The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking

Tom Lumley and Barry O'Sullivan

Language Testing 2005; 22; 415
DOI: 10.1191/0265532205lt303oa

The effect of test-taker gender, audience and topic on task performance in tape-mediated assessment of speaking

Tom Lumley *Australian Council for Educational Research* and **Barry O'Sullivan** *University of Roehampton*

Performance in tests of spoken language can be influenced by a wide range of features of both task and participants. This article hypothesizes that there may be effects on performance attributable to an interaction of variables such as the task topic, the gender of the person presenting the topic and the gender of the candidate. In contrast to previous studies, which have examined speaking tests involving face-to-face interaction, this study considered the issue in a tape-mediated test delivered in a language laboratory, with no interlocutor present, but where stimulus material is presented by one or more speakers, one of whom acts as 'audience' for the candidate's speech. The test was taken by 894 students graduating from Hong Kong universities. In an advice-giving task, the last of a series involving different situations and audiences, topics considered stereotypically 'male'-oriented or 'female'-oriented were presented with systematic changes in the roles taken by the male and female speakers. A multi-faceted Rasch analysis examined the interaction of test-taker gender, task topic, and gender of presenter/audience. The results showed small effects for some, but not all, of the hypothesized interactions. Examples of differential performance by male and female candidates on other tasks are also presented. The article concludes with discussion of the implications for task design and test content.

I Introduction

Performance in language test tasks can be influenced by a wide range of features, which can interact unpredictably with characteristics of individual test-takers (O'Sullivan, 2000a). Collectively, these influences can be considered as contributing to task difficulty, a topic that has attracted a lot of interest recently (Iwashita *et al*., 2001;

Address for correspondence: Tom Lumley, Australian Council for Educational Research, 19 Prospect Hill Road, Private Bag 55, Camberwell, Victoria, Australia 3124; email: lumley@acer.edu.au

Language Testing 2005 22 (4) 415–437 10.1191/0265532205lt303oa © 2005 Edward Arnold (Publishers) Ltd

Bachman, 2002; Brindley and Slatyer, 2002; Elder *et al*., 2002; Norris *et al*., 2002; Fulcher and Márquez Reiter, 2003; Tavakoli and Skehan, 2003). There remains an assumption in many language testing contexts that test tasks are interchangeable for all sections of the test population. This article explores the question of whether tasks used may result in bias towards or against particular groups of test-takers, in the context of the speaking component of the English version of the Graduating Students' Language Proficiency Assessment (GSLPA) (Lumley and Stoneman, 2000; Lumley and Qian, 2003).[1] In contrast to previous studies, which have examined speaking tests involving face-to-face interaction, this study considers the issue in a tape-based test delivered in a language laboratory, where no interlocutor is present.

II Background to the test

The GSLPA is an exit-level examination for students graduating from universities in Hong Kong, 'designed primarily to provide potential employers a statement about students' English language proficiency around the time of graduation' (Lumley and Stoneman, 2000: 54). A 1995 survey less conducted by the Student Affairs Office of employment choices of graduates of the Hong Kong Polytechnic University, where the GSLPA was developed, showed that the great majority (over 80%) obtained employment in industry, commerce or professional firms. Employment in certain professions, notably education, requires separate assessment (Falvey and Coniam, 2000), making the GSLPA redundant in these contexts. The test, which is intended to make predictions about candidates rather than to look back on their academic career (Lumley and Qian, 2003), concentrates generally on the kinds of communication required for business, with content situated generally in the professional workplace domain in Hong Kong, and as such is considered suitable for a graduating cohort irrespective of the academic background of that group.

The test has a written and a spoken component. In the written component, test-takers are required to write two texts and complete a proof-reading task. The test of spoken language is administered in a language laboratory, requiring students to speak their responses

[1] A Chinese GSLPA was also developed. This article discusses only the GSLPA-English.

onto tape in response to input that they hear on an audiotape and read in their test booklets. The test therefore assesses only the spoken production of the test-taker rather than the interaction between test-taker and interlocutor found in interviews, role plays and other tests of speaking involving multiple speakers.

Most of the text types represented in the test are consistent with its general business-oriented focus:

- writing: memos, business letters, sections of reports;
- speaking: presentations to business colleagues, workplace-related phone messages, work-related interviews.

However, as Davies (2001: 138) among others points out, 'sampling for a proficiency test should not be restricted to the work domain', and he argues convincingly for the inclusion of test tasks relevant to a broader, social domain. In fact, the needs analysis for the GSLPA uncovered precisely the same view. Employers, when consulted about the anticipated language needs of their newly graduated employees, explicitly raised the need for social English to be represented in the spoken component of the test (Hong Kong Polytechnic University, 1996). The speaking test, therefore, includes two tasks relevant to this requirement. In the first of these, test-takers listen to a radio interview on a topic related to popular culture, entertainment or leisure activities (rather than to business contexts), which they then have to summarize for the benefit of a friend. The final task requires them to join in a conversation between a colleague and a visitor to Hong Kong, and to offer advice, opinions or recommendations about some aspect of life in Hong Kong. In this sense, then, the GSLPA does not fall foul of Davies' complaint that needs analyses are over-restrictive: the test samples a fairly broad domain of language use situations, more and less specifically related to the business environment.

The Hong Kong focus of the test content is a second indication of its specific nature, and its relevance to its target test population of graduates from universities in Hong Kong. In this it stands in marked contrast to a range of commercial tests vying for widespread acceptance and recognition in the territory as part of the government-sponsored 'Workplace English Campaign' (Education and Manpower Bureau, 2000). This campaign was launched to promote large-scale test-taking by employees in a range of different occupational categories as a means of assessing whether they had achieved expected standards (or benchmarks) of spoken and written English. The tests recognized under this campaign include the Test of English for

International Communication (TOEIC) (developed in the USA and mainly used in Japan and Korea), the English Language Skills Assessment (ELSA), the Business Language Testing Service (BULATS) and a suite of the City and Guilds Pitman tests (these last three developed in the UK). All of these tests claim to be international in focus, and hence make no explicit assumptions about knowledge of any particular country. By contrast, the GSLPA capitalizes on the test-takers' background to provide some parts of the content of some of the tasks, particularly the final task, as we have seen.

If a specific purpose test is to fulfil both its stated purpose and the exhortations of writers such as Widdowson (1983), Davies (2001) and others, as well as the demands of employers, it seems necessary to include language tasks that require the use of social English in unrestricted domains. This the GSLPA does, especially in the two tasks described above: summarizing a radio interview for the benefit of an English-speaking friend, and offering advice to a visitor to Hong Kong in a social context. The context domain then becomes very broad, and here is clearly not defined in terms of subject-specific content or knowledge. The test is specific in purpose, setting and target population, but much less so in terms of content domain, although it does require a background knowledge of Hong Kong, as well as some ability to relate to popular culture. However, given this broad assumption of common background knowledge (popular culture, for example, or aspects of Hong Kong), it is also necessary for the test developers to investigate the possibility of bias towards or against particular groups. This is consistent with the kind of investigation that considers the effect of test-taker background knowledge on a subject-specific test such as IELTS (see, for example, Clapham, 1996) and TOEFL. Of the many studies undertaken on TOEFL, a number looked at the impact of candidate-related variables on performance: for example, native language (Angoff and Sharon, 1974; Hosley, 1978; Swinton and Powers, 1980; Alderman and Holland, 1981), sex of test-taker (Blanchard and Reedy, 1970; Odunze, 1980; Wilson, 1982) and student major and text content in the reading section of the test (Hale, 1988).

III Variation in task performance

One type of potential bias that it is relevant to investigate is that relating to the sex of the test-taker. There is quite a large second language acquisition literature on the subject of how language tends to vary systematically under particular conditions (see, for example, Dickerson, 1975; Schmidt, 1980; Tarone, 1985; 1988; Ellis, 1989; Smith, 1989; Tannen, 1990; Coates, 1993), with the focus tending to be on factors such as topic and contextual change, and on speaker-related variables (age, sex and social class, for example). In other words, these studies have adapted a speaker-centred approach to performance. In addition, there are a small number of studies that focus on the effect of the audience, or perception of the audience, on linguistic performance (see, for example, Douglas-Cowie, 1978; Russell, 1982; Thelander, 1982; Bell, 1984).

The studies referred to here would suggest that Sunderland's (1995: 25) assertion that there is evidence that 'a test or exam can favour female or male testees in three possible ways: "Topic", "Task" and "Tester"' is both supportable and in need of some investigation.

From the literature, we can hypothesize that it is possible for particular topics to prove more accessible or familiar to males or females, although to date the only evidence in support of this claim remains anecdotal. There is also a possibility that women may feel more comfortable talking to a female audience; this is a suggestion supported by the results of Buckingham (1997), who found evidence of a same-sex effect in her interviews involving Japanese learners of English. This is, however, contradicted by Porter (1991a; 1991b) and O'Sullivan (2000b), who found that participants tended to achieve higher scores in interview-type tests when the interviewer was a man (with Arab students) or a woman (with Japanese students), irrespective of the sex of the interviewee. On the other hand, there may be some more complicated interaction amongst these and other facets of the interaction, as has been suggested by O'Sullivan (2000a).

Both Buckingham (1997) and O'Sullivan (2000a) examined Japanese candidates, but in different contexts: the former group were relatively low level students studying English at a UK university, while the latter group were upper intermediate students in Japan. On the other hand, Porter (1991a; 1991b) was primarily working with upper intermediate Arabic students studying in the UK.

Young and Milanovic (1992), in a qualitative study, suggested that the gender of both the interviewer and the candidate may be among the factors that account for discourse variation within the interview. However, they concluded that 'we still do not know whether or not any of this makes a difference to the examiner's assessment of the learner's oral proficiency' (Young and Milanovic, 1992: 421).

Unlike the studies referred to above, O'Loughlin (2002) found no evidence of a gender effect either on the scores achieved by male and female candidates or on features of the discourse hypothesized to be affected by the changes in condition he explored (where eight male and eight female candidates were each interviewed twice, once by a man and once by a woman). O'Loughlin (2002: 171) suggests that the inconsistency of the results in the studies reported by himself, Buckingham (1997), O'Sullivan (2000b) and Porter (1991a; 1991b), and the "unstable nature of gender in interaction" observed, may result from an interaction of gender of test-takers with other dimensions of the testing context, including the tasks used as well as the cultural background of participants.

Following a focus on the effect of task topic in learner production in the second language acquisition literature has come a growing interest in the language testing literature on how tasks used in tests affect learner performance on that test. This has led to unexpected findings, by researchers such as Hamp-Lyons and Mathias (1994) and McNamara and Lumley (1997), that the harder a task or its performance conditions appear to be, the easier it can turn out to be in measurement terms, perhaps as the result of compensation by raters. More recently, Skehan (1998), among numerous others (Porter, 1991a; 1991b; Porter and Shen Shu Hung, 1991; Foster and Skehan, 1994; 1996; 1999; Porter and O'Sullivan, 1994; 1999; Skehan and Foster, 1995; 1997; Wigglesworth, 1997; Mehnert, 1998; Bygate, 1999; Ortega, 1999), has attempted to manipulate psycholinguistic aspects of tasks in order to modify or predict difficulty. However, Iwashita *et al*. (2001) appear to cast doubt on the possibility of categorizing tasks in terms of difficulty following this approach (although their claims have been questioned by Tavakoli and Skehan, 2003), and we are still a long way from making accurate predictions of task difficulty.

Clearly, the issue is either not settled, or impossible to settle. What is suggested by the literature is that context and the profile of test-takers, or an interaction between them, may play a significant role in the findings.

One feature of the studies referred to above is that they have examined face-to-face interactions, where interlocutors may display differences in features such as personal style and degree of accommodation or support provided to test-takers (Ross, 1992; Ross and Berwick, 1992; Lumley and Brown, 1996; Reed and Halleck, 1997; Brown, 2003; 2004), or the degree of rapport established (Lazaraton, 1996; McNamara and Lumley, 1997; Morton *et al.*, 1997), all of which have the potential to influence test performance. An area that remains unexplored, in the literature is the degree to which factors that have been shown to impact on live (i.e. face-to-face) performance can affect performance on a tape-mediated test such as the GSLPA.

IV Research questions

Experience in Hong Kong learning contexts suggested, for example, that test conditions would be affectively different (and potentially less face-threatening) for test-takers whose interlocutor was a native speaker of Cantonese and therefore sharing a similar cultural background, compared to those whose interlocutor was an expatriate, typically a native speaker of English, who spoke little or no Cantonese. In contrast to tests of face-to-face interaction, no interlocutor is actually present in the GSLPA; instead, recorded stimulus material is presented by one or more speakers. The audience for the candidate's speech is either one of these speakers or an alternative fictional person or group described in the test rubric, but all test-takers have the same audience for any particular task in each test administration session.

In such a context, we might predict that any effect caused by audience is likely to be less significant, since the candidate is required to talk to a disembodied, fictional person. On the other hand, test developers take considerable care to create situations that are as realistic or 'authentic' as possible. In tests of this kind it is therefore important to consider the impact of gender of participants in 'interpersonal' tasks (such as giving advice, making suggestions, disagreeing) where candidates have to construct their discourse to suit the 'audience'. While the studies relating to changes in language production referred to above do not specifically focus on Chinese, we know from sources such as Hu (1991) and Chan (1998) that not only are there measurable differences in the way in which Chinese men and women speak, but there is also evidence of changes in speech (particularly of children and young women) depending on the addressee (Farris, 1995).

In this tape-mediated situation, how do topics that might potentially advantage either males or females in fact operate? One of the difficulties of writing materials for a test such as the GSLPA is finding credible situations that do not obviously cause advantage or disadvantage to any clearly identified group. The test is designed to be relevant for graduates of any academic discipline, as noted, but the issue of its fairness for males and females requires investigation.

Materials writers ‘dress tasks up’ in order to provide contexts, so that a task might superficially appear to be concerned with soccer, which in stereotypical terms is clearly a male-oriented topic, since males tend to show more interest in soccer than females. However, the task candidates are actually faced with might not require any particular knowledge or familiarity with soccer. The question here is whether or not female candidates are to some extent put off by the apparent presentation of a topic of which they may be quite ignorant, leading them to perform less well than they should.

In the advice-giving task of the GSLPA, Task 5, topics considered stereotypically ‘male’ or ‘female’ oriented, in the judgement of the test development team, are presented with systematic changes in the roles taken by the male and female speakers. Sometimes the test-takers were required to speak to a male audience and sometimes to a female audience, but in other respects the tasks were identical.

Because this is potentially a high-stakes test (since it has the possibility to influence employment decisions), we need as an issue of fairness and validity to consider various ways in which test-takers might be disadvantaged or advantaged. Therefore, this article aims to investigate two questions:

- Does the gender of the intended audience for the candidate response in a tape-mediated test make a difference to the test scores?
- Do stereotypically male-oriented or female-oriented tasks have an observable effect on scores?

V The study

1 The test population

A group of 894 students from two Hong Kong universities sat the GSLPA during April 1999. Graduating students from a pair of universities in Hong Kong might superficially appear to constitute a homogeneous group, especially by comparison with the populations of international tests such as TOEFL or IELTS, but, as with most tests, the reality is more complex. In fact the group varies in terms of the following factors:

- academic background, experience and discipline;
- course types (full-time degree-level students mostly, but also significant numbers of full-time and part-time Higher Diploma students);

- age;
- level of professional experience: both pre-experience (students straight from school) and post-experience (those with varying time in the workplace).

2 *The raters*

A group of 30 raters (13 native speakers of English and 17 speakers of other languages, predominantly Chinese) rated the taped performances. As the test was at the final stages of development at the time of this study, the raters used in the study were drawn from the existing group of trained and accredited examiners.

Before they were selected for training, it was established that all GSLPA raters met specified qualifying criteria (which included professional ESL teaching qualifications and experience as well as experience as a rater). Rater training consisted of a one-day seminar in which the test purpose and rating scale were debated and then a series of performances, deemed by an 'expert' panel to be at a range of levels, were discussed and rated.

Following this procedure, raters were asked to rate a set of 10 test performances. Their performance was monitored in terms of agreement and consistency. Raters found to have met the set accreditation standards participated in the scoring procedure.

3 *The test tasks*

While a 'content-free' test might be considered ideal in order not to disadvantage particular test-takers, this is not possible to achieve. Therefore test designers are forced to seek topics of general interest, and to guess at commonly assumed knowledge and interests. In the GSLPA, the unifying feature is the test-takers' experience of Hong Kong, as the vast majority were born and raised there, and all had lived in the territory for at least two years, offering a context for test content. However, amongst topics that satisfy these demands, there are those that are likely to be more or less accessible, interesting and familiar to males or females. Such perceptions were repeatedly voiced by members of the test development team as well as teachers and others with an interest in the test.

The needs analysis conducted with employers during test development showed that in their use of spoken English, employees needed not only to cope with formal situations, but also to mix with international clients and colleagues, from the same company or from other businesses in more informal situations; for example, in marketing, in project management, and in hospitality and tourism. The test therefore includes a range of tasks, varying in formality, setting and number of participants. All tasks require comprehension and use of written and/or spoken input of various kinds, so that each tests more than the single macro skill of speaking (see Table 1). A second issue is that of audience for the test-taker's response. Audience is always stated in task prompts, in the following ways:

- If the audience is a single person, the test-taker is given the name, role, type of relationship (e.g., friend, colleague in same/other workplace, client, potential employer), and one piece of contextual information to assist in developing a purpose for the task (such as interlocutor's interests, previous telephone call to which you are replying).
- If the speaking performance is to be addressed to a group, the test-taker is given information about the purpose of talk, level of familiarity (e.g., colleagues, students from former university), and of the formality of the interaction.

Since we were interested in the effect on the performance of male and female candidates of systematic variations in test topic and audience, tasks were manipulated across these two variables. An example of this is Task 5, where candidates are asked to listen to a short interaction between a work colleague and an international visitor and then to give advice to the visitor. The task was varied, as shown in Table 2, with topics expected to suit either male candidates or female candidates, or to be neutral. Two versions of each task topic were recorded. The scripted task prompts included wording that would be suitable for both male and female speakers. This was important as the wording of the input should not change from version to version.

Table 1 Outline of the five GSLPA tasks

	Task	Input	Audience
1	Summary	listening (5-minute dialogue)	friend
2	Interview	reading (advertisement)	interviewer (e.g., employer)
3	Presentation	reading (graphic)	group (peers, colleagues, students)
4	Phone message	listening/reading (minimal)	colleague, client, etc.
5	Giving advice	listening/reading (minimal)	international visitor (colleague, client)

Table 2 Six variations on Task 5

Task ID	Audience	Topic orientation
Accommodation/housing		
5.3	female	neutral
5.6	male	neutral
Entertainment/horse racing		
5.2	female	male
5.4	male	male
Leisure/places to visit		
5.1	female	neutral
5.5	male	neutral

In one version of the topic, the female speaker asked the question of the candidate (indicating that the audience was female), while in the other the male speaker asked the question. Similar procedures were repeated for the other tasks.

Performance on each task was rated using an analytic rating scale, with six levels and including the following scoring criteria (though not all criteria were used for every task):

- task fulfilment and relevance;
- clarity of presentation;
- grammar and vocabulary;
- pronunciation;
- clarity of presentation.

VI Data analysis

A total of 894 students each took one of 14 versions of the test.[2] Each individual task appeared in several forms of the test, allowing analysis that compares tasks across forms. The method of analysis used was multi-faceted Rasch analysis, which provides probabilistic estimates of (in this study):

- ability of test-takers;
- difficulty of each task;
- challenge/difficulty of each of the scoring criteria;
- harshness/leniency of each rater.

[2] Each of these versions was compiled from a bank of test tasks. The confidential nature of the test means that actual test tasks cannot be included in this article.

These estimates contribute to the calculation of test-taker scores, thus assisting in reducing the problem of differences amongst test forms and raters. This analysis also offers a kind of analysis that detects bias against or in favour of any specified group. It does this by comparing the scores obtained by that group with the scores the model would have predicted them to receive, on the basis of what is known about the other features mentioned (test-taker ability, task difficulty, rater harshness).

VII Results

The summary diagram (Figure 1) is a visual representation of the analysis of all of the facets included in the analysis. To read the summary diagram it is best to view each vertical 'ruler' as a separate result, while the first column (Logit Measr) is a common scale upon which all results have been placed. Column 2 shows a graph of the distribution of the candidates' scores, where it can be seen that the distribution is normal. Column 3 shows the distribution of the raters, in terms of harshness/leniency. Myford and Wolfe (2000: 11) suggest that the ratio of the spread of these two columns to one another represents an estimation of the impact on candidate performance of rater harshness, suggesting that where the candidate range is greater than the rater range by a factor of at least two, then the impact is not significant. From the chart we can see that, even allowing for outliers, this ratio is substantially exceeded.

In Column 4 we can see that the female students tended to slightly outperform male students, although the actual difference was not significant, while from Column 5 it is clear that there is a range of task difficulty. The final column represents the actual six-point scale used in the study.

Table 3 presents the results of a bias analysis examining whether particular tasks favoured candidates of one gender. The two facets analysed were gender of candidate and task difficulty as measured by the individual assessment criteria used for each task. In this type of bias analysis, significant bias is indicated by a Z-score exceeding plus or minus 2, as shown in Column 4 of the table; a positive Z-score indicates bias against members of the sex stated in Column 6, hence an advantage to the opposite sex; a negative Z-score bias indicates bias in favour of members of the sex in Column 6. The table shows that there are only four tasks, out of 27 in all, where

```
---------------------------------------------------------
|Logit |High        |Harsh     |Sex  |Difficult  |Scale |
|Measr |Achieving   |Raters    |High |Tasks      |      |
|      |Students    |          |Score|           |      |
---------------------------------------------------------
+  10  +            +          +     +           + (6)  +
|      |            |          |     |           |      |
|      |            |          |     |           |      |
+   9  + .          +          +     +           +      +
|      |            |          |     |           |      |
|      |            |          |     |           |      |
+   8  +            +          +     +           +      +
|      |            |          |     |           |      |
|      |            |          |     |           |      |
+   7  +            +          +     +           +      +
|      | .          |          |     |           |      |
|      | .          |          |     |           |  --- |
+   6  + .          +          +     +           +      +
|      | .          |          |     |           |      |
|      | .          |          |     |           |      |
+   5  + *.         +          +     +           +      +
|      | *.         |          |     |           |  5   |
|      | *.         |          |     |           |      |
+   4  + **.        +          +     +           +      +
|      | *.         |          |     |           |      |
|      | ***        |          |     |           |  --- |
+   3  + ****.      +          +     +           +      +
|      | ******     |          |     |           |      |
|      | ******     |          |     |           |      |
+   2  + ********   +          +     +           +      +
|      | ********   | *        |     | .         |  4   |
|      | *********. | **       |     | *.        |      |
+   1  + *********. + *        +     +           +      +
|      | *****.     | *****    |     | ***.      |      |
|      | *******.   | ****     | F   | ******    |  --- |
*   0  * *******.   * ****     *     * *******.  *      *
|      | ******.    | ******** | M   | ********  |      |
|      | ****.      | **       |     | *****.    |  3   |
+  -1  + ****.      + *        +     +           +      +
|      | ***        |          |     | .         |      |
|      | ***.       |          |     |           |  --- |
+  -2  + *.         + *        +     +           +      +
|      | *.         | *        |     |           |      |
|      | .          |          |     |           |  2   |
+  -3  + .          +          +     +           +      +
|      | .          |          |     |           |  --- |
|      | .          |          |     |           |      |
+  -4  +            +          +     +           +  1   +
|      | .          |          |     |           |      |
|      | .          |          |     |           |      |
+  -5  + .          +          +     +           +  --- +
|      |            |          |     |           |      |
|      |            |          |     |           |      |
+  -6  + .          +          +     +           + (0)  +
---------------------------------------------------------
|Logit |Low         |Lenient   |Sex  |Easy       |Scale |
|Measr |Achieving   |Raters    |Low  |Tasks      |      |
|      |Students    |          |Score|           |      |
|      | * = 8      | * = 1    |     | * = 2     |      |
---------------------------------------------------------
```

Figure 1 Results summary (all facet vertical 'rulers')

Table 3 Summary of bias analysis results

Raw average obs-exp	Logit bias+ measure	Bias model S.E.	Bias Z−score	Bias infit MnSq	Sex of candidate	Task ID	Advantage
−.09	.24	.12	2.04	0.8	male	3.1	F
.10	−.28	.12	−2.32	1.4	female	4.3	F
.10	−.28	.13	−2.13	0.9	male	5.1	M
−.15	.38	.19	2.03	1.6	female	5.4	M

the bias for or against candidates of either sex was identified as significant. Of these, two were found in two versions of the final task, number 5 (Tasks 5.4, 'Entertainment/Horse racing', male audience: advantage to males; and 5.5,[3] 'Leisure/Places to visit, male audience: advantage to males), where both topic and audience (a single person in each case) might potentially play a part. The influence of topic is considered further below.

Of the remaining two instances of bias, however, one occurred with one of the presentation tasks (Task 3.1, 'Office/Computers', group audience: advantage to females), where the audience is a group of colleagues rather than a single person. The other instance occurred with one of the phone message tasks (Task 4.3, 'Airport meeting', female audience: advantage to females), where no effect had been anticipated by the test development team, because of the brevity and apparently routine nature of the task. In this task test-takers had one minute to plan, and 40 seconds to speak; the questions to be addressed were specified and all the necessary information was provided. We can see that each gender appeared to gain advantage on two occasions, and that where there was a male audience, males had an advantage, whereas where there was either a female or a group audience, there was an advantage to females. As noted, however, these effects were not repeated across the great majority of tasks. Column 1 in Table 3 shows that the differences between observed and expected scores, although significant, were small: between .09 and .15 of a score point on average.

However, further exploration of the data was conducted, to see if any trends modified this preliminary view. In addition to examining significant bias, therefore, an attempt was made to detect observable differences in scores obtained by candidates of each sex. An observable effect was defined as one where there was on average

[3] Tasks are identified by Task Type and Task Form, so Task 5.4 means Task 5, Form 4.

an overall discrepancy between the average scores obtained by males and females of at least 0.1 of a raw scale point on any scoring category. This level of difference was deemed to be the smallest that might make a noticeable difference to test-takers' final scores. Data for Task 5 were examined first. In this task (see Table 4), it was anticipated that there might be observable effects (even though not identified in the analysis as significant) on performance related to task topic, where the gender of the audience had been systematically varied. The anticipated bias is indicated under the 'Topic' column and the actual bias found as a result of the analysis is shown in the 'Advantage' column.

The first pair of versions of Task 5 (in each pair one had a female audience, the other a male audience, as stipulated in instructions to test-takers) contained a topic related to accommodation and housing, which was deemed by the test development team to be of equal relevance to both sexes, and was therefore categorized as neutral. Unsurprisingly, there was no evidence of advantage to candidates of either sex, regardless of the gender of the audience, for any of the scoring criteria.

There was clearer evidence in the second pair of tasks of a trend supporting our expectation, with one of them clearly biased towards males and against females. In these tasks the topic was entertainment and horse racing. In Hong Kong, horse racing and the gambling associated with it are commonly discussed in the media and in public life. However, many more males attend horse races than do females, and it is a more common topic of conversation amongst males than females, so that although the task as presented in the test does not

Table 4 Task 5: anticipated task orientations, audience and actual bias

Task ID	Topic	Audience	Actual advantage	Scoring category
Accommodation/housing: easiest of tasks				
5.3	neutral	female	none	–
5.6	neutral	male	none	–
Entertainment/horse racing: medium diffculty				
5.2	male	female	male	task fulfilment and relevance
5.4	male	male	male	clarity of presentation; task fulfilment and relevance; grammar and vocabulary (but NOT pronunciation)
Leisure/places to visit: hardest of tasks				
5.1	neutral	female	male	clarity of presentation
5.5	neutral	male	none	–

require any detailed knowledge of horse racing, this was categorized by the test development team as essentially male-oriented. With a female audience, there was a slight advantage to males on one of the scoring criteria, ‘task fulfilment and relevance’. However, when the audience was male, the difference became much more marked, with advantages for male test-takers on three of the four scoring criteria, most markedly for ‘task fulfilment and relevance’ again. There appears to be an interaction here, where the combination of a male-oriented topic and a male audience renders the task harder for females.

Suggestions were included in the task instructions for all versions of Task 5 for points students might wish to discuss, but it was made clear these were not obligatory. In rater training, it was made clear that raters were to take a liberal view of relevance. Nevertheless, this task seems to have caused more difficulties for the female students in terms of providing a relevant response. From what is observed here, it does seem to be the case that topic is more significant than audience, while an interaction of the two compounds the effect.

The final pair of tasks was concerned with leisure and common places to visit in Hong Kong. There was no obvious topic orientation towards males or females here. Although some suggestions were made in the task rubric for places test-takers might wish to talk about, these included typical places of interest to both local and international visitors, visited equally commonly by males and females. This topic was therefore anticipated by the test developers to be equally familiar to both sexes. One instance of significant bias in this task appeared, with males outperforming females on the scoring category of ‘clarity of presentation’. This finding is hard to explain, since what is observed is neither obviously an issue of task nor of the gender of the participants.

The second task type investigated was Task 1, where topic might be expected to play a role because of the substantial nature of the input: a five-minute dialogue that test-takers are required to summarize for a friend (Table 5).

As in Table 4, Table 5 shows the predicted orientation of the topic in the judgement of the test development team, the audience, and the actual direction of any bias found. Although soccer and fashion tend to be of more interest to males and females, respectively, no clear prediction could be made about the orientation of basketball. In most cases, there was no evidence of a difference, but in the task superficially to do with fashion, there was slight advantage to

Table 5 Task 1: anticipated and actual bias

Task 1	Topic	Audience	Actual advantage	Criteria
Basketball	neutral	not clear	female	clarity of presentation; grammar and vocabulary
Soccer	male	male	none	–
Fashion	female	female	female	clarity of presentation

females for just one of the four scoring categories, 'clarity of presentation'. In the task on basketball, the effect size for the two scoring categories was even smaller, only just meeting the level determined as 'observable'. Interestingly, the topic superficially about soccer had no adverse effect on female test-takers.

Most importantly, most of the bias terms are small. Where there is evidence of any measurable bias at all, it tends to be for one scoring category only. There were a few conflicting examples, where a task that was relatively easy in one test form for females was harder for them in another test form, where only minor details of fact had changed, but not the situation or audience. This suggests that the effect is insufficiently reliable to indicate systematic bias.

There was one further, confusing result. For one of the phone message tasks (Task 4), where no difference in performance was expected, the largest absolute difference between male and female students was observed. This amounted to an average of about .25 of a scale point, which is clearly measurable. This appeared consistent with the earlier observation, noted in Table 3, of a significant bias against males on a different phone message task. Although the differences tended to be very small, in all cases the females performed better on this task. In the absence of a better explanation, what we seem to have here is a task type, rather than instances of tasks, on which females appear to perform better. Such an observation is likely to indicate not unfairness, but real differential ability on this task type by males and females: females simply appear to do better at this task, and the test is merely revealing this truth. The question here is whether or not we consider it necessary to include such a task in the test. This is a validity issue, not one of fairness. The authenticity of the task of leaving a phone message provides a strong argument for retaining it. The task is arguably the only one where the tape-mediated test method replicates real life; in addition, it is a skill that just about everyone in Hong Kong could be expected to demonstrate high competence in, and one that is unquestionably necessary in today's professional environment.

VIII Implications

This study has used a statistical approach to examining the issue of the influence on task difficulty of topic and audience in a new context, that of a tape-based speaking test. It found only limited evidence that the gender of the hypothetical interlocutor in a tape-mediated test plays much of a role, although this is apparently not always the case, and it cannot reliably be predicted.

A slightly more significant part seems to be played by task topic, and sometimes an interaction of topic and the gender of the audience leads to bias. In the instance observed here, it appears likely that the female students had less expertise than the males about horse racing, or else were to some extent put off by the topic. We may speculate that when required to talk about a topic they were unfamiliar with to a (hypothetical) foreign male, this is more face-threatening than showing their ignorance to an absent female.

However, the bias is not always predictable: instances are observed here where there was unexpected bias, and others where it failed to be seen where it might be anticipated. In addition, interactions seen in this study are by and large small, and although not negligible, do not seem to pose a huge concern. The results are perhaps not unexpected, but appear reasonably reassuring in this context.

The somewhat inconclusive results suggest that further exploration, or at least monitoring, is warranted, although the expectation should be more that these findings will be confirmed than that major differences will be found. Test developers should be not complacent but cautiously optimistic. This study offers further evidence that at all stages of the test development process there is a need for careful judgement, for which there would appear to be no substitute. Despite the inconclusiveness of earlier studies – or more accurately, consistent with it – this study suggests the need for questions like this to be investigated again in each new context.

The study suggests that test-takers are to some extent affected by the contextual features of tasks, even in a tape-mediated test: the fictional, disembodied audiences do take on a certain reality in the perception of the test-takers. The test-takers do appear to be accepting the authenticity of the stimulus not only as a test task but as a communicative event, to some extent at least. Even in a tape-mediated test the notion of interaction is apparently not extinguished.

A final observation is that this study points in similar directions as previous studies: task difficulty is too complex to be categorized in such simplistic terms as stereotypical notions of what is interesting or familiar to men or women. Tasks are more likely to affect individuals differentially, rather than at the level of groups defined at this level. It should be noted that use of statistical analysis alone, as in this article, is limited in its potential to uncover differences in how task features may influence performance for individuals. As noted by Fulcher and Márquez Reiter (2003: 324):

> it is no longer adequate to attach a difficulty statistic to a speaking task; rather, difficulty is determined by a whole range of task features and conditions that must be manipulated independently in order to investigate their impact upon discourse variation and test score variance.

Approaches that combine statistical analysis with analysis of discourse, as in Brown's (2003; 2004) study of the influence on performance of variability in interlocutor style, are therefore more likely to shed light on the complexities of features that affect task difficulty.

Acknowledgement

Thanks are due to three Language Testing anonymous reviewers for their helpful comments on earlier drafts of this article.

IX References

Alderman, D.L. and **Holland, P.W.** 1981: *Item performance across native language groups on the test of English as a foreign language*. TOEFL Research Report No. 9. Princeton, NJ: Educational Testing Service.

Angoff, W.H. and **Sharon, A.T.** 1974: The evaluation of differences in test performances of two or more groups. *Educational and Psychological Measurement* 34, 807–16.

Bachman, L.F. 2002: Some reflections on task-based language performance assessment. *Language Testing* 19, 453–76.

Bell, A. 1984: Language style as audience design. *Language in Society* 13, 145–204.

Blanchard, J.D. and **Reedy, R.** 1970: *The relationship of a test of English as a second language to measures of achievement and self-concept in a sample of American-Indian students*. Research and Evaluation Report Series No. 58. Washington, DC: Bureau of Indian Affairs, US Department of Interior.

Brindley, G. and **Slatyer, H.** 2002: Exploring task difficulty in ESL listening assessment. *Language Testing* 19, 369–94.

Brown, A. 2003: Interviewer variation and the co-construction of speaking proficiency. *Language Testing* 20, 1–25.

—— 2004: Interviewer variability in oral proficiency interviews. Unpublished PhD dissertation, University of Melbourne, Melbourne.

Buckingham, A. 1997: Oral language testing: do the age, status and gender of the interlocutor make a difference? Unpublished MA dissertation, University of Reading.

Bygate, M. 1999: Quality of language and purpose of task: patterns of learners' language on two oral communication tasks. *Language Teaching Research* 3, 185–214.

Chan, M.K.M. 1998: Gender differences in the Chinese language: a preliminary report. In Hua Lin, editor, *Proceedings of the ninth North American Conference on Chinese Linguistics*. NACCL 9, May 1997. Volume 2. Los Angeles, CA: GSIL, University of Southern California, CA, 35–52.

Clapham, C. 1996: *The development of IELTS: a study of the effect of background knowledge on reading comprehension*. Cambridge: Cambridge University Press.

Coates, J. 1993: *Women, men and language*. Second edition. First edition 1986. Harlow: Longman.

Davies, A. 2001: The logic of testing languages for specific purposes. *Language Testing* 18, 133–48.

Dickerson, L. 1975: The learner's interlanguage as a system of variable rules. *TESOL Quarterly* 9, 401–07.

Douglas-Cowie, E. 1978: Linguistic code-switching in a Northern Irish village: social interaction and social ambition. In Trudgill, P., editor, *Sociolinguistic patterns in British English*. London: Edward Arnold, 37–51.

Education and Manpower Bureau 2000: Hong Kong Government information homepage. Workplace English Campaign. Hong Kong: Education and Manpower Bureau. Available at http://www.english.gov.hk (November 2004).

Elder, C., Iwashita, N. and **McNamara, T.** 2002: Estimating the difficulty of oral proficiency tasks: what does the test-taker have to offer? *Language Testing* 19, 347–68.

Ellis, R. 1989: Sources of intra-learner variability in language. In Gass, S., Madden, C., Preston, D. and Selinker, L., editors, *Variation in second language acquisition: psycholinguistic issues*. Volume 2. Clevedon: Multilingual Matters, 22–45.

Falvey, P. and **Coniam, D.** 2000: Establishing writing benchmarks for primary and secondary language teachers in Hong Kong. *Hong Kong Journal of Applied Linguistics* 5, 128–59.

Farris, C.S. 1995: A semiotic analysis of *sajiao* as a gender marked communication style in Chinese. In Johnson, M. and Chiu, F.Y.L., editors, *Unbound Taiwan: closeups from a distance*. Select Papers. Volume 8. Chicago: Center for East Asian Studies, University of Chicago, 1–29.

Foster, P. and **Skehan, P.** 1994: The influence of planning on performance in task-based learning. Paper presented at the annual meeting of the British Association of Applied Linguistics, Leeds.

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

—— 1999: The influence of source of planning and focus of planning on task-based performance. *Language Teaching Research* 3, 215–47.

Fulcher, G. and **Márquez Reiter, R.** 2003: Task difficulty in speaking tests. *Language Testing* 20, 321–44.

Hale, G.A. 1988: *The interaction of student major-field group and text content in TOEFL reading comprehension*. TOEFL Research Report No. 25. Princeton, NJ: Educational Testing Service.

Hamp-Lyons, L. and **Mathias, S.P.** 1994: Examining expert judgments of task difficulty on essay tests. *Journal of Second Language Writing* 3, 49–68.

Hong Kong Polytechnic University 1996: Graduating Students Language Proficiency Assessment (GSLPA). English Strand. A UGC funded project. Central allocation vote 1992–95. Project code HKP 17. Report submitted to the UGC. Hong Kong: English Department, Hong Kong Polytechnic University.

Hosley, D. 1978: Performance differences of foreign students on the TOEFL. *TESOL Quarterly* 12, 99–100.

Hu, M. 1991: Feminine accent in the Beijing vernacular: a sociolinguistic investigation. *Journal of the Chinese Language Teachers Association* 36, 49–54.

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 test design. *Language Learning* 51, 401–36.

Lazaraton, A. 1996: Interlocutor support in oral proficiency interviews: the case of CASE. *Language Testing* 13, 151–72.

Lumley, T. and **Brown, A.** 1996: Specific-purpose language performance tests: task and interaction. In Wigglesworth, J. and Elder, C., editors, *The testing cycle: from inception to washback. Australian Review of Applied Linguistics, Series S,* Volume 13. Canberra: Australian Review of Applied Linguistics, 105–36.

Lumley, T. and **Qian, D.** 2003: Assessing English for employment in Hong Kong. In Coombe, C.A. and Hubley, N., editors, *Assessment practices*. Alexandria, VA: TESOL, 135–47.

Lumley, T. and **Stoneman, B.** 2000: Conflicting perspectives on the role of test preparation in relation to learning. *Hong Kong Journal of Applied Linguistics* 5, 50–80.

McNamara, T.F. and **Lumley, T.** 1997: The effect of interlocutor and assessment mode variables in offshore assessments of speaking skills in occupational settings. *Language Testing* 14, 140–56.

Mehnert, U. 1998: The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition* 20, 83–108.

Morton, J., Wigglesworth, G. and **Williams, D.** 1997: Approaches to the evaluation of interviewer performance in oral interaction tests. In Brindley, G. and Wigglesworth, G., editors, *Access: issues in English language test design and delivery*. Sydney: National Centre for English Language Teaching and Research, Macquarie University, 175–96.

Myford, C.M. and **Wolfe, E.W.** 2000: *Strengthening the ties that bind: improving the linking network in sparsely connected rating designs*. TOEFL Technical Report No. 15. Princeton, NJ: Educational Testing Service.

Norris, J.M., Brown, J.D., Hudson, T.D. and **Bonk, W.** 2002: Examinee abilities and task difficulty in task-based second language performance assessment. *Language Testing* 19, 395–418.

O'Loughlin, K. 2002: The impact of gender in oral proficiency testing. *Language Testing* 19, 169–92.

O'Sullivan, B. 2000a: Towards a model of performance in oral language testing. Unpublished PhD dissertation, University of Reading.

—— 2000b: Exploring gender and oral proficiency interview performance. *System* 28, 1–14.

Odunze, O.J. 1980: Test of English as a Foreign Language and first year GPA of Nigerian students. Unpublished doctoral dissertation, University of Missouri-Columbia, MO.

Ortega, L. 1999. Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition* 20, 109–48.

Porter, D. 1991a: Affective factors in language testing. In Alderson, J.C. and North, B., editors, *Language testing in the 1990s*. London: Macmillan, Modern English Publications in association with The British Council, 32–40.

—— 1991b: Affective factors in the assessment of oral interaction: gender and status. In Anivan, S., editor, *Current developments in language testing*. Singapore: SEAMEO Regional Language Centre, 92–102.

Porter, D. and **O' Sullivan, B.** 1994: Writing for a reader: does the nature of the reader make a difference? Paper presented at the RELC regional seminar, Singapore, April.

—— 1999: The effect of audience age on measured written performance. *System* 27, 65–77.

Porter, D. and **Shen Shu Hung** 1991: Gender, status and style in the interview. *The Dolphin 21*. Aarhus University Press, 117–28.

Reed, D.J. and **Halleck, G.B.** 1997: Probing above the ceiling in oral interviews: what's up there? In Kohonen, V., Huhta, A., Kurki-Suonio, L. and Luoma, S., editors, *Current developments and alternatives in language assessment: proceedings of LTRC 96*. Jyväskylä: University of Jyväskylä and University of Tampere, 225–38.

Ross, S. 1992: Accommodative questions in oral proficiency interviews. *Language Testing* 9, 173–86.

Ross, S. and **Berwick, R.** 1992: The discourse of accommodation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 159–76.

Russell, J. 1982: Networks and sociolinguistic variation in an African urban setting. In Romaine, S., editor, *Sociolinguistic variation in speech communities*. London: Edward Arnold, 125–40.

Schmidt, M. 1980: Coordinate structures and language universals in interlanguage. *Language Learning* 30, 397–416.

Skehan, P. 1998: *A cognitive approach to language learning*. Oxford: Oxford University Press.

Skehan, P. and **Foster, P.** 1995: Task type and task processing conditions as influences on foreign language performance. *Thames Valley University Working Papers in English Language Teaching* 3, 139–88.

—— 1997: The influence of planning and post-task activities on accuracy and complexity in task-based learning. *Language Teaching Research* 1, 185–211.

Smith, J. 1989: Topic and variation in ITA oral proficiency: SPEAK and field-specific oral tests. *English for Specific Purposes* 8, 155–68.

Sunderland, J. 1995: Gender in language testing. *Language Testing Update* 17, 24–35.

Swinton, S.S. and **Powers, D.E.** 1980: *Factor analysis of the Test of English as a Foreign Language for several language groups*. TOEFL Research Report No. 6. Princeton, NJ: Educational Testing Service.

Tannen, D. 1990: *You just don't understand: women and men in conversation*. New York: William Morrow.

Tarone, E. 1985: Variability in interlanguage use: a study of style shifting in morphology and syntax. *Language Learning* 35, 373–404.

—— 1988: *Variation in interlanguage*. London: Edward Arnold.

Tavakoli, P. and **Skehan, P.** 2003: Planning, narrative task structure and performance. Paper presented at the Language Testing Research Colloquium, University of Reading, July.

Thelander, M. 1982: A qualitative approach to the quantitative data of speech variation. In Romaine, S., editor, *Sociolinguistic variation in speech communities*. London: Edward Arnold, 65–83.

Widdowson, H.G. 1983: *Learning purpose and language use*. London: Oxford University Press.

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

Wilson, K.M. 1982: *GMAT and GRE aptitude test performance in relation to primary language and scores on TOEFL*. TOEFL Research Report No. 12. Princeton, NJ: Educational Testing Service.

Young, R. and **Milanovic, M.** 1992: Discourse variation in oral proficiency interviews. *Studies in Second Language Acquisition* 14, 403–24.