Can We Predict Task Difficulty in an Oral Proficiency Test? Exploring the Potential of an Information-Processing Approach to Task Design

Noriko Iwashita and Tim McNamara
University of Melbourne

Catherine Elder
University of Auckland

This study addresses the following question: Are different task characteristics and performance conditions (involving assumed different levels of cognitive demand) associated with different levels of fluency, complexity, or accuracy in test candidate responses? The materials for the study were a series of narrative tasks involving a
The present study investigates predictors of task difficulty in the measurement of speaking proficiency. It adopts an information-processing approach and draws on the framework proposed by Peter Skehan (Skehan, 1996, 1998). In part, the study is a replication of important aspects of Skehan’s research as applied to the testing situation. It aims to examine whether Skehan’s framework can help define an interpretable scale of oral proficiency, in the context of the development of a semi-direct test of speaking.

Since the mid-1980s, the characteristics of oral tasks in language learning have been the subject of intensive attention among second-language acquisition (SLA) researchers. Understanding the cognitive demands different kinds of tasks make on participants has been seen as necessary for the management of the conditions for successful language acquisition, a process that is assumed to depend on the engagement of “malleable interlanguage systems” that are potentially modified as a result of participation in tasks (Skehan & Foster, 1997). Tasks are thus seen as the foundation for a psycholinguistically motivated communicative language syllabus, and research has focused on task characteristics (e.g., type, format, performance conditions) that might form the basis for sequencing and planning task use in a task-based syllabus.

Research on the characteristics of tasks also has potential for the definition of a variable of difficulty in tests of oral language ability. This potential has so far been suggested (Skehan, 1998;
and some promising work has been emerging (e.g., Norris, Brown, Hudson, & Bonk, 2000; Slatyer, Brindley, & Wigglesworth, 2000). The study reported in this paper proposed to initiate the research by looking at the possibility of implementing under test conditions a cognitive framework for understanding the relationship between task characterization and task performance. In particular, it investigated the relationship between task (e.g., type, format, performance conditions) and response characteristics (e.g., accuracy, fluency, complexity), as shown in an analysis of test discourse on the one hand and of test outcomes (i.e., test scores) on the other.

This attempt to define the construct of oral ability in terms of the characteristics of test tasks represents a new departure for the field. In current approaches, the oral ability continuum is not well defined by differences in task difficulty. Henning and Davidson (1987) first pointed to this phenomenon in their early use of Rasch measurement approaches in the testing of speaking and writing skills. They rapidly discovered that the ability continuum was more adequately defined by different score points (1–5) on the rating scale used for each speaking test item than by any differences in task difficulty. Although in their study what were defined as tasks were not so much separate tasks as separate aspects of the performance being rated (“accuracy,” “organization,” and so on), it was possible that this finding would be confirmed even when rating scales were used as the basis for the rating of performance on different tasks. In fact, this has been found to be the case repeatedly in subsequent analyses of performance-based test data. For example, the current Test of Spoken English (TSE) tasks themselves do not differ greatly in difficulty and do not currently offer a basis for defining a variable of oral language ability by differences among the tasks themselves. Although TSE items and tasks are specified in terms of seven contextual characteristics (test interviewer, audience, setting, topic, purpose, function, and visuals), and these characteristics are cross-referenced to facets of the test method, there is no hierarchy of difficulty implicit in the way the tasks are chosen. Douglas and Smith are prepared to say
no more than that “it is expected that the range of situations presented in each version of the test will tap a similar range of situational abilities in the examinee” (1997, p. 16).

It has been a great challenge for both language educators and test developers to define task difficulty, because tasks do not lend themselves readily to categorization for test purposes. Robinson (1996a) points to the problem faced by those wishing to operationalize the notion of task difficulty for language testing purposes, remarking that research has so far established no clear criteria for grading and sequencing tasks. Before discussing task difficulty further, the main approaches to the characterization of tasks that have emerged in the literature will be presented.

The three main approaches to the characterization of tasks are interactional approaches, information-processing approaches, and test-method approaches. One tradition, stemming from the work of interactionists (e.g., Long, in press; Pica, 1994; Pica, Kanagy, & Falodun, 1993), has attempted to determine the interactional characteristics of tasks and their impact on negotiation in dyadic communication. A second tradition, the information-processing approach, perhaps best known in the work of Skehan and Robinson, has examined the impact on performance of the cognitive characteristics of tasks, and links have been explicitly made with predictions of task difficulty (Skehan, 1996, 1998; Robinson, 1995, 1996a, in press-a, in press-b). A third tradition, in which test task characteristics are defined as test methods (Bachman & Palmer, 1996), has suggested that the impact of task differences can best be understood as a question of test-method effects. This latter framework for characterizing test methods is also considered in the context of this project.

Information-Processing Approaches and Implications for Language Testing

From the point of view of proficiency testing, the second, information-processing approach offers a more promising framework within which to view performance on oral language tasks. A
number of researchers have begun a systematic program of work exploring the effects of aspects of task structure and performance conditions on performance on second language oral tasks (Brown, Hudson, & Norris, 1999; Norris, Brown, Hudson, & Yoshioka, 1998; Robinson, 1996a, in press-a, in press-b; Skehan, 1996, 1998). To date, research has attempted to take a number of factors into account, and these have differed to some extent among researchers. For example, three factors—cognitive load, planning time, and prior information—are discussed by Robinson, Tin, and Urwin (1995), whereas Norris et al. (1998) and Brown et al. (1999) have considered the impact of linguistic code, cognitive complexity, and communicative demand on task difficulty. Following Brown, Anderson, Shilcock, and Yule (1984), Skehan (1998) proposes a number of factors relevant to task difficulty, grouped under the broad headings of “task characteristics” and “task implementation conditions” (or “processing conditions”). Skehan (1998) recognizes the potential of this work for defining an ability/difficulty continuum using Item Response Theory (IRT) approaches:

The issues reviewed here are highly relevant to testing. One view of testing would be that there is a scale of difficulty and that students with greater levels of underlying ability will then be able to successfully complete tasks which come higher on such a scale of difficulty. Such an approach underlies, for example, the postulation by item-response theorists of a unidimensional scale for language proficiency against which individual test takers can be located. The discussion of task difficulty presented earlier is, in many ways, targeted at the same place, except that it tries to achieve its aims by means external to the analysis of test results. (p. 184)

Skehan argues that if tasks are to attain a position of importance in the testing context, it is necessary “to know more about the ways tasks themselves influence (and constrain) performance” (1998, p. 169). Following Kenyon (1992) and McNamara (1996), he proposes an expanded model of oral test performance, based on a psycholinguistic processing perspective, in which the notion of task has a central place. The value of such a model, according to
Skehan, is that it becomes an organizing framework, which can reveal the precise nature of the mediation that occurs between any underlying abilities and the way a task is transacted and hence avoids the view that tasks are neutral devices for testing. According to Skehan (1998), research conducted within this framework has the potential to

(a) inform task selection (by allowing us to predict the relative difficulty of each task and to ensure the full range of candidates’ ability will be tapped);

(b) assist test developers in structuring test tasks and the conditions under which these tasks are performed in appropriate ways;

(c) inform development of rating scale descriptors;

(d) facilitate interpretation of test scores (which may differ according to task).

On the basis of previous studies by himself and others (e.g., Foster & Skehan, 1996; Skehan, 1996; Skehan & Foster, 1997, 1999; Mehnert, 1998), Skehan has suggested a number of factors having an impact on task difficulty. A number of task dimensions (e.g., abstractness of task information, familiarity of task information) are considered, and within each, relevant contrasting performance conditions (e.g., concrete vs. abstract information, familiar vs. unfamiliar information) are held to affect the difficulty of the task. All this is summarized in Table 1. In the table, two values for each factor are contrasted, with the performance conditions on the right associated with greater task difficulty.

The above findings need further empirical confirmation and should therefore be seen as tentative. A further range of possible influences canvassed by Skehan (1998, pp. 176–177) are also considered likely to affect difficulty. Among these, two seem to have particular relevance to the development of a semidirect test of speaking proficiency, given the implementation constraints of computer-mediated delivery. These are time pressure (as pressure
increases, difficulty increases) and the degree of visual support provided (greater support leading to easier tasks; cf. Crookes & Rulon, 1988; Iwashita, 1998). Other aspects of performance conditions may affect task difficulty in a more complex way, in that tasks emphasizing particular types of attentional demand may have a measurable impact not on task difficulty as a whole but on particular aspects of performance. The following components of cognitive processing in performance on speaking tasks have been so identified in research: for example, tasks and performance conditions that direct attentional resources to form and rule. These tasks and conditions may induce “risk-avoiding” or “risk-taking” behavior, yielding variation in measures of accuracy and complexity of language, respectively (e.g., Rahimpour, 1997; Robinson, 1995). Robinson (1995) explains the difference between easy and hard tasks in terms of Givon’s taxonomy of syntactic and pragmatic modes. Easy tasks that involve contextual support may lead learners to stay within the pragmatic mode and require interlocutors to fill in large quantities of linguistically uncoded

Table 1

Task dimensions and performance conditions influencing task difficulty

<table>
<thead>
<tr>
<th>Task dimension</th>
<th>Performance conditions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Less difficult</td>
</tr>
<tr>
<td>Number of participants or elements</td>
<td>few</td>
</tr>
<tr>
<td>Abstractness of information or task</td>
<td>concrete</td>
</tr>
<tr>
<td>Type of task information</td>
<td>immediate, here-and-now</td>
</tr>
<tr>
<td>Nature of operation required on task information</td>
<td>retrieval</td>
</tr>
<tr>
<td>Familiarity of task information</td>
<td>familiar</td>
</tr>
</tbody>
</table>

Note. This table is based on Skehan, 1998, p. 174, Table 7.2.
information from the context. On the other hand, in harder tasks in which contextual support is not available, learners can no longer confine themselves to the pragmatic mode and can be expected to draw on a greater range of syntactic resources.

Furthermore, tasks and performance conditions that focus attentional resources on meaning and real-time processing have been found to yield variation in measures of fluency. The bulk of research has gone into planning time (e.g., Crookes, 1989; Ellis, 1987; Mehnert, 1998; Ortega, 1999; Ting, 1996; Wigglesworth, 1997). These studies in general hypothesized that planning time would enable L2 learners to produce more accurate, higher levels of lexical complexity and fluent speech. However, results are rather mixed, and this is particularly true for accuracy. While some studies showed more accurate production under the planned condition (e.g., Ortega, 1999; Ting, 1996), other studies found that this increased accuracy was evident only for certain tasks (e.g., Foster & Skehan, 1996; Mehnert, 1998) and on certain measures (e.g., Crookes, 1989). Reviewing empirical studies on pretask planning, Ortega (1999) argues that it is difficult to evaluate the effect of pretask planning, because of the range of interlanguage measures and the different operationalizations of planning employed across the studies.

Only recently have researchers have begun to explore the impact of cognitive demands of tasks in language-testing contexts. For example, Wigglesworth (1997) has investigated the impact of planning time on measures of accuracy, fluency, and complexity, with mixed results. The results showed that subjects who were given planning time improved their performance in terms of complexity (amount of subordination), fluency (number of self-repairs), and accuracy (suppliance of plural-s, verbal morphology, and indefinite article). More recently, Slatyer et al. (2000) have examined the role played by certain key variables in the assessment of listening tasks and found that there is a complex interaction between the text and other components of the task and that these in turn interact with the attributes of individual candidates. They conclude that the task characteristics are more usefully
conceived of as degrees on a continuum rather than in terms of a simple dichotomy, and also that statements about the effect of any single task or item characteristic on difficulty would “need to be carefully qualified in the light of what is known about its interaction with other variables” (Slatyer et al., 2000, p. 19). Their work differs from our study in that difficulty factors were identified post hoc rather than on the basis of a preexisting taxonomy or framework. Their findings, however, suggest that accurate a priori estimation of task difficulty may be extremely hard to achieve.

Work by a research group in Hawai’i has built specifically on Skehan’s framework in the design of syllabus-related performance assessments (Norris et al., 1998; Brown et al., 1999; Norris et al., 2000). Skehan’s dimensions of cognitive demand have been adopted and modified in the design of specific tasks. The researchers report some difficulty in operationalizing the proposed task dimensions in actual task design, an issue to be revisited below. Moreover, they find only weak support for the proposed relationships between the combinations of cognitive factors with particular task types and actual task difficulty as manifest in task performance by candidates at a range of ability levels. The tasks classified as Type 6 (those characterized as the most difficult in terms of code, cognitive complexity, and communicative demand) did in fact distinguish between NS candidates and all others, but were not associated with performance differences among lower-proficiency candidates. Type 2 tasks (those predicted to the least cognitively demanding) resulted in widely varying performances within all proficiency subgroups (Norris et al., 2000). This work is the closest to that carried out in the present study, yet differs from it in a number of ways:

- Skehan’s framework has been heavily modified and a new framework for characterizing task difficulty developed, related to specific tasks in the curriculum.

- The tasks are interactive and heavily contextualized and are therefore unsuitable for use in standardized proficiency testing.
Although some promising work has emerged based on Skehan’s framework, there are aspects of this work that are problematic: (1) the notion of difficulty, etc.; (2) the measures of candidate performance, etc.; (3) findings to date that have not always been consistent; (4) the complexity of task performance.

The notion of difficulty has not been properly operationalized in these studies. No single measure of difficulty has been used, and instead, there has been an unacknowledged shift from measures of accuracy, complexity, and fluency to discussion of task difficulty. While provisionally one can say that a task that makes it harder to be accurate, or more fluent, is a more difficult task, it seems that the area of complexity of language produced needs to be understood a little differently. More cognitively demanding tasks may elicit a greater range of complexity of language, and if this is the case, the difficulty of the task may need to be defined in terms of the failure of weaker candidates to produce more complex language. This contrasts with accuracy and fluency, where, as tasks become more difficult, weaker candidates will become less accurate, rather than fail to become more accurate. Practically this means that measures of difficulty will have to be thought through carefully, with the goal of a single measure of difficulty remaining a target for research. Researchers to date (e.g., Wigglesworth, 1997) have entered “difficulty” as a facet in multi-faceted Rasch analyses of performance data and have derived empirical measures of difficulty accordingly, but the interpretation of this difficulty in terms of Skehan’s framework remains unclear. Such an a posteriori approach is to be supplemented by an a priori analysis of task difficulty in the current work. Robinson (in press-a) has proposed the independence of the dimensions of complexity and difficulty, with complexity being a feature of the task, and difficulty operationalized in terms of perceptions of task difficulty on the part of test-takers. While a distinction between the task and the test-taker is important to recognize in principle, in practice, as the research of Slatyer et al. (2000) and Norris et al. (2000) indicates, there is likely to be a strong relationship between the two. Scored performance on tasks will, in other words, reflect the
interaction between task complexity on the one hand and the abilities and attitudes of the test taker on the other. The latter “interactive” approach to task difficulty is the one relevant in the present research context.

(1) The measures of candidate performance used in research to date have been complex and time-consuming, involving close examination of the output, and are not therefore most effective in operational testing conditions. The need for practical and efficient measures of ability that are sensitive to the variations in task structure and processing conditions proposed in the research remains an important issue. It would also seem that what Skehan describes as “breakdown fluency” (i.e., hesitation) does not relate to “repair fluency” (i.e., repetition, false starts). This particular finding has implications for rating scale design, because it implies that the incorporation of both these aspects of fluency in a single scale descriptor may be problematic.

(2) Findings to date have not always been consistent. For example, Robinson (1996a) reports that the assumption made in all these papers is that greater complexity of tasks in terms of their cognitive demands will facilitate greater attention to form and planning of production and lead, therefore, to greater accuracy in the use of morphosyntax and greater complexity of production (p. 5).

Skehan (1996, 1998), on the other hand, associates greater accuracy with easier (less complex) tasks and distinguishes accuracy from complexity of production. Existing research findings need exploration and confirmation.

(3) The complexity of task performance and the manner in which factors interact with each other in hard-to-control ways is an issue increasingly acknowledged in language testing research (O’Loughlin, 1997; McNamara, 1996). It remains to be seen whether a stable environment for the assessment of
ability in relation to tasks of known demand is achievable in a language-testing environment.

Despite the complexity of the research task, research on the implications for testing of cognitive approaches to task characterization seems of great interest and practically worthwhile. As Skehan (1998) puts it:

The conditions under which tasks are done and the way conditions interact with performance are a fertile area for research. Already we can see how, in areas such as pre-task (test) planning, there are findings which clarify how conditions of task elicitation influence performance . . . If we are to understand how testing conditions are to be standardized and/or how an adequate range of sampling conditions is to be identified, an understanding of this set of influences is essential. Unless research bearing on these factors becomes available, generalizations based on actual test performance will be a hazardous and chance-dominated undertaking. (p. 177)

In this study our aim was therefore to explore, first of all, whether we could replicate Skehan's findings in a test (rather than classroom) situation, or, more precisely, when subjects were completing speaking tasks in a semidirect test format. To this end, we adopted the same measures (of fluency accuracy and complexity) as have been used by Skehan, and this involved transcribing and carefully counting selected features of the discourse produced by candidates when performing tasks in different dimensions and under different performance conditions. However, as discourse analysis is not a practical methodology under operational test conditions, we also set out to explore the effect of different performance conditions on scores assigned to candidates' task performance by trained raters. This score data was also analyzed, by using IRT methods.
In the present study the following research question was addressed:

Are different task characteristics and performance conditions (involving assumed different levels of cognitive demand) associated with different levels of fluency, complexity, or accuracy in test candidate responses? The research question was examined through two different kinds of analysis:

1. a quantitative analysis of candidates' test discourse, using measures derived from Skehan's work;
2. a Rasch analyses of candidates' test scores assigned by trained raters using analytical rating scales.

The first approach is that commonly used in SLA studies; the second is now a standard procedure in language testing research.

Method

Participants

We used data from 193 students in the study. There were 71 men and 122 women. The majority (approximately 80%–90%) of the participants were students currently enrolled in an ESL course in Melbourne to prepare for tertiary study in Australia, while the remainder were already studying at a tertiary institution in Melbourne.

The mean age of the participants was 21.9 years ($SD = 4.5$), and the mean length of residence in Australia was 4.3 months. The L1 of subjects varied, but the majority were speakers of Asian languages (e.g., Chinese, Vietnamese, or Japanese). The mean length of time they had been studying English was 6.9 years ($SD = 4.6$); many had additionally studied foreign languages other than English at some time. Most participants spoke English at home in Australia. Participants' mean score on the Institutional
TOEFL test was 493.1, with a SD of 45.8, and a range between 427 and 670.

Materials

Speaking tasks used in the study involved a single type of stimulus, of the kind routinely used in the Test of Spoken English (TSE), namely, a narrative task based on a sequenced set of picture prompts. The stories were all simple narratives with point of climax and resolution of the kind regularly included in EFL books. Through a complex process of materials development, prepiloting on native and nonnative speakers, teacher/researcher workshops, and expert consultation, a final research design was determined. (Details of this process are provided in Iwashita, McNamara, & Elder, 1998). In the final design, four dimensions of narrative tasks, with accompanying performance conditions (+ or –), were used. It was anticipated that varying the performance conditions within each dimension would make the tasks easier (i.e., less cognitively demanding) or more difficult (i.e., more cognitively demanding) for the participants. The various task dimensions and performance conditions are described below.

Perspective. Participants were required to tell a story either as if it happened to the participant (– condition) or from someone else’s point of view (+ condition). The hypothesis implicit in our operationalization of this dimension is as follows:

A story told as if it happened to the participant will be less difficult; that is, it will be associated with a more accurate, more fluent, but less complex response from candidates than a story told from someone else’s perspective.

Immediacy (Foster & Skehan, 1996; Robinson, 1995). Participants were required to tell the story with and without the pictures in front of them (i.e., in the “here and now” [– condition] or in the “there and then” [+ condition]). In this case, the underlying hypothesis is as follows:

A story told with a set of accompanying pictures in view will be less difficult; that is, it will be associated with a more accurate,
more fluent, but less complex candidate response than a story told after these pictures have been removed.

**Adequacy.** Participants were required to tell the story with a complete set of six pictures (– condition) or with an incomplete set of pictures (i.e., with two of the six pictures missing [+ condition]). The hypothesis underlying our operationalization of the adequacy dimension is as follows:

A story told with the aid of a complete set of six pictures will be less difficult; that is, it will be associated with a more accurate, more fluent, but less complex candidate response than a story based on an incomplete set of pictures.

**Planning time.** Under the “– planning time” condition, participants were given 3 minutes of planning time in addition to 0.5 minutes to read the instructions and look at the pictures, whereas under the “+ planning time” condition, participants were given only 0.5 minutes total to read the instructions and look at the pictures. The hypothesis to be tested here is as follows:

A story told after 3.5 minutes planning time will be less difficult; i.e., it will be associated with a more accurate, more fluent, but less complex candidate response than a story told with only 0.5 minutes’ total planning time.

In each case, two exemplars of tasks (1 and 2) in each dimension were used to investigate any effect for specificity of task or, in other words, to determine whether our operationalization of the experiment was generalizable across different task exemplars (see Table 2).

**Data Collection Procedure**

Data were collected in November 1998 and April 1999 at the University of Melbourne. The speaking test was administered in a university language laboratory. Before commencing the tests, participants were asked to fill in a questionnaire about their background. All participants were randomly assigned to one of four experimental groups; they completed the speaking test first, and then took the Institutional version of the TOEFL test. All
Table 2

*Design of tasks*

<table>
<thead>
<tr>
<th>Task code</th>
<th>Subcode</th>
<th>Dimension</th>
<th>Requirements of dimension</th>
<th>Predicted difficulty (according to assumed degree of cognitive demand)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1 and A2</td>
<td>–</td>
<td>A. Perspective</td>
<td>tell a story as it happened to you</td>
<td>less difficult (lighter cognitive demand)</td>
</tr>
<tr>
<td>A1 and A2</td>
<td>+</td>
<td></td>
<td>tell a story from someone else's point of view</td>
<td>more difficult (heavier cognitive demand)</td>
</tr>
<tr>
<td>B1 and B2</td>
<td>–</td>
<td>B. Immediacy</td>
<td>here and now (with pictures)</td>
<td>less difficult (lighter cognitive demand)</td>
</tr>
<tr>
<td>B1 and B2</td>
<td>+</td>
<td></td>
<td>there and then (without pictures)</td>
<td>more difficult (heavier cognitive demand)</td>
</tr>
<tr>
<td>C1 and C2</td>
<td>–</td>
<td>C. Adequacy</td>
<td>tell a story from a set of six pictures</td>
<td>less difficult (lighter cognitive demand)</td>
</tr>
<tr>
<td>C1 and C2</td>
<td>+</td>
<td></td>
<td>invent a picture story based on six pictures, but with two of the pictures missing</td>
<td>more difficult (heavier cognitive demand)</td>
</tr>
<tr>
<td>D1 and D2</td>
<td>+</td>
<td>D. Planning time</td>
<td>0.5 min only for reading the instruction and for looking at the pictures</td>
<td>more difficult (heavier cognitive demand)</td>
</tr>
<tr>
<td>D1 and D2</td>
<td>–</td>
<td></td>
<td>0.5 min for reading the instruction and for looking at the pictures, and 3 mins planning time</td>
<td>less difficult (lighter cognitive demand)</td>
</tr>
</tbody>
</table>
participants were required to perform eight speaking tasks (maximum of 3 minutes for each task). After completing the first four tasks, participants had a 10-minute break before resuming the test, and a further break of 10 minutes before taking the TOEFL test. The data collection procedure is summarized in Table 3.

Data

Discourse analysis. An analysis of test discourse was undertaken to determine whether measures of accuracy, fluency, and complexity in participant output differ according to the hypothesized difficulty of performance conditions. We used data from the test performances of 36 participants. These participants were more or less randomly selected, except that care was taken to produce equivalence across conditions in terms of gender and TOEFL score. This number, while somewhat arbitrary, was chosen for the following reasons: (1) it constituted a little under 20% of

Table 3

Data collection procedure

<table>
<thead>
<tr>
<th></th>
<th>Group 1 (n = 37)</th>
<th>Group 2 (n = 58)</th>
<th>Group 3 (n = 51)</th>
<th>Group 4 (n = 47)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1st task</td>
<td>Task A1+</td>
<td>Task A2−</td>
<td>Task A2+</td>
<td>Task A1−</td>
</tr>
<tr>
<td>2nd task</td>
<td>Task B1+</td>
<td>Task B2−</td>
<td>Task B2+</td>
<td>Task B1−</td>
</tr>
<tr>
<td>3rd task</td>
<td>Task B2−</td>
<td>Task B1+</td>
<td>Task B1−</td>
<td>Task B2+</td>
</tr>
<tr>
<td>4th task</td>
<td>Task A2−</td>
<td>Task A1+</td>
<td>Task A1−</td>
<td>Task A2+</td>
</tr>
<tr>
<td></td>
<td>BREAK (10 mins)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5th task</td>
<td>Task C1+</td>
<td>Task C2−</td>
<td>Task C2+</td>
<td>Task C1−</td>
</tr>
<tr>
<td>6th task</td>
<td>Task D1+</td>
<td>Task D2−</td>
<td>Task D2+</td>
<td>Task D1−</td>
</tr>
<tr>
<td>7th task</td>
<td>Task D2−</td>
<td>Task D1+</td>
<td>Task D1−</td>
<td>Task D2+</td>
</tr>
<tr>
<td>8th task</td>
<td>Task C2−</td>
<td>Task C1+</td>
<td>Task C1−</td>
<td>Task C2+</td>
</tr>
<tr>
<td></td>
<td>BREAK (10 mins)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>TOEFL TEST</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
the entire data; (2) the number needed to be limited because of the work involved in transcription; (3) the data set is significantly larger than that reported in most previous studies, where transcripts of 10 students for each dimension are common. The data were transcribed with standard orthography and entered with the computer program from the CHILDES project (MacWhinney, 1995). The measures used in Skehan and Foster (1999) and Foster and Skehan (1996), with minor modifications, were adopted in analyzing the test discourse. These measures are as follows:

*Fluency*, operationalized as the number of repetitions, false starts, reformulations, and hesitations and pauses, divided by the total amount of speech

*Accuracy*, operationalized as the percentage of error-free clauses in the total number of clauses

*Complexity*, operationalized as the number of clauses divided by the number of c-units

*Inter-rater reliability* for coding, calculated by getting an independent researcher to code 10% of the data

The results for each of the categories coded are given in Table 4, in which percentage agreement between the two raters and $\kappa$ statistics are reported. Acceptably high levels of agreement were reached.

*Quantitative analysis of test ratings.* The second analysis was an IRT analysis of test ratings by trained raters using analytical rating scales for fluency, accuracy, and complexity specifically developed for the study (see Appendix). Performances of all 193 subjects were then rated against these categories and the scores analyzed by multifaceted Rasch analysis implemented by the FACETS program (Linacre, 1992; see McNamara, 1996, for a detailed explanation of this approach).
Results

Impact of Conditions on Discourse Measures

The research question posed in this study was whether a range of different performance conditions (involving assumed different levels of cognitive demand) in each dimension was associated with different levels of fluency, complexity, or accuracy in candidates’ test responses. The first approach to answering this question was to consider the results of a detailed discourse analysis of data from a subset of 36 students. A summary of the results is given in Table 5. No significant effect for any of the measures was found, with the single exception of an effect for accuracy in the immediacy dimension. Descriptive statistics of fluency, accuracy, and immediacy measures are given in Table 6.

Fluency

As explained earlier, five measures of fluency were used. As shown in Table 6, a relatively higher amount of repair fluency (reformulation, repetition, and pauses) was observed than false starts and hesitations. Results of the statistical analysis (repeated
Table 5

Summary of findings from the analysis of candidate discourse

<table>
<thead>
<tr>
<th>Task dimension</th>
<th>Performance condition</th>
<th>Task version</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Perspective</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Accuracy</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Complexity</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td><strong>Immediacy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Accuracy</td>
<td><em>p &lt; .048</em></td>
<td>ns</td>
</tr>
<tr>
<td>Complexity</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td><strong>Adequacy</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Accuracy</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Complexity</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td><strong>Planning time</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fluency</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Accuracy</td>
<td>ns</td>
<td>ns</td>
</tr>
<tr>
<td>Complexity</td>
<td>ns</td>
<td>ns</td>
</tr>
</tbody>
</table>

*Note. ns = not significant.*

Table 6

Performance conditions and task performance in Immediacy dimension task: Descriptive statistics for fluency, accuracy, and complexity (mean, SD)

<table>
<thead>
<tr>
<th>Condition</th>
<th>False starts</th>
<th>Hesitation</th>
<th>Reformulation</th>
<th>Repetition</th>
<th>Pauses</th>
<th>Accuracy</th>
<th>Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td>+</td>
<td>.02</td>
<td>.04</td>
<td>.18</td>
<td>.22</td>
<td>.17</td>
<td>74.26</td>
<td>1.45</td>
</tr>
<tr>
<td></td>
<td>(.03)</td>
<td>(.07)</td>
<td>(.14)</td>
<td>(.22)</td>
<td>(.16)</td>
<td>(13.82)</td>
<td>(.21)</td>
</tr>
<tr>
<td>−</td>
<td>.01</td>
<td>.04</td>
<td>.12</td>
<td>.22</td>
<td>.19</td>
<td>67.66</td>
<td>1.41</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.06)</td>
<td>(.11)</td>
<td>(.23)</td>
<td>(.20)</td>
<td>(17.88)</td>
<td>(.24)</td>
</tr>
</tbody>
</table>
measures MANOVA) revealed no effect of performance conditions and task version, nor any interaction of performance condition and task version for the five fluency measures (Table 7). The results, however, should be interpreted with caution, because the data used for analysis were transformed twice—once to take account of varying lengths of utterance, and a second time to compensate for lack of homogeneity in variance.

**Accuracy**

The measure of accuracy used for test candidates’ task performance was the percentage of error-free clauses in the total number of clauses. On the basis of Skehan’s work (e.g., Foster & Skehan, 1996; Skehan & Foster, 1997, 1999), it was assumed that easier tasks (– performance condition in Table 2) would be associated with more accurate speech than harder tasks (+ performance condition in Table 2). However, in terms of raw figures, as presented in Table 6, candidates produced more accurate speech under the condition presumed to be more difficult.

Two-way repeated measures ANOVA analyses showed that a significant effect was found only for the performance conditions in the Immediacy dimension (Table 8). It should be noted, moreover, that the only significant result was in the opposite direction to that hypothesized.

Table 7

*Impact of performance conditions on fluency measures for each task dimension (two-way repeated measures MANOVA, Wilks test)*

<table>
<thead>
<tr>
<th>Source</th>
<th>Value</th>
<th>Hypothesized df</th>
<th>Error df</th>
<th>p</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Immediacy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td>.78421</td>
<td>1.651</td>
<td>5.00</td>
<td>.177</td>
<td>.216</td>
</tr>
<tr>
<td>Condition</td>
<td>.81784</td>
<td>1.336</td>
<td>5.00</td>
<td>.276</td>
<td>.156</td>
</tr>
<tr>
<td>Condition * version</td>
<td>.84364</td>
<td>1.112</td>
<td>5.00</td>
<td>.375</td>
<td>.182</td>
</tr>
</tbody>
</table>
Complexity was measured by calculating the number of clauses per c-unit. Contrary to our hypothesis, we found that raw complexity measures were higher under the more demanding condition for the Perspective, Immediacy, and Adequacy dimensions. Statistical analysis (two-way repeated measures ANOVA), however, revealed no significant effect of performance conditions, task versions, or interaction of performance condition and version across the four task dimensions (Table 9).

### Impact of Conditions on Operational Ratings

The above analysis essentially addressed the question of the replicability of the findings of previous research (i.e., Skehan and Foster, 1997, 1999; Foster and Skehan, 1996) when data were gathered under test conditions rather than under conditions resembling those of classroom interaction. Whatever the results may have been, this is not a feasible approach in actual testing situations, because it requires time-consuming discourse analysis. The usefulness of Skehan’s approach in a testing context depends on its being able to be realized operationally, that is, with real time ratings. To this end, rating scales were developed for each of the rating categories (fluency, accuracy, complexity) (see Appendix).
Performances of all 193 subjects were then rated in this way, and the scores were analyzed by multifaceted Rasch analysis implemented by the FACETS program (Linacre, 1992; see McNamara, 1996, for a detailed explanation of this approach and for interpretation of output from such analyses). First, four analyses were run, with “task version” identified as a separate facet. Then the same procedure was conducted again, this time identifying “performance condition” as a facet. In total, eight separate analyses were carried out, and, given space constraints, they are presented in summary form only here. Full details of these analyses are provided in McNamara, Elder, and Iwashita (1999).

The first set of analyses was designed to identify the impact of task version, or, in other words, to establish the equivalence of the two task exemplars for each dimension. Findings revealed a statistically significant difference between the two exemplars of Immediacy, but for all other dimensions the differences between task versions were nonsignificant, which indicated that the two task versions were making similar demands on candidates’ ability and could therefore be regarded as equivalent for measurement purposes.

The second set of FACETS analyses, which are the ones most pertinent to this study, yielded very similar findings to those derived from the discourse analyses reported above. The conditions under which tasks were performed did not, with one

<table>
<thead>
<tr>
<th>Source</th>
<th>Sum of squares</th>
<th>df</th>
<th>Mean square</th>
<th>F</th>
<th>p</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Immediacy</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Version</td>
<td>.11</td>
<td>1</td>
<td>.11</td>
<td>1.51</td>
<td>.228</td>
<td>.033</td>
</tr>
<tr>
<td>Condition</td>
<td>.03</td>
<td>1</td>
<td>.03</td>
<td>1.17</td>
<td>.287</td>
<td>.000</td>
</tr>
<tr>
<td>Condition * version</td>
<td>.00</td>
<td>1</td>
<td>.00</td>
<td>.01</td>
<td>.929</td>
<td>.042</td>
</tr>
</tbody>
</table>

Performances of all 193 subjects were then rated in this way, and the scores were analyzed by multifaceted Rasch analysis implemented by the FACETS program (Linacre, 1992; see McNamara, 1996, for a detailed explanation of this approach and for interpretation of output from such analyses). First, four analyses were run, with “task version” identified as a separate facet. Then the same procedure was conducted again, this time identifying “performance condition” as a facet. In total, eight separate analyses were carried out, and, given space constraints, they are presented in summary form only here. Full details of these analyses are provided in McNamara, Elder, and Iwashita (1999).

The first set of analyses was designed to identify the impact of task version, or, in other words, to establish the equivalence of the two task exemplars for each dimension. Findings revealed a statistically significant difference between the two exemplars of Immediacy, but for all other dimensions the differences between task versions were nonsignificant, which indicated that the two task versions were making similar demands on candidates’ ability and could therefore be regarded as equivalent for measurement purposes.

The second set of FACETS analyses, which are the ones most pertinent to this study, yielded very similar findings to those derived from the discourse analyses reported above. The conditions under which tasks were performed did not, with one
exception, have a systematic impact on the scores assigned by raters when judging the accuracy, fluency, or complexity of candidate responses. Again, the one exception was in the Immediacy dimension, where, contrary to what we hypothesized, the ratings were significantly higher (i.e., the task was easier) when candidates told the story without the pictures in front of them.

Two figures are produced here to illustrate the nature of the output yielded by the FACETS program. Figure 1 is a visual representation of the impact of performance conditions for the Perspective dimension. This and all other facets of the analysis (i.e., raters, candidates, rating categories) are positioned on a common measurement scale that facilitates comparisons across and within facets. This common scale appears as the first column in Figure 1. The next column shows the various raters involved in scoring candidate responses, with the harshest rater (ID 4) appearing at the top and the most lenient (ID 14) at the bottom. (As is typical in tests of speaking proficiency, the raters vary considerably in severity, but the FACETS program is able to adjust ability estimates accordingly.) The asterisks in Column 3 represent the candidates, who are spread across the ability continuum, with the most able ones at the top and the weakest at the bottom of the scale. The next column (fourth from the left) is the most interesting from the point of view of this study, because it shows the two performance conditions (plus and minus) that were hypothesized to differ from one another in the demand they place on candidate ability. In fact this proves not to be the case, as illustrated by the fact that the two conditions do not occupy different points on the measurement scale but are perfectly aligned with one another. The same is true for the items or rating categories (accuracy, complexity, and fluency) represented in the next column (although this finding is not relevant to the research question posed here). The right-hand column graphically describes the combined rating scales with the numbers 1 to 5, indicating the most likely scale score for each ability level.

Figure 2 presents an analogous map for the Immediacy dimension. In this case, it can be seen that the plus performance
conditions (represented in the center column) are at different levels with the minus condition (hypothesized to be less difficult than the plus condition) appearing higher up the scale (i.e., at the more difficult end of the ability continuum). This indicates what we have already reported, namely, that contrary to what was predicted, telling the story with the pictures in front of them (minus condition) proved to be more difficult for candidates than telling it after the pictures had been removed (plus condition).

**Figure 1.** FACET map, data from ratings of performances (Perspective).
The FACETS program also yields several statistical measures of the differences among the elements of each facet. One such measure is the fixed (all the same) $\chi^2$, which tests the null hypothesis that all elements of the facet are equal. Findings for each of the performance condition analyses are summarized in Table 10 below. Only the Immediacy dimension showed a significant difference in terms of difficulty.
In sum, we can state that the directional hypotheses underlying our operationalization of the performance conditions within each task dimension are not sustained by the analysis of either the candidate discourse or the subjective ratings assigned to their performance. Where systematic variation does occur, the trend is in the opposite direction from what was predicted.

Discussion

This study represents an attempt to apply in a test setting work in Second Language Acquisition, previously carried out in pedagogic settings, on the cognitive demands of oral communicative tasks. Partly, this study was motivated by a suggestion by Skehan (1998) that such an approach had potential for the assessment of oral proficiency. Skehan has also proposed that models of performance assessment should take into account the impact of task characteristics and performance conditions in estimates of the speaking ability of candidates. This study used data gathered under the conditions of a semidirect test in order to attempt to replicate the findings of previous research. Participants were required to produce oral narratives from picture prompts that had been designed to differ in their cognitive demands in ways that previous research had suggested would result in measurable differences in performance. These potential differences were investigated in two ways: (1) by counting discourse features

### Table 10

<table>
<thead>
<tr>
<th>Dimension</th>
<th>$\chi^2$</th>
<th>df</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perspective</td>
<td>.7</td>
<td>1</td>
<td>.40</td>
</tr>
<tr>
<td>Immediacy</td>
<td>20.5</td>
<td>1</td>
<td>.001</td>
</tr>
<tr>
<td>Adequacy</td>
<td>.8</td>
<td>1</td>
<td>.37</td>
</tr>
<tr>
<td>Planning</td>
<td>.1</td>
<td>1</td>
<td>.8</td>
</tr>
</tbody>
</table>

In sum, we can state that the directional hypotheses underlying our operationalization of the performance conditions within each task dimension are not sustained by the analysis of either the candidate discourse or the subjective ratings assigned to their performance. Where systematic variation does occur, the trend is in the opposite direction from what was predicted.
produced under the various conditions, through use of procedures familiar from the Second Language Acquisition literature; and (2) by subjectively rating the discourse features directly from performances, using trained raters. The first method of investigation was intended to tie the methodology of this research closely to the existing tradition of research in Second Language Acquisition; the second was intended to see whether a more feasible procedure for a large-scale test would yield similar results.

In general, the results failed to confirm the findings of existing research, although it should be noted that the finding for the Immediacy dimension does confirm the results of Robinson (1995), who found that accuracy increases rather than decreases when the task is performed under more cognitively demanding conditions (i.e., in the “there and then” mode, without the support of pictures). He explains that the “here and now” condition is context supported and involves a simple pragmatic mode requiring the interlocutor to fill the gap of linguistic resources from context, whereas the “there-and-then” condition requires learners to be more accurate because of the lack of contextual support (Robinson, 1995, p. 104). Robinson’s conceptualization of this finding is interesting, and the fact that his results are replicated here indicates that we need to think more carefully about what constitutes cognitive complexity in a task and how this is likely to manifest itself in candidate performance.

Apart from this one significant effect, we found no systematic discourse variation associated with the various task dimensions for performance conditions. The findings of the exact discourse analysis were matched by the judge-mediated data; that is, with one exception, no systematic variation was found.

It seems that, at least with simple picture-story tasks of the kind used in this study, it is not possible to specify task demands for speaking in a way that will lead to the definition of a scale of increasing mastery of speaking tasks. If we conceived of narratives as typical of all speaking tasks, we might conclude from this study that speaking is more of a unitary skill, deployed in similar ways across different task conditions. However, such a conclusion would,
in our view, be premature, both because it fails to accord with the findings of previous research and because the choice of task in this study, as already stated, was very much constrained by what we considered to be feasible in a tape-based test of speaking used for the purpose of generating large-scale standardized proficiency measures. In particular, for the category of complexity, the narrative task adopted in this research was insufficiently productive of complex sentence structure, a fact that raters commented on. This could be remedied by conceiving of narratives more broadly than has been done here and by involving candidates in recasting more complex input of the kind, for example, provided by television “soaps” containing dialogic sequences. Furthermore, there are many other task types used in tape-based oral proficiency testing, and replication of this study using different task types may yield more promising results. Robinson (in press-a) has discussed a map task whose demands may be varied in systematic ways, which could be explored in future research. Such a task could be carried out in a semidirect test format.

It should also be noted that there are almost certainly different ways of operationalizing the performance conditions under which tasks in this study were performed, and the lack of precise specification of what is meant by the various task dimensions and conditions can be regarded as a weakness of Skehan’s framework, at least in terms of its applicability to testing. A major problem we faced was that of achieving equivalence across task exemplars in all respects other than the condition we were manipulating. For example, the dimension of Familiarity of Task Information (see Table 1) in Skehan’s framework was operationalized in our study by manipulating the perspective from which the story was told (it being assumed that a first-person narrative, albeit based on a standard set of pictures, would allow the candidate to elaborate on the story by drawing on personal information, whereas when recounting the same story in the third person, the candidate would be unlikely to embellish the story to the same degree). We could think of no other way of manipulating the familiarity dimension without at the same time
changing other task input characteristics, but there may well be other means of implementing this dimension in a way that is compatible with an experimental research design. A similar problem was faced with the dimension of Adequacy, which we attempted to render cognitively complex by removing two pictures from the sequence of six provided to the candidates in the less cognitively demanding condition. We had some doubts about which pictures to remove (some may have made it more difficult to construct the story than others) and also about how many pictures we could take out without rendering the task impossible for candidates to perform. Although the decisions we arrived at were based on extensive piloting with both native speakers and nonnative speakers, it is nevertheless conceivable that a different solution might have generated a different outcome. Similarly, our decision to opt for 3.5 minutes’ planning time, while not arbitrary, in that it was based partly on the findings of previous research (reported earlier in this paper) and partly on the practicality constraints of the testing situation, may nevertheless have influenced our results. Greater amounts of planning time, perhaps in combination with a more elaborate narrative task of the kind mentioned above, may well have resulted in more fluent and accurate speech. However, the absence of any sign in the results of consistent differences in performance related to different elicitation conditions suggests that there are lacunae in the information-processing framework on which the study was based. Unless these problems are addressed, greater effort in the production of more varied tasks may not reverse the finding of this study. The inconclusive findings of Norris et al. (2000), whose research was based on similar assumptions, but who used a greater range of tasks and manipulated a larger number of variables than we have done here, does not give grounds for optimism.

An alternative explanation for the overall lack of consonance between the results of this study and those reported in previous SLA research is that the differences in testing and pedagogic contexts are so great as to alter the cognitive focus of the tasks. In an interactive task carried out with a classmate, the focus will be
on the completion of the task. In a test, where tasks are carried out alone in a computer-mediated environment and hence lack an interactive dimension, the cognitive focus may be on display, and this may alter the relation between task characteristics and language output. For example, a focus on accuracy may be paramount in the testing situation regardless of the conditions under which the task is performed, and this in turn may affect the fluency and complexity of candidates’ speech. Delivery may be halting whether the task is easy or difficult, because the candidates are focusing primarily on correctness. The lack of complexity in candidates’ production may likewise be due to their anxiety about how their speech is being evaluated, making them reluctant to venture beyond what they know how to say properly even when the task conditions allow for this. If this is the case, it of course casts doubt on the validity of the testing of speaking in a semidirect format and indeed the validity of oral proficiency testing more generally as an index of “real-life” communication.

Clearly, much further work is needed to clarify the significance of the findings of this project for work in the field of language testing on performance conditions and their interaction with different task characteristics. The present study has provided important evidence regarding the applicability of this work to the semidirect testing of speaking, or more specifically to monologic narrative tasks. The broader contribution of this research to both language testing and to second-language acquisition has yet to be fully explored, but the possibilities are intriguing.

Revised version accepted 07 March 2001

Notes

1The notion that visual support is necessarily facilitative of task performance has, however, been called into question by Gruba (1997), at least with respect to listening comprehension activities.

2This research is continuing, with somewhat mixed results (cf. Norris et al., 2000).

3There is, however, some disagreement as to whether more difficult tasks can indeed be expected to elicit more complex language. Skehan and Foster
(1999), for example, found that more complex language was produced under a less cognitively demanding task condition.

4Most students lived with Australian host-families, and so unless their host families spoke the students’ native language, they had to speak English at home.

5As for the TSE, see http://www.toefl.org/edpubs.html#user-man

6We used the term “task dimension” to refer to task characteristics, as described in the literature (e.g., Skehan, 1996, 1998) and “performance condition” to refer to the condition imposed on test candidates within each dimension. The “+” condition is the one hypothesized to be more cognitively demanding, and the “-” condition less cognitively demanding.

7A c-unit was defined as a simple clause, or an independent subclausal unit, together with the subordinate clauses associated with them (Foster, Tonkyn, & Wigglesworth, 2000).

8Relatively large SDs are routinely observed in discourse studies of individual performances (e.g., Skehan & Foster, 1999).

9Bartlett tests of sphericity indicated that the five fluency measures (i.e., false starts, hesitation, pause, reformulation, and repetition) were correlated.

10As mentioned above, the MANOVA analysis was performed four times separately (once for each task dimension). Since all the subjects performed all the tasks across the task dimensions, the results of each task are somewhat related, and so the alpha level for each result would have needed to be adjusted if any significant effects for performance conditions or task versions had been identified. In the event, this issue did not arise.

11The exception was for Planning Time, where it was assumed that having planning time (+ condition) would be easier than performing the task without any preparation time (– condition).

12The units of measurement are called “logits.” The average item difficulty is set by convention at zero.

References


**Appendix:**

**Rating Scales**

**Fluency**

5. Speaks without hesitation; speech is generally of a speed similar to a native speaker.

4. Speaks fairly fluently with only occasional hesitation, false starts and modification of attempted utterance. Speech is only slightly slower than that of a native speaker.

3. Speaks more slowly than a native speaker due to hesitations and word-finding delays.

2. A marked degree of hesitation due to word-finding delays or inability to phrase utterances easily.

1. Speech is quite disfluent due to frequent and lengthy hesitations or false starts.

**Accuracy**

5. Errors are barely noticeable.
4 Errors are not unusual, but rarely major.

3 Manages most common forms, with occasional errors; major errors present.

2 Limited linguistic control; major errors frequent.

1 Clear lack of linguistic control even of basic forms.

**Complexity**

5 Confidently attempts a variety of verb forms (e.g., passives, modals, tense, and aspect), even if the use is not always correct. Regularly takes risks grammatically in the service of expressing complex meaning. Routinely attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is occasionally awkward or incorrect.

4 Attempts a variety of verb forms (e.g., passives, modals, tense, and aspect), even if the use is not always correct. Takes risks grammatically in the service of expressing complex meaning. Regularly attempts the use of coordination and subordination to convey ideas that cannot be expressed in a single clause, even if the result is awkward or incorrect.

3 Mostly relies on simple verb forms, with some attempts to use a greater variety of forms (e.g., passives, modals, more varied tense and aspect). Some attempt to use coordination and subordination to convey ideas that cannot be expressed in a single clause.

2 Produces numerous sentence fragments in a predictable set of simple clause structures. If coordination and/or subordination are attempted to express more complex clause relations, this is hesitant and done with difficulty.

1 Produces mostly sentence fragments and simple phrases. Little attempt to use any grammatical means to connect ideas across clauses.