example, operationalises structure in terms of the amount of information which is provided, an approach not dissimilar to Iwashita et al.'s (2001) treatment of *adequacy*. Both approaches certainly contrast with Skehan and Foster's (1999) treatment of structure in terms of information organisation, as well as the approach taken in this chapter. But there are other variables which are operationalised similarly in both research contexts, such as Iwashita et al.'s (2001) use of immediacy, compared to Robinson's (2001c) "here and now" vs. "there and then" conditions, or alternatively, the approaches to planning in the various studies, and here the testing studies, with lack of significance, compare markedly with the non-assessment studies which do report significances. Once again we are left with a puzzle.

that scoring procedures differ in the two contexts, with A/L researchers tending to use detailed measures and assessment researchers using ratings. While it is certainly true that A/L researchers do not use ratings of performances (although it would not be a bad thing if they did), the assessment researchers have fre-

quently used a range of measures and the lack of significances are reported with *both* methods of evaluating performance.

Perhaps the conclusion to draw here is the familiar one that more research is needed. Although it would be worrying to conclude that it is the Observer's Paradox that strikes at the heart of testing, so to speak, it may be premature to come to that conclusion yet. It appears that the discrepancies in the results, while unlikely to be the result of the differences in scoring procedure, could well be produced by the different foci and motivations for experimental variables, as well as the different operationalisations that have been used. The present study does provide some very relevant results, indicating clear differences between conditions, and appreciable effect sizes, and these in a context which was approached as an assessment environment. Perhaps the best recommendation would be that, with research in this area still developing techniques and understandings, the role of A/L work will be to offer suggestions for relevant variables and the standardization of operationalisations of variables, but that these insights and findings will need to be confirmed within assessment contexts. It is to be hoped that the approach to structure and pre-task planning portrayed in the present chapter can make a contribution to this.

Notes

* The authors would like to thank Rod Ellis and two anonymous reviewers for reading earlier drafts of this chapter, and offering comments which have enabled us to strengthen it considerably.

1. Even so, it should be noted that Robinson might well have predicted this result, since he argues (see above), that language complexity is driven by functional complexity.