Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?

Catherine Elder, Noriko Iwashita and Tim McNamara

Language Testing 2002; 19; 347
DOI: 10.1191/0265532202lt235oa

Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer?

Catherine Elder *University of Auckland*, **Noriko Iwashita** and **Tim McNamara** *University of Melbourne*

This study investigates the impact of performance conditions on perceptions of task difficulty in a test of spoken language, in light of the cognitive complexity framework proposed by Skehan (1998). Candidates performed a series of narrative tasks whose characteristics, and the conditions under which they were performed, were manipulated, and the impact of these on task performance was analysed. Test-takers recorded their perceptions of the relative difficulty of each task and their attitudes to them. Results offered little support for Skehan's framework in the context of oral proficiency assessment, and also raise doubts about *post hoc* estimates of task difficulty by test-takers.

I Introduction

The current popularity of performance testing as opposed to multiple-choice and other discrete-point item types has resulted in a growing interest in tasks as a vehicle for assessing learner ability. Task-based assessment requires the test-taker to engage in the performance of tasks which simulate the language demands of the real world situation with the aim of eliciting an "authentic" sample of language from the candidate. The properties of such tasks and the influence of these properties on learner performance are now being widely researched, with some scholars focusing on strengthening the links between test tasks and their real world counterparts (e.g., Bachman and Palmer 1996; Douglas, 2000) and others on the effect on candidate production of manipulating different task characteristics in the test situation (e.g., Norris *et al.*, 1998; Norris *et al.*, 2000; Slatyer *et al.*, 2000). The present study falls within this latter tradition and attempts to operationalize Skehan's (1998) framework of task complexity – derived from second language research (SLA) research conducted in a second language (L2) classroom environment – in the context of

Address for correspondence: Noriko Iwashita, Language Testing Research Centre, The University of Melbourne, Victoria, 3010, Australia; email: norikoi@unimelb.edu.au

 10.1191/0265532202lt235oa

a semi-direct test of speaking. The study also explores the issue of test-taker perceptions of task difficulty and the extent to which these perceptions correspond to the assumptions of task complexity underlying the Skehan framework, on the one hand, and to the quality of task performance, on the other.

II Estimating task difficulty

One of the challenges facing those concerned with gauging the influence of task characteristics and performance conditions on candidate performance is how to determine the difficulty of tasks. A greater understanding of the factors affecting task difficulty could assist in the choice of a suitable range of tasks for assessment purposes and also has the potential to influence the way levels of test performance are described. Some research on this issue has been initiated by SLA researchers (e.g., Robinson, 1995; 1996; 2001; Skehan, 1996; 1998), and a number of factors that appear to influence task performance in the classroom have been identified (although these have differed to some extent among researchers).

Only recently have attempts been made to apply insights from SLA research to explore the issue of task difficulty in the testing situation. Wigglesworth (1997), for example, explored the effects of planning time on the discourse produced by test candidates and found that one minute of pre-task planning resulted in measurable improvements in the complexity, fluency and accuracy of their speech (although, interestingly, this was not reflected in scores assigned by raters). In a subsequent study Slayter *et al.* (2000) examined the role played by certain key variables in the design 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.

The difficulty factors in the Slatyer *et al*. study were, however, identified *post hoc* rather than on the basis of a pre-existing taxonomy or framework as is the case with the current study. One such framework has been developed by Robinson (2001; in press) who identifies two sets of factors as contributing to task complexity. These are "resource-directing" factors (e.g., number of task elements, reasoning demands of the task, immediacy of information provided) and "resource-depleting" factors (e.g., planning time, number of tasks, prior knowledge) (Robinson, 2001: 30). Robinson claims that by manipulating these factors, the cognitive demand (e.g., amount of attention, memory, reasoning and other information processing) required for task performance will vary, leading to variation in the quality of language produced. A somewhat different model has been

proposed by Skehan (1998) who defines task difficulty in terms of three different factors:

- code complexity: incorporating both linguistic complexity/variety and vocabulary load/variety;
- cognitive complexity: involving cognitive processing factors such as information type and organizational structure as well as the familiarity of task topic discourse and genre; and
- communicative stress: referring to the logistics of task performance e.g., time pressure, nature of the prompt and number of participants.

Skehan and his colleagues (e.g., Foster and Skehan, 1996; Skehan and Foster, 1997; 1999) have proposed that more complex tasks direct learners' attention to context and divert attention away from form. Simple tasks therefore generate more fluent and more accurate speech, as opposed to more complex tasks which generate more complex speech at the expense of accuracy and fluency.

Both Skehan and Robinson claim that their respective models have the potential to reveal the precise nature of the mediation that occurs between any underlying abilities and the way a task is transacted. Such frameworks would appear to hold considerable promise for language testing in so far as they allow us to make predictions, and therefore to select and sequence test tasks according to their difficulty (i.e., the challenge they pose to test candidates) by manipulating a number of the factors mentioned above.

Work by Norris *et al*. (1998; 2000) and Brown *et al*. (1999) has built specifically on Skehan's framework in the design of syllabus-related performance assessments. Skehan's dimensions of cognitive demand have been adopted and modified in the design of specific tasks. However, Norris and his colleagues have found only moderate 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.[1] One possible explanation, which would need to be explored empirically, is that learner factors – such as anxiety, confidence and motivation – produce different levels of stress and engagement during task performance and that, as Slatyer *et al*. (2000) conclude, these interact in complex ways with the characteristics of the tasks themselves. Accordingly, following previous work on tasks (e.g., Brindley, 1987; Nunan, 1989), Robinson has proposed the independence of the

[1]This research is continuing, with somewhat mixed results (cf. Norris *et al*., 2000).

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 learners (2001; in press).

III Test-taker perceptions

Given the potential impact on performance of learner perceptions of task difficulty it seems that, in the testing situation, there may be some value in canvassing test-takers' perceptions of task difficulty to determine how influential these are in test performance. Furthermore, if test-takers can predict what makes a task difficult, it may be wise for us to access their views during the test design stage to determine whether they correspond to the hunches of test-developers and with existing theories about what makes a task more or less complex. It is conceivable that test-takers may be able to identify additional features of the task, or additional challenges involved in performing such tasks other than those visible to the test-developer or to the rater.

While learner perceptions are accorded a central place in many SLA studies, this tends not be the case in the field of language testing. Test-taker reactions have traditionally been associated with face validity, or "appearance of validity" and are therefore not seen as central to the test validation process (see, e.g., Bachman, 1990). The traditional view has been that test validation is more properly left to experts with relevant training in test development and analysis. Interest in how test-takers feel about a test is usually motivated by a desire to ensure that the test is acceptable to its users, and therefore that its results are taken seriously by all concerned (Davies *et al.*, 1999). It has nevertheless been pointed out that if test-takers have negative attitudes to the test then they are less likely to perform to their best of their abilities. This has obvious implications for test validity. If test attitudes interfere with test performance this may result in unwarranted inferences being drawn from test scores (see, e.g., Nevo, 1985; Spolsky, 1995; Elder and Lynch, 1996). Messick (1989) in fact explicitly recommends including test-taker perceptions as a crucial source of evidence for construct validity.

Of relevance to the present study is research comparing test-taker reactions to different test formats or task types. This research shows clearly that test-takers have preferences for certain types of test and that some tasks are perceived to be easier, more interesting or more acceptable as measures of ability than others (see, e.g., Shohamy, 1982; Scott, 1986; Zeidner and Bensoussan, 1988; Bradshaw, 1990; Zeidner, 1990). Research focusing specifically on semi-direct oral tasks of the kind used in the research reported in this article is of particular interest. Findings indicate that test-takers tend to find this

format more difficult and/or more stressful than the live interview situation (see, e.g., Clarke, 1985; Stansfield *et al.*, 1990; McNamara 1990; Stansfield, 1991; Brown, 1993; Hill, 1998), although opinions were not in all cases uniform across different kinds of test-taker. Interestingly, in Hill's (1998) study, the factor most strongly associated with perceptions of test difficulty in the tape-based format was preparation time. It might therefore be hypothesized that providing greater amounts of time for pre-task planning on a tape-based oral test will minimize stress and result in a reduction in the perceived difficulty of test tasks. Brown's (1993) informants attributed difficulty of the tape-based format to a range of different factors, including inadequate response time, unfamiliar vocabulary, speed of voices on the tape, lack of clarity in instructions, unclear prompts, too much input material to process and lack of familiarity with the task type. While this kind of information may be useful in making revisions to a test (and, on this issue, see also Alderson, 1988; Kenyon and Stansfield, 1991) her findings suggest that in practice it may be difficult to separate features of the task and attributes of the candidate in any operationalization of task difficulty.

The issue of whether learner perceptions of tasks are related to actual task performance has been explored by a number of the above researchers, with the majority of studies showing a clear relationship between the two. It should, however, be noted that while some focus specifically on perceptions of task difficulty, most deal with attitudes to the task more generally. Scott and Madsen (1983) showed that learners with low levels of proficiency rated oral interview tasks less favourably than did more proficient learners. Iwashita and Elder (1997) found that language proficiency was a more powerful factor than any other background variable in determining their participants' reactions to the listening component of a Japanese proficiency test for teachers. Brooks (1999) likewise noted a relationship between attitudes towards different assessment types (portfolio presentations and class participation) and levels of performance on the respective tasks. Shohamy (1982), Zeidner (1988; 1990), Bradshaw (1990) and Brown (1993) all found significant relationships between scores obtained by candidates and their attitudes to one or more features of the test task, with weaker candidates tending to respond less positively than the those with higher levels of proficiency. Owing to the fact that in all the above cases attitudes to the test were canvassed following its administration, it is difficult to interpret the meaning of the attitude/score relationship. Attitudes may be influenced by the experience of taking the test (with those performing well feeling more positive about the experience and vice versa for less proficient

students), or test results may be determined by attitudes (i.e., perceptions of the test-taking experience may have a facilitating or a debilitating effect on test performance). If the latter proves to be true, then it would seem appropriate to include learner perceptions of and attitudes towards the task as a factor in any model of task difficulty.

A study which is not dissimilar in purpose and design (albeit more limited in scope) to the research reported below is that of Robinson (2001) in so far as it compares reactions of ESL learners to two different versions of a single task (simple and complex). The findings showed that:

- task complexity, as operationalized in his study, has a significant influence on various aspects of learner production;
- task complexity is significantly associated with learner perceptions of task difficulty; and
- learners' affective responses are related to certain aspects of their test performance.

It remains to be seen whether these findings are replicable with other kinds of tasks and in the more constrained context of a semi-direct oral test of speaking.

The current study differs from many of those reviewed above in that our operationalization of task complexity involves the manipulation of one variable at a time, i.e., we have taken four different dimensions of task complexity (perspective, immediacy, adequacy and planning time) and investigated each of these in turn. In the present study, three research questions are addressed.

- Are hypothesized differences in task complexity associated with differences in task difficulty as reflected in scores assigned to learner performance?: Here our aim is to test the applicability of the Skehan framework in the test environment. Implicit in this framework is the hypothesis that simple tasks will be easier for test-takers than more complex ones, and will therefore yield relatively higher mean scores at least for fluency and accuracy, if not for complexity.
- Are hypothesized differences in task complexity associated with differences in test-taker attitudes and perceptions of task difficulty?: Our hypothesis is that complex tasks (operationalized in terms of Skehan's framework) will be perceived to be more difficult than simple ones, and that attitudes towards a task may also be affected by perceptions of its difficulty.
- Are differences in test-taker attitudes and perceptions of difficulty associated with actual differences in task difficulty as reflected in scores assigned to learner performance?: Here it is hypothesized,

in keeping with the findings reported from the literature on test-taker reactions, that learner proficiency estimates, as reflected in actual scores assigned to task performance, will be negatively associated with learner perceptions of task difficulty.

IV Methodology

1 Speaking tasks

Speaking tasks used in the study involved a single type of stimulus, of the kind used routinely in the Test of Spoken English (TSE), namely: a narrative task based on a sequenced set of picture prompts. A number of pilot narrative tasks were developed as means of operationalizing different dimensions of the Skehan model. Through a complex process of materials development, pre-piloting on native and nonnative speakers, teacher/researcher workshops and expert consultation, it was agreed that four dimensions (immediacy, adequacy, perspective and planning time) were the ones that lent themselves most readily to experimental manipulation. The 'immediacy' dimension, for example, could be easily manipulated by asking candidates to tell a story with a set of pictures in front of them, or to tell the same story after the pictures had been removed. Likewise comparable groups of candidates could tell the same story with and without preparation time (planning dimension) and with and without the provision of a complete sequence of pictures (adequacy dimension). The 'perspective' dimension could also be manipulated easily by asking candidates to tell a story from their own and then from another person's point of view.

The rationale for varying the performance conditions within each dimension was that this would either make the tasks easier (i.e., less cognitively demanding) or more difficult (i.e., more cognitively demanding) for the candidates. There were two performance conditions – labelled plus (+) or minus (–) according to their predicted difficulty for the candidates – within each dimension.[2] In order to investigate the effect of specificity of task and the generalizability of the experimental condition across task exemplars (see Table 1), two exemplars of each task (1 and 2) were used in each experimental condition. Thus, 8 different story tasks were developed for this experiment, and two different performance conditions for each story. All tasks were piloted, and expert judgements were canvassed in an

[2]We used the terms 'task dimension' to refer to task characteristics as described in the literature (e.g., Skehan, 1996; 1998) and 'performance condition' to refer to the conditions within each dimension imposed on test candidates.

Table 1 Design of tasks

Dimension	Predicted difficulty (according to assumed degree of cognitive demand)	
	+	–
Perspective	Tell a story from someone else's point of view	Tell a story as it happened to you
Immediacy	There and then (without pictures)	Here and now (with pictures)
Adequacy	Tell a story from a set of five pictures (the third picture in the set is missing)	Tell a story from a set of six pictures
Planning time	30 seconds for reading instruction and looking at the pictures and 3 minutes planning time	30 seconds only for reading instruction and looking at the pictures

Note: For the perspective, immediacy and adequacy dimensions, the + condition was assumed to be more difficult than the – condition, whereas for the planning dimension, the + condition was expected to be easier than the – condition.

attempt to ensure that the two exemplars of each task dimension resembled each other as closely as possible in terms of their linguistic demands and likely level of familiarity to the test-takers. In the actual study, each participant performed 8 different narrative tasks, representing each of the four dimensions in both the 'easy' (–) and 'difficult' (+) condition (4 × 2). No student told the same story twice.

2 *Participants*

The study consisted of 201 student participants. The majority (80–90% of the participants) was currently enrolled in an ESL course in Melbourne to prepare for study at university in Australia, while the remainder was already studying at a tertiary institution in Melbourne. The mean age of the participants was 21.6 years (*SD* 4.5), and the mean length of residence in Australia was 4.3 months. The first language of participants 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; many had also studied foreign languages other than English at some time. Most participants spoke English at home in Australia. (Most students lived with Australian host families, and so unless their host families spoke the students' native language, they had to speak English at home.) The mean score of the participants on the Institutional version of the TOEFL test was 493.1, with a *SD* of 45.8 and a range of 427 to 670.

3 Data collection

The speaking test was administered in a university language laboratory. All participants were randomly assigned to one of four experimental groups. They completed a language background questionnaire, followed by the speaking test. They then took the institutional version of the TOEFL test (so that we could control for any differences in ability from one group to the other). The speaking test was made up of 8 narrative tasks (3 minutes maximum for each task) with participants granted a 10-minute break after the first 4 tasks. All participants had the experience of telling 2 stories for each dimension, one in the + condition and one in the – condition. The order of presentation of the 4 task dimensions and of the + and – conditions was counterbalanced across the 4 experimental groups. After completing each task, test-takers completed a one-page questionnaire on their perceptions of the task.

4 The data

a Quantitative analysis of test ratings: Performances by all 201 participants were rated using analytical rating scales for fluency, accuracy and complexity specifically developed for the study (Appendix 1). In total, 14 raters were recruited for the assessment of the speaking tasks. All raters had some experience in rating speaking tests (e.g., IELTS, Occupational English Test for medical professionals, TSE) as well as teaching ESL at a level similar to the participants in the present study. Before assessing the speaking tasks, all raters participated in a rater training session, and then were asked to rate sample tasks for accreditation. Each of the 201 performances received two independent ratings, from any pair of raters drawn from the pool of 14.

These data were analysed using the IRT-based program FACETS (Linacre, 1992) in order to determine whether there was any impact of the imposed conditions on the scores assigned by raters to task performance. As distinct rating scales with distinct wording were used for each of the aspects of performance being assessed (fluency, accuracy and complexity), it was decided to use the Partial Credit model for these three aspects, or items. (In the analysis that follows, each separate aspect of performance is referred to as a test 'item', scored on a 5-point scale.) The Rating Scale model was used for judges. The quality of rater judgements needed to be controlled, as raters needed to achieve consistent judgement for stable measures of item difficulty to be achieved. This was achieved by examining the fit statistics provided by Rasch measurement, which summarize the consistency of

the measurement of facets. The fit statistic used to evaluate rater consistency was Infit Mean Square, with an expected value of 1 and an acceptable range of 2 standard deviations around the mean (McNamara, 1996: 181). Where the 'fit' or consistency of judges was outside this range, their ratings were eliminated from the analysis and the data re-analysed; this was an iterative process, until only raters showing acceptable levels of consistency were left. The impact of the imposed conditions was evaluated by means of a *t*-test for differences in the estimates of the difficulty presented by each condition.

b Questionnaire feedback: All participants were asked to complete a questionnaire after each task. This contained questions about their perceptions of the difficulty of each task, their familiarity with this type of task and their attitudes towards it (defined in our study as enjoyment). Answers were given on a five-point Likert scale. In the analysis of responses, answers were coded on a scale of 1 to 5, with 5 representing the most favourable response (easiest, most liked) and 1 the least favourable (hardest, least liked).

- Perceptions of the difficulty of the task:
  Q1 Did you find the task easy or difficult?
- Attitudes towards each test task:
  Q5 I enjoyed telling the story

The above questions were designed to tap test-takers' reactions to the story-telling experience, without drawing their attention to the particular condition under which it was performed.

First of all, mean scores were calculated for each question under the two conditions in each task; *t*-tests were then carried out to examine whether the task conditions had a significant impact on the test candidates' perception of task difficulty and familiarity and their level of enjoyment.

In addition, further analyses were carried out to investigate the relationship between the test scores and learner perceptions using the results (in logits) yielded from FACETS analysis and the questionnaire feedback (Q1 and Q5). A series of rank correlations (Kendall's tau) were performed to determine whether there was an association between test-takers' ability as indicated by their overall performance across tasks and their perceptions of the difficulty and familiarity of each individual task and their task enjoyment. Open-ended questionnaire responses were cross-referenced to the quantitative analyses in the hope that this qualitative feedback might shed further light on the findings.

Table 2 Impact of performance conditions on ratings assigned to candidate responses in each task dimension

Dimension	Condition	Measure (logits)	Standard error	*t*	*p*
Perspective	+	−0.03	0.06	0.50	ns
	−	0.03	0.06		
Immediacy	+	−0.19	0.06	3.17	<.01
	−	0.19	0.06		
Adequacy	+	0.05	0.06	0.83	ns
	−	−0.05	0.06		
Planning	+	−0.04	0.06	0.67	ns
	−	0.04	0.06		

V Results

1 Are hypothesized differences in task complexity associated with differences in task difficulty as reflected in scores assigned to learner performance?

The scores of the 201 task performances were analysed using multi-faceted Rasch analysis (Linacre, 1992; for a detailed explanation of this approach and for interpretation of output from such analyses, see McNamara, 1996). The FACETS program (Linacre, 1992) was run four times for each task dimension, separately identifying 'performance condition' as a facet. Results are presented in Table 2. The table provides estimates (in logits) of the difficulty associated with each of the performance conditions for each task dimension, together with the standard errors of those estimates. In order to evaluate whether the measures in each case for each dimension are significantly different, a *t* value for the difference is provided.[3] It can be seen that for three (perspective, adequacy, planning) of the four dimensions the two performance conditions are not associated with significant differences in difficulty. There is a significant difference associated with the conditions in the 'immediacy' dimension – telling the story without the pictures present (–0.19) and telling it with the pictures (+0.19) – but the size of the impact on 'immediacy' is modest, under 0.4 logits or 0.1 of a score point.

It is worth noting at this point that the discourse analysis of candidate performance that has been reported in detail elsewhere (Iwashita *et al.*, 2001) was found to parallel the above analyses, with the only significant difference in the quality of candidate production emerging

[3]The formula for calculating *t* from these estimates and their standard errors is given in Linacre (1992: 17).

again in the 'immediacy' dimension, where performance was superior (there was a higher incidence of error-free clauses) when candidates told the story without the pictures in front of them. On the bases of these findings our hypothesis that posited differences in task complexity would be reflected in actual differences in task performance cannot be confirmed.

2 Are hypothesized differences in task complexity associated with differences in test-taker attitudes and perceptions of task difficulty?

The mean scores and *SD*s of candidates' response to their perception of task difficulty (Q1) and task enjoyment (Q5) under the two conditions are given in Table 3. In general, very little difference between task conditions was observed. However, it is worth noting that in the 'immediacy' dimension the mean scores for perceptions of task

Table 3 Perceptions of task difficulty and attitude to task (mean score and SD)

Task dimension and version		Task difficulty (Q1)				Task enjoyment (Q5)			
		+	–	*t*	*d*	+	–	*t*	*d*
Perspective									
1	M	3.00	2.93	0.55	0.08	2.95	2.92	0.22	0.03
	SD	0.94	0.83	$p = 0.57$		0.98	0.87	$p = 0.82$	
	n	97	100			97	99		
2	M	3.01	2.99	0.15	0.02	2.56	2.55	0.11	0.02
	SD	0.96	0.95	$p = 0.88$		0.94	0.89	$p = 0.91$	
	n	100	97			98	97		
Immediacy									
1	M	2.80	2.69	0.91	0.13	3.04	2.72	2.28	0.33
	SD	0.84	0.92	$p = 0.36$		1.03	0.94	$p = 0.02$	
	n	97	100			97	100		0.16
2	M	3.04	3.17	0.98	0.14	3.02	3.18	1.14	
	SD	0.84	0.95	$p = 0.32$		0.96	0.94	$p = 0.25$	
	n	99	96			100	97		
Adequacy									
1	M	2.32	2.43	1.04	0.15	2.81	2.85	0.26	0.04
	SD	0.81	0.67	$p = 0.30$		1.00	0.90	$p = 0.79$	
	n	97	100			97	100		
2	M	2.67	3.18	3.54	0.50	2.77	3.09	2.37	0.34
	SD	1.06	0.95	$p = 0.01$		0.99	0.93	$p = 0.02$	
	n	99	97			99	96		
Planning									
1	M	2.74	2.65	0.070	0.10	2.34	2.35	0.06	0.01
	SD	0.95	0.91	$p = 0.49$		0.74	0.78	$p = 0.95$	
	n	97	100			96	100		
2	M	2.47	2.52	0.36	0.05	2.80	3.01	1.56	0.22
	SD	0.75	0.82	$p = 0.72$		0.91	0.99	$p = 0.12$	
	n	99	97			99	97		

difficulty (Q1) in Task 1 and task enjoyment (Q5) in both Tasks 1 and 2 are slightly higher in the + condition (i.e., when the pictures were removed) than in the – condition (i.e., when they had the pictures in front of them). These differences were significant for Task enjoyment (Q5) only ($t = 2.28$, $p = 0.02$) but the effect size is small ($d = 0.33$). In the 'adequacy' dimension, in Task 2, candidates perceived the task to be significantly easier when they had a full set of six pictures ($t = 3.54$, $p = 0.01$, $d = 0.50$) than when two pictures were missing. They also found this task significantly more enjoyable under the same condition (i.e., when all six pictures were present) than when two pictures were missing ($t = 2.37$, $p = 0.02$, $d = 0.34$). Again, the effect sizes for these latter findings were modest. For the Planning dimension, in Task 1 test candidates found telling the story slightly easier when they had three minutes of planning time than when they did not. This difference was not however significant and the reverse trend (also nonsignificant) was observed for Task 2. Likewise for 'perspective', performance conditions made no difference to candidates' perceptions of task difficulty or enjoyment.

3 Are differences in test-taker attitudes and perceptions of difficulty associated with actual differences in task difficulty as reflected in scores assigned to learner performance?

The correlation analyses in the table indicate that, with regard to attitudes (Question 5) there is a nonsignificant relationship between enjoyment of the story-telling experience and level of proficiency as measured by scores assigned to task performance. This finding is consistent across all dimensions and all task exemplars.

The relationship between performance and perceptions of difficulty (Question 1) is somewhat more complex. The data (i.e., questionnaire response and task performance) reveal that for all four dimensions there is a significant relationship between perceptions of task difficulty and task performance, but that this relationship is not consistent across task exemplars. For 'perspective', those who did Task 2 in the + condition (i.e., telling the story from someone else's point of view) and found it difficult were more likely to be the low scoring candidates ($tau = .164$, $p = .034$). For 'immediacy' there was again a significant relationship between perceptions of difficulty and scores for Task 1, but this applied to those who performed the task without looking at the pictures (+ condition) ($tau = .154$, $p = .048$). For 'adequacy' there was a relationship between perceptions of difficulty and performance for those who performed Task 1 and Task 2 in the + condition (with an incomplete set of pictures) ($tau = .181$, $p = .045$; $tau = .18$, $p = .027$). For 'planning', significant relationships were

found in both + and – conditions in Task 1 only (*tau* = .193, p = .017; *tau* = .204, p = .008). Given that all correlations are weak and the relationship between scores and perceptions of difficulty is not generally consistent across task exemplars, our hypothesis regarding the relationship between task performance and test-taker attitudes and perceptions of task difficulty cannot be sustained.

Qualitative feedback from test-takers confirms what was revealed by the quantitative analyses; namely, the lack of any systematic relationship between task difficulty and hypothesized task complexity, on the one hand, and actual test performance, on the other.

In general, learner comments about the impact of the task conditions (+/–) in the 'planning' dimension corresponded with our initial predictions about their impact. In other words, having 3 minutes of planning time (+ condition) was generally seen as making it easier to tell a story than having no planning time (– condition), regardless of proficiency. Thus, a high scoring candidate (ranked 30) commented that 'with planning time, I could put ideas together and wrote them down' and a similar comment was made by a low performer (ranked 144). The fact that these comments were not reflected in actual performance differences under the two different conditions may have to do with individual differences in the quality of planning undertaken.

In the 'perspective' dimension, the learner who was ranked 165 found telling a story as it happened to her was easier and gave the following explanation: 'Because if it happens to me I could understand what my feeling [is].' A similar remark was made by a high scorer whose performance was ranked 9: 'Because we can express our feeling about what we feel, etc.' This explanation corresponds to our initial assumption that telling the story as it happened to them would be easier than telling it from another person's perspective. However, as found in the quantitative analysis of questionnaire responses, it appears that these perceptions about the relative facility of the – condition were shared by only a portion of the candidature.

For 'immediacy', there were a number of candidates who indicated (in accordance with the initial task complexity hypothesis) that removing the pictures was an obstacle, and that their inability to remember the details of the pictures detracted from their telling the story. In contrast, a number of candidates who found the + (no pictures) condition easier than the – (pictures present) condition explained that they could concentrate more on telling the story without the pictures in front of them. These polarized views were not related in any systematic way to candidate's actual story telling ability, as judged by the raters, which suggests that a range of individual factors quite independent of language ability such as visual acuity,

memory, personality or learning style may influence learner perceptions of task difficulty.

For 'adequacy', the hypothesized difficulty of the task was associated with the absence of key pictures depicting important components of the story. While the test-takers' task ratings support this hypothesis, there appears to be no relationship between these estimates and their level of proficiency. For example, both a high scorer (ranked 7) and a low scorer (ranked 132) reported that it was easier to tell the story when all pictures were present because they did not have to spend time thinking about what was missing. On the other hand, some found it easier to tell the story even if two pictures were missing as this allowed them to fill the gap in whatever way they liked. Here it seems that there is an interaction between two different factors: the creativity factor in the latter case and the immediacy factor in the former. Whether these factors are 'resource directing' or 'resource depleting' (Robinson, 2001; in press) may be an entirely individual manner which in turn makes their likely impact on actual performance extremely difficult to predict. On the other hand, it may be that the general lack of systematicity in candidates' responses to the various tasks may be a function of our failure adequately to operationalize the various task conditions in the tasks chosen for this study. The fact that the effect of these conditions was generally inconsistent from task to task lends some support to this possibility.

VI Discussion and conclusions

The present study operationalized task difficulty by attempting to apply to an L2 assessment context the insights from research on the cognitive demands of oral communicative tasks previously carried out in pedagogic settings. In addition, test-taker feedback was sought to examine if it was legitimate to consider test-taker perceptions as a component in any model of task difficulty and whether such perceptions could be useful in the design of test tasks. Students 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 by subjectively rating the performance using trained raters, and also by exploring student attitudes to the task and their perceptions of the difficulty of taking the tasks under various performance conditions.

The results showed no systematic variation associated with the various performance conditions for each task dimension, except in the case of 'immediacy' where the differences were in the opposite direction to what had been predicted. As in earlier studies, differences

in ability were found to be associated primarily with steps on the rating scale, rather than with differences in task demand. Student perceptions of difficulty, too, appeared not to be related to the predicted difficulty of the performance conditions for each task dimension.

The fact that our results differ so consistently and markedly from those of previous SLA research (e.g., Skehan and Foster, 1997; 1999; Robinson, 2001) may have to do with differences between testing and pedagogic contexts, with the former producing a cognitive focus on display rather than on task fulfilment or getting the message across. Under testing conditions, which in this case involved speaking in a language laboratory, candidates may concentrate on producing accurate speech, regardless of the conditions under which tasks are performed and may therefore be unable or unwilling to exploit the possibilities offered by varying the task conditions. This raises the issue of the validity of the testing of speaking in a semi-direct format and indeed of oral proficiency testing more generally.

It is also possible that the conditions of the experiment itself were not conducive to producing marked differences in the quality of candidate performance. Raters, for example, commented on the fact that the narrative task did not generate complex sentence structures, which raises the question of whether another kind of task might have produced a different finding. In addition, the fact that candidates were required to tell 8 different stories one after the other, and to report their perceptions of each task on each occasion, may have resulted in a certain perfunctoriness in their responses. Although one could argue that any such fatigue effect, or lack of motivation, would be offset by our counterbalancing the order of task presentation across the four different test-taker groups, it may nevertheless have resulted in an overall reduction in both the variability of candidate performance, on the one hand, and the intensity of candidates' affective reactions to the tasks, on the other.

Another, more pessimistic, interpretation of these findings is that it may simply not be possible to establish a hierarchy of task difficulty based on different task conditions or, indeed, to make reference to task conditions in any characterization of the ability levels on a scale of speaking proficiency. The absence, in this study's results, of consistent performance differences supporting our initial difficulty hypotheses gives no grounds for believing that greater effort in the production of more varied tasks would reverse the finding of this study.

In this regard it is worth noting that our attempt to isolate particular task complexity variables in isolation from others seems not to have been entirely successful. The fact that the different task exemplars within each dimension elicited different candidate reactions suggests

that there were unanticipated dimensions of complexity/difficulty embedded within each task. Such features as the topic and structure of each narrative or the clarity of the picture prompts appear to have influenced task performance in unpredictable and highly individual ways, despite our best efforts to produce equivalent task demands.[4] In some cases these difficulty factors were noticed by candidates (and were reflected in their ratings and comments about task difficulty), but in other cases they may have passed unnoticed but nevertheless have had an impact on the quality of their language production.

As far as the utility of test-taker feedback is concerned, it was reported above that test-takers perceptions of task difficulty did not generally correspond to the hypothesized difficulty of the different task conditions. The one exception was 'adequacy' where test-takers reacted more favourably and found the task easier when all pictures were provided. Putting a more positive spin on these findings, we could make a case to the effect that test-taker perceptions were in fact accurate for each of the remaining three dimensions in the sense that they were in line with actual differences (or lack of differences) in test performance as revealed by the FACETS analysis. For 'perspective' and 'planning', the condition under which the task was performed appeared to make no difference to perceptions of difficulty (corresponding to what was revealed in the actual performance data), whereas for 'immediacy', the majority perception that the presence of the pictures made one of the two tasks within this dimension less enjoyable was reflected in significant differences in the overall difficulty of this particular task condition as measured by scores assigned to test performance. Test-takers, in other words, may have some insight into whether a particular task feature or performance condition makes it easier to perform the task, and should perhaps – as Alderson (1988), Stansfield (1991) and Brown (1993) suggested – be consulted at the early stages of test development, along with other parties, to give their feedback on task selection and task design.

On the other hand, the erratic pattern of test-taker perceptions across the different task exemplars and the unsystematic relationship between perceptions of difficulty and proficiency (as measured by candidates' performance across the range of narrative tasks) suggests that we should not rely too heavily on test-taker feedback, either as a basis for test design or in mounting test validation arguments. Test-taker reactions and attitudes may be conditioned by a range of different attributes (e.g., gender, social class, professional experience,

[4]It is, moreover, worth noting that previous research on task complexity/difficulty tends to be based on single task exemplars rather than on multiple versions of a particular type of task.

proficiency) as our earlier review would indicate, as well as by features of the task itself.

In sum, this study has demonstrated the on-going difficulty of making *a priori* estimates of task difficulty in oral proficiency assessment using models such as those of Skehan, which have until recently been applied primarily in pedagogic contexts. In addition it has been shown that perception of task difficulty is a multidimensional phenomenon, resulting from a series of complex and unstable interactions between different task features and different test-taker attributes.[5] Until we know more about these interactions it seems unlikely that task difficulty can be accurately estimated after the event on the basis of subjective impressions of test-takers.

Acknowledgements

An earlier version of the article was presented at the Applied Linguistics Association of Australia 25th Annual Congress, Melbourne, Australia, 2001. The research paper reported in this article was funded by the Educational Testing Service (ETS) for the work of the TOEFL 2000 Speaking Team. We wish to acknowledge the generous input of the following people in the evaluation of this article: Peter Skehan, Peter Robinson and three anonymous ETS reviewers. We would also like to thank the editor of this issue and anonymous reviewers of *Language Testing* for their helpful suggestions for its improvement. Any errors that may remain are the sole responsibility of the authors.

VIII References

Alderson, J.C. 1988: New procedures for validating proficiency tests of ESP? Theory and practice. *Language Testing* 5, 220–32.

Bachman, L.F. 1990: *Fundamental considerations in language testing*. Oxford and New York: Oxford University Press.

Bachman, L.F. and **Palmer, A.** 1996: *Language testing in practice*. Oxford and New York: Oxford University Press.

Bradshaw, J. 1990: Test-takers' reactions to a placement test. *Language Testing* 7, 13–30.

Brindley, G. 1987: Factors affecting task difficulty. In Nunan, D., editor, *Guidelines for the development of curriculum resources*. Adelaide: Adelaide National Curriculum Resource Centre, 45–56.

Brooks, L. 1999: Adult ESL student attitudes towards performance-based assessment. Unpublished MA thesis, University of Toronto.

[5]The results of this study also suggest that generalizations made on the basis of learner performance on single task exemplars should be treated with extreme caution and that the findings of SLA research should also be revisited with this caveat in mind.

Brown, A. 1993: The role of test-taker feedback in the test development process: test-takers' reactions to a tape-mediated test of proficiency in spoken Japanese. *Language Testing* 10, 277–303.

Brown, J.D., Hudson, T. and **Norris, J.M.** 1999: Validation of test-dependent and task-independent ratings of performance assessment. Paper presented at the 21st Language Testing Research Colloquium, Tsukuba, Japan, July

Clarke, J.L.D. 1985: Development of tape-mediated, ACTFL/ILR scale-based test of Chinese speaking proficiency. In Stansfield, C.W., editor, *Technology and language testing*. Princeton, NJ: Education Testing Service, 129–46.

Davies, A., Brown, A., Elder, C., Hill, K., Lumley, T. and **McNamara, T.** 1999: *Dictionary of language testing*. Cambridge: UCLES, Cambridge University Press.

Douglas, D. 2000: *Language testing for specific purpose*. Cambridge: Cambridge University Press.

Elder, C. and **Lynch, B.** 1996: Public perceptions of basic skills tests and their ethical implications. Paper presented at the 20th Language Testing Research Colloquium, Monterey, March.

Foster, P. and **Skehan, P.** 1996: The influence of planning and task type on second language performance. *Studies in Second Language Acquisition* 18, 299–323.

Hill, K. 1998: The effect of test-taker characteristics on reactions to and performance on an oral English proficiency test. In Kunnan, A.J., editor, *Validation in language assessment*. Mahwah, NJ: Lawrence Erlbaum, 209–29.

Iwashita, N. and **Elder, C.** 1997: Expert feedback? Assessing the role of test-taker reactions to a proficiency test for teachers of Japanese. *Melbourne Papers in Language Testing* 6, 53–67.

Iwashita, N., McNamara, T. and **Elder, C.** 2001: Can we predict task difficulty in an oral proficiency test? Exploring the potential of an information processing approach to task design. *Language Learning* 21, 401–36.

Kenyon, D. and **Stansfield, C.** 1991: A method for improving tasks on the performance assessments through field testing. Paper presented in the annual meeting of the National Council on Measurement in Education, Chicago, IL, April 1991.

Kessler, S. 1984: *AMEP wastage survey*. Sydney, Australia: AMES.

Linacre, J.M. 1992: *FACETS Computer program for many faceted Rasch Measurement*. Chicago, IL: Mesa Press.

McNamara, T.F. 1990: Assessing the second language proficiency of health professionals. Unpublished PhD Dissertation, The University of Melbourne.

—— 1996: *Measuring second language performance*. London and New York: Addison Wesley Longman.

McNamara, T., Elder, C. and **Iwashita, N.** in preparation: Investigating

predictors of task difficulty in the measurement of speaking proficiency. Final Report, TOEFL 2000 Research Project. Princeton; NJ: Educational Testing Center.

Messick, S. 1989: Validity. In Linn, R.J., editor, *Educational measurement*. 3rd edition. New York: American Council on Education/Macmillan.

Nevo, B. 1985: Face validity revisited. *Journal of Educational Measurement* 22, 287–93.

Norris, J.M., Brown, J.D., Hudson, T.D. and **Bonk, W.** 2000: Assessing performance on complex L2 tasks: investigating raters, examinees and tasks. Paper presented at the 22nd Language Testing Research Colloquium, Vancouver, March.

Norris, J.M., Brown, J.D., Hudson, T. and **Yoshioka, J.** 1998: *Designing second language performance assessments. Technical Report 18, Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.* Honolulu: University of Hawaii Press.

Nunan, D. 1989: *Designing tasks for the communicative classroom*. Cambridge: Cambridge University Press.

Robinson, P. 1995: Task complexity and second language narrative discourse. *Language Learning* 45, 141–75.

—— 1996: Connecting tasks, cognition and syllabus design. In Robinson, P., editor, *Task complexity and second language syllabus design: data-based studies and speculations*. University of Queensland Working Papers in Applied Linguistics (Special Issue). Brisbane: University of Queensland, 1–16.

—— 2001: Task complexity, task difficulty and task production: Exploring interactions in a componential framework. *Applied Linguistics* 21, 27–57.

—— in press: Attention and memory during SLA. In Doughty, C. and Long, M., editors, *Handbook of research in second language acquisition*. Oxford: Blackwell.

Scott, M.L. 1986: Student affective reactions to oral language tests. *Language Testing*, 3, 99–118.

Scott, M.L. and **Madsen, H.S.** 1983: The influence of retesting on test affect. In Oller, J.W., editor, *Issues in language testing research*. Rowley, MA: Newbury House, 270–79.

Shohamy, E. 1982. Affective considerations in language testing. *The Modern Language Journal* 66, 13–17.

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

—— 1998: *A cognitive approach to language learning.* Oxford: Oxford University Press.

Skehan, P. and **Foster, P.** 1997: Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research* 1, 185–211.

—— 1999: The influence of task structure and processing conditions on narrative retellings. *Language Learning* 49, 93–120.

Slatyer, H., Brindley, G. and **Wigglesworth.** 2000: Task difficulty in ESL

listening assessment. Paper presented at the 22nd Language Testing Research Colloquium, Vancouver, March.

Spolsky, B. 1995: *Measured words: the development of objective language testing*. Oxford: Oxford University Press.

Stansfield, C. 1991: A comparative analysis of simulated and direct oral proficiency interviews. In Anivan, S., editor, *Current developments in language testing*. Singapore: SEAMEO RELC, 199–209.

Stansfield, C.W., Kenyon, D.M., Paiva, R., Doyle, F., Ulsh, I. and **Cowles, M.A.** 1990: The development and validation of the Portuguese speaking test. *Hispania* 73, 641–51.

Widdowson, H. 2001: Communicative language testing. In Elder, C., Brown, A., Grove, E., Hill, K., Iwashita, N., Lumley, T., McNamara, T. and O'Loughlin, K., editors, *Experimenting with uncertainty: essays in honour of Alan Davies*. Cambridge: UCLES, 12–21.

Wigglesworth, G. 1997: An investigation of planning time and proficiency level on oral test discourse. *Language Testing* 14, 85–106.

Zeidner, M. 1988: Sociocultural differences in examinees' attitudes toward scholastic ability exams. *Journal of Educational Research* 80, 352–258.

—— 1990: College students' reactions towards key facets of classroom testing. *Assessment and Evaluation in Higher Education* 15, 151–69.

Zeidner, M. and **Bensoussan, M.** 1988: College students' attitudes towards written versus oral of EFL. *Language Testing* 5, 100–14.

Appendix 1 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 attempt 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.