An investigation into the comparability of direct and semi-direct versions of an oral interaction test in English

Gillian Wigglesworth & Kieran O'Loughlin

In this paper we investigate the comparability of two versions of an oral interaction test—a direct (live interview) and semi-direct (tape-based). This is one part of a four-skill English proficiency test to be administered to certain categories of intending migrants to Australia. The test is designed to be taken by these intending migrants in their country of origin; two versions are necessary for economy and flexibility, since the human and/or technical resources available in each overseas test centre may mean that only one version can be administered.

In investigating the issue of comparability we examine (1) the extent to which test items were of similar difficulty across both versions of the test; (2) whether the same candidates perform similarly on the direct and semi-direct versions of the test; (3) the extent to which the performance of the candidates in each version correlates with their ratings on another well-established scale of language proficiency, and (4) the effect test method on success or failure of the candidates on the test.

Previous research investigating the comparability of candidate scores on direct and semi-direct tests has been undertaken on a variety of tests. However, the numbers under investigation have generally been small (N sizes of 10 to 30), and thus replication studies are crucial for further validation. Stansfield (1991) reports correlations of between r = 0.89 and r = 0.95 for the scores of candidates on semi-direct and direct oral proficiency tests in various languages. However, Shohamy (1982, 1992) argues high correlations between the two types of tests may not provide sufficient evidence for test substitution i.e. they may not necessarily be measuring the same kind of language. She also suggests that direct tests may be more valid as face-to-face conversation is the dominant mode of oral interaction.

Feedback from test-takers on both types of tests is also relevant to an investigation of their comparability, with Stansfield (1991)

***Melbourne Papers in Language Testing* 1993 Volume 2.1 pp. 61-71. The Language Testing Research Centre, The University of Melbourne.**

finding that while most preferred the live format, about a quarter either preferred the tape format or had felt there was no difference. Where the semi-direct version was preferred it was because test-takers were not so nervous talking to a tape-recorder as to an interviewer. On the other hand, in our study, we found that our subjects strongly favoured the interview format. However, this may in part be due to the fact that all the candidates were taking the test as a trial, and there were no consequences resting upon their performance in the test. Stansfield also found that raters considered the tape version to be easier to rate because the format is consistent throughout, and it is predictable at what point on the tape the candidate will begin speaking. Our raters were in agreement with this.

Stansfield (1991) argues that given that the reasons for the administration of the test are very important, then it may be preferable to use a semi-direct test since the format remains the same and ensures control over reliability as well as the validity of the score. This is certainly a consideration with the **access: Australian assessment of communicative English skills**[1] oral interaction test where the control over any aspect of interlocutor behaviour is limited, since, while the test is developed in Australia, interlocutors are chosen in the country in which the test is undertaken. We are currently running a study of interlocutor behaviour as assessed by raters who are assessing the overseas administered tapes.

While this paper is based on an investigation of the comparability of the direct and semi-direct oral tests through quantitative statistical analyses only, we nonetheless take the view that qualitative analyses are essential in assessing the construct validity of two forms of oral test. This analysis should be considered to be a first step in the complex process of validating the two types of test, and further research involving detailed evaluation of the language samples obtained from the two versions of the test will be reported later this year. For the purpose of this study a number of research questions were addressed. We begin by investigating how well the two versions discriminate between

[1] This test is being developed under the aegis of the National Centre for English Language Teaching and Research at Macquarie University, Sydney, Australia, as a project sponsored by the Australian Commonwealth Department of Immigration and Ethnic Affairs

candidates and items. Our next concern is whether the overall item difficulty was the same across both versions of the test. This was investigated by correlating the difficulty measures for each matching item on the two versions of the test. The next issue examined considers whether the overall performance of candidates is correlated on the two versions of the test. Following this, we investigate the concurrent validity of the two versions of the test by comparing the performance of candidates in each version of the test with their rating on another well-established scale of language proficiency [ie. the ASLPR (roughly equivalent to the FSI scale) in the live version of the test][2]. This ASLPR rating was made on the performance of each candidate on the live version of the test. Finally we considered the degree to which the test method determined success or failure on the test.

Method

The Test

The two versions were developed in parallel. A series of matching tasks was developed which were intended to be similar in their language focus, although different in content. The exception was the final section which was identical on the two versions in order to have one point of direct comparabilty across both tests. A wide variety of task types, topics and language functions were chosen for each version of the test in an attempt to elicit a rich language sample from candidates. The types of tasks which appeared on each version of the test are outlined in Table 1 overleaf:

[2] The ASLPR can be assessed from a wide range of interview procedures

	Live	Tape
Description	✓	✓
Narration (picture sequence)	✓	✓
Exposition (data based)	✓	✓
Summary (of taped conversation)		✓
Role play	✓	
Telephone answering message		✓
Extended discussion	✓	✓

Table 1. Task Types

In phase one of this project, the tests were required to identify candidates at the vocational level which corresponds to ASLPR 3 (roughly equivalent to FSI 3). The findings in this paper result from the analyses of the first trial of this test which was held in Melbourne in December 1992.

In order to standardise the input of the interlocutor as much as possible, a booklet for interlocutors detailing the requirements of the tasks was developed for the live version of the test. The booklet provided the interlocutors with detailed instructions concerning language input to be used with candidates. Since this test will be conducted overseas with interlocutors who may or may not be trained language teachers, we considered it essential to reduce the potential variability in interlocutor behaviour to the greatest possible extent[3]. Test booklets were also provided to candidates for both the live and tape-based versions. Both booklets included the stimuli for the test tasks. In addition, in the tape version, the instructions for the tasks were written, as well as spoken, since the candidate would not have the advantage of clarifying misunderstandings with the interlocutor. The tape version was longer than the live version (45 minutes compared to 30 minutes).

[3] This measure is complemented by a video training package which all prospective interlocutors are required to view.

The scoring criteria adopted were identical for matching tasks and the other tasks in the two versions. The criteria used included fluency, grammar, vocabulary, coherence and cohesion, appropriacy of language, intelligibility, overall communicative effectivencess and, in the case of the live version, comprehension. Each of these criteria was assessed on a six-point scale with accompanying descriptors for each level.

A trial was conducted in which 94 candidates attempted both versions of the test. Of these, half were administered the live version first, and half were administered the tape version first. These performances were all audio-taped so that they could be rated retrospectively. Ten tape-based recordings were unsuccessful due to technical faults in the recording equipment in the language laboratory and these candidates were therefore excluded. (We have now minimised the potential for technical problems in this mode.) One additional candidate was excluded because the live version recording became unintelligible after only one rating. Thus this analysis is based on the direct and semi-direct version of the test from a total of 83 candidates.

Thirteen teachers were recruited for the purpose of rating the tapes. They were all trained teachers of English as a Second Language (ESL) and had considerable experience teaching a range of levels of ESL. Prior to carrying out this task, each participated in a comprehensive rater-training session. The rating process was conducted in two phases. Initially each tape was rated by two raters. Thus, for the first stage, nine raters (A–I) rated approximately 20 tape versions and 20 live versions each (see Table 2 below). The rating design was such that no rater assessed any particular candidate on either version more than once.

	LIVE		TAPE	
Candidate no.	Rating 1	Rating 2	Rating 1	Rating 2
1–10	A	G	F	D
11–20	B	H	G	E
21–30	C	I	H	F
31–40	D	A	I	G
41–50	E	B	A	H
51–61	F	C	B	I
62–72	G	D	C	A
73–83	H	E	D	B
84–94	I	F	E	C

Table 2. Rating Design (9 Raters)

In the second stage of the rating, a further four raters (J–M) rated every tape (83 x 2=166). Two of the raters rated the tape version first, and two rated the live version first.

All data were entered into a FACETS format for analysis (Linacre 1989–1992). There were three data files; one for the live data, one for the tape-based data, and one for the combined data. In the combined data analysis, the items were treated as a single test with items from the live test numbered from 1–23 and items from the tape test numbered from 24–47.

Rasch analysis (or Item Response Theory) has been very useful in the area of language testing because it provides information about candidate ability in relation to item difficulty. A new programme, FACETS, extends IRT to include rater characteristics (and other facets of the test situation) (Linacre 1989–1990). This ability level is expressed as the probability of a candidate obtaining a certain score on a particular task given the ability of the candidate, the difficulty of the item, the harshness of the rater and the effect of any additional facets (Linacre 1992). The decision to use the FACETS analysis was taken, in particular, because the programme compensates for variability in rater harshness in producing ability-estimates for each candidate. This analysis also provides an estimate of the relative difficulty of each of the test "items" i.e. each of the scoring criteria for individual tasks. In addition, the output from this program also provides information about misfitting candidates and items i.e. those items and candidates for whom the pattern of scoring is inconsistent with overall trend for the other items or candidates. In the analyses for this project five facets were included. These were (1) candidate (2) rater (3) tape or live (4) order[4] (5) items.

Results

The FACETS analysis indicated that both test versions taken either separately or combined, reliably distinguished a range of candidate abilites (see Table 3 below). For example, the figures for the live version reveal that the test reliability of person separation (the Rasch equivalent of conventional reliability estimates such as KR-20) was 1.00.

	Reliability Estimate
Live	1.00
Tape-Based	0.99
Combined	1.00

Table 3. Test Reliability Of Person Separation.

	Reliability Estimate
Live	0.98
Tape-Based	0.98
Combined	0.98

Table 4. Test Reliability Of Item Separation.

[4] Given that each candidate took both the direct and semi-direct versions, order indicates whether each particular version was taken first or second.

In addition, both test versions taken either separately or combined reliably distinguished a range of item difficulties (see Table 4 above).

A number of statistical analyses were performed subsequent to the FACETS analysis. The first of these was a Spearman rank order correlation between the logit scores obtained for the task items in each version. The correlation was significant but not especially high (**rho** = 0.797). However, examination of the item fit analysis revealed that there were several misfitting items i.e. the pattern of scores on these items was inconsistent with the pattern of scores on the majority of criteria. Raters had been required to attempt to measure both pragmatic and sociolinguistic skills through the criteria 'appropriacy' on two tasks in each version of the test. As well, raters were required to assess the degree of 'intelligibility' exhibited by candidates overall in each version of the test. All four appropriacy criteria, and both intelligibility criteria proved to be significantly misfitting. These were therefore removed from both versions of the test since we were concerned with measuring only items which discriminated effectively among candidates. In addition, rater feedback indicated that two of the tasks on each version of the test should be excluded on the grounds that they were perceived as problematic by the candidates, and in some cases by the interlocutors. In the original design of the test, it had been anticipated that items might need to be removed, and consequently the number of items included in the trial version was greater than those which would be required for the final administration. These tasks and items were removed and the resultant data sets were subjected to further FACETS analyses. A Spearman correlation was run on the resultant logit values obtained for the items. This improved the correlation significantly (**rho**= 0.952).

The crucial question concerns the relative validity of the two tests conditions. Current and future administrations of this test are likely to require that parallel versions of the tests be available since overseas centres which administer these tests cannot always avail themselves of both options. The central question, then, pertains to the ability estimate obtained for each individual candidate. It is essential that no candidate is disadvantaged or advantaged by the fact that s/he is obliged to undertake a direct test. We now focus on the candidates.

An estimate of ability (expressed as a logit value) was obtained for each candidate from both the direct version of the test, and the semi-direct version of the test. A Pearson's correlation run on the logit values obtained for each candidate on each of the two tests was r = 0.916. In addition, the intra-class correlation was calculated (r_I = 0.905). This is a measure of actual agreement as opposed to simple linearity.

In order to determine whether the strength of this correlation was unduly influenced by the fact that four of the raters rated all tapes, separate correlations were calculated on the logit measures from the rating from all four raters and, separately, the nine raters. The correlation for the four raters was r=.885 and that of the nine raters was r=.836. This suggests that the characteristics of the four raters was not disproportionately influential.

A further step in this investigation was to examine the concurrent validity of both versions of the test. This process involved measuring the strength of relationship between one or both of them and some other established external benchmark of language proficiency. To this end, the sets of ability estimates obtained from the Rasch analyses for each version were correlated with the ASLPR ratings. These had been provided by the interlocutor and an observer (both of whom were experienced users of this procedure) to candidates on the basis of their performance in the live version immediately after the interview. Spearman correlations were calculated and the results (**rho**= 0.89 for the live version and 0.87 for the tape-based version) suggest that both versions share a reasonably high degree of concurrent validity .

The final issue we investigated related to cut-off scores. We were concerned whether the same candidates who achieved a vocational level on one version of the test, also reached vocational level on the other version of the test. In this case the cut-off for vocational level was calculated for each version by entering the ASLPR scores as separate items for each candidate, and on both versions of the test. This provided a logit value equivalent to ASLPR 3 (ie the vocational level). This differed slightly for each version of the test and is a function of the overall item difficulty in each case. Thus the cut-off for the live version was 0.10, and that for the tape version was -0.10. For each candidate there were four possible options : pass live and pass tape; pass live and fail tape; fail live and pass tape;

fail live and fail tape. Table 5 summarises these results for the eighty-three candidates:

		TAPE	
		Pass	Fail
LIVE	Pass	53	7
	Fail	3	20

Table 5. Results For Candidates With Respect To Vocational Level

As can be seen from the table, there were only ten candidates who were problematic (ie. they passed on one version, whilst failing on the other. In every case the candidate failed the version of the test which they had undertaken first. To explain this, we suggest that the effect of order acted as an intervening variable for these candidates to the extent that it significantly influenced their second performance in a positive way. It is possible, and even likely, that this would have affected other candidates, who did not, however, stand out because they obtained the same level in respect to vocational level on both versions. We are currently investigating this issue further, in the light of some recent results we have obtained from a trial in which candidates undertook live and tape version of the test, where both versions contained identical items. A correlation run on the logit measures for the candidates on live vs. tape yielded a coefficient of .48, much lower than that for the parallel versions discussed here, and that which we had expected. We postulate that when the items on the test are identical, the practice effect will be so strong as to make meaningful comparison between the results on the two kinds of tests impossible.

Conclusion

The results of this study suggest that the two versions of the test were highly comparable in so far as candidates performed similarly and that test items were of similar difficulty in each case. However, there may be other aspects of test performance which differ but were not measured by the criteria used here. It should also be stressed that the similarity between the two versions may be a function of the fact that the live version lacks a certain interactive element due to the manner in which it was constructed. This may not be the case in other studies which examine this issue.

In general, it appears that candidates are not disadvantaged by sitting for either version. However, given the opportunity to 'practice', it may be that a certain percentage of candidates will improve their performance sufficiently to move from failure to success on the kinds of parallel versions used in this study. This is clearly an area that requires further investigation.

References

Linacre, J.M. (1989–1992) FACETS, a computer program for the analysis of multi-faceted data. Chicago : MESA press

Linacre, J.M. (1992) A user's guide to FACETS. Chicago : MESA press

Shohamy, E. (1982) Predicting speaking proficiency from Cloze tests : theoretical and practical considerations for test substitutions. Applied Linguistics, 3 : 161–171

Shohamy, E. (1992) The validity of concurrent validity of direct vs. semi-direct tests of oral proficiency. Paper presented at the AERA Convention, San Francisco, CA. April

Shohamy, E. & Stansfield, C. W. (1991) The Hebrew Oral Test : an example of international cooperation. AILA Bulletin, no 7

Shohamy, E., Gordon, C., Kenyon, D & Stansfield, C. (1989) The development and validation of a semi-direct test for assessing oral proficiency in Hebrew. Bulletin of Higher Hebrew Education, Vol. 4

Stansfield, C. W. (1991) A comparative analysis of simulated and direct oral proficiency interviews. In S. Anivan (ed) Current Developments in Language Testing. Singapore : SEAMEO RELC