The Testing of Oral Interaction:  
A Research Perspective

Michael Milanovic

Active research into the issues surrounding the testing of spoken language has been going on since the late 1970’s. It is interesting to note that the topic of the first Language Testing Research Colloquium, held in Boston in 1979, had as its theme ‘The construct validation of tests of communicative competence’. While the theme suggests a somewhat broad area of interest, the vast majority of the papers presented focused on the testing of speaking. Madsen and Jones (1981) were asked to review oral testing to date and in their paper ‘The Classification of Oral Proficiency Tests’ they wrote:

During the past few decades oral language testing has had a great deal in common with physical fitness. Everyone thinks that it is a wonderful idea, but few people have taken the time to do anything about it.

During the eighties interest increased and a substantial amount of research was carried out primarily in an American context, related to the Foreign Service Institute/Interagency Language Roundtable (FSI/ILR) interview and the ACTFL/ETS Oral Proficiency Interview for the most part. This work focused on:

- content validity (Lowe 1981) in relation to the FSI oral interview;
- construct validity either in relation to the FSI oral interview, (Bachman and Palmer 1981, 1983) or the ACTFL oral proficiency guidelines and oral interview procedure (Dandonoli and Henning 1990);
- reliability and rating procedures in relation to the Ilyin Oral Interview (Engelskichen, Cottrell and Oller 1981), a Hebrew oral interview (Shohamy 1983), and the ACTFL Oral Interview (Magnan 1987);

While covering a broad area, it is now clear that this work had a rather limited focus. Little attention was paid to the nature of candidate discourse, the task, the examiner or the nature of the ratings.

It was not until the late eighties that the focus began to broaden. This development is not limited to the testing of spoken language although it is this area that I will focus on here.

It now seems clear that we can consider the testing of spoken language across a number of facets, and that these facets interact in complex ways. The challenge for the future will be to identify them and develop ways of understanding, explaining and accounting for these interactions. The most obvious participant in the testing process is the candidate. Each candidate brings a complex set of variables to the testing situation. They relate at least to the following:

i. who the candidate is: age, gender etc.;
ii. what the candidate knows in general terms: knowledge schemata;
iii. how the candidate behaves and what s/he feels: affective schemata;
iv. how the candidate thinks: cognitive and metacognitive skills;
v. the sample of language elicited.

In order to understand and explain performance these features must be investigated. An instrument for the investigation of areas iii and iv outlined above has been developed in a joint UCLES/UCLA research project. While not directly related to the testing of spoken language it is as relevant there as in any other area of language testing. Bachman (1990) outlines three categories of influences on test score: personal characteristics, test method facets and random factors, and suggests that the study of the effects of personal attributes and test method facets on language test performance is an essential part of the process of construct validation. Research into the effects of these factors can also have important practical implications for controlling variation in and reducing bias due to these factors, thus improving the usefulness of language tests. This research project has as its aim the construction of a questionnaire bank that focuses on the following:
Area 1: Socio-psychological Factors
1. Attitudes;
2. Motivation;
3. Anxiety;
4. Effort.

Area 2: Strategic Factors
5. Cognitive strategies;
6. Metacognitive strategies;
7. Communication strategies.

Each of the Area 2 categories is further subdivided into a number of subcategories such as memory strategies, retrieval and using strategies, etc. Both qualitative and quantitative analyses have been carried out on the questionnaire bank and we are now in the process of designing research projects to investigate the relationships between candidate characteristic and performance.

Interest in the nature of the language elicited in oral interviews was stimulated by the publication of van Lier's paper in 1989. This paper questions our assumptions about the language of oral interviews and conversation. Some work has taken place since then. For example, the nature of the discourse in the University of Cambridge First Certificate in English oral proficiency interview, provided the basis for an investigation of a number of features of language produced by candidates and how candidate language varied in relation to theme, topic and gender (Young and Milanovic 1992).

Following the work of van Lier (1989), a model of dyadic native–non-native speaker discourse was proposed in which interview discourse is described in terms of three features: interactional contingency, goal orientation of participants and dominance. The first two are based on a model proposed by Jones and Gerard (1967). This model describes dyadic interaction in terms of two dimensions: contingency and goal orientation.

Interactional contingency relates to how people react to each other. It represents a way of looking at the local dynamic of how participants create shared meanings in any one exchange. It is seen as a property of adjacent turns in dialogue, when participants in an interaction search for mutually agreed meaning, as in a negotiation, or when superficially different topics are linked by conversational implicature. In the Young and Milanovic ver-
sion of the model, a contingent utterance is one in which the content and often the form of an utterance depend in some way on the previous utterance. In other words, the extent to which what is said in the interview context depends on or relates to what has been said before, and to what extent the discourse represents a series of disjointed speech events.

Goal orientation reflects the internal goals of each speaker separately. It focuses on how shared meanings evolve and change over longer stretches of discourse. Goal orientation is evident when a given topic persists over an extended number of turns, or when a speaker returns to a topic raised earlier after a number of intervening turns on a different topic. Goal orientation is particularly evident in interview situations when one speaker follows a predetermined agenda.

These dimensions are then combined in four styles of dyadic interaction. These are:

*Pseudocontingency*, where there is a high degree of goal orientation but little reactivity;
*Assymetrical contingency*, where there is a high degree of goal orientation by one party and a high degree of reactivity by the other;
*Reactive contingency*, where there is little goal orientation by either party but a high degree of reactivity by both;
*Mutual contingency*, where there is a high degree of reactivity and goal orientation by both parties.

In addition to the above, the dominance variable was introduced. Dominance may be defined as the tendency for one participant to control the discourse by various means such as holding the floor, initiating and terminating topics and controlling the other participants access to the floor by means of interruptions and questions.

The Young and Milanovic (1992) study set out to develop a model that allowed for the description of oral interview discourse in relation to the features described above. It provides an approach that allows for the comparison of oral proficiency interview discourse and other genres of spoken language. It also provides a way of relating variation in the structure of discourse in oral proficiency interviews to features of context such as theme of interaction, task and gender relationship of participants. The long term aim is to investigate the effects of these features on performance and score
obtained in oral assessments.

Related to the work described above, in that it focuses on the language elicited in the context of oral interviews, are the studies that draw on accommodation theory. In the early eighties Beebe and Zuengler (1983) looked at the ways in which aspects of phonology varied when different interviewees were paired with different interviewers. They found that there were significant differences in performance depending on the pairing and that these related to the extent to which individuals were prepared to accommodate to each other. Beebe and Zuengler's work was not specifically related to language testing contexts. However, Ross and Berwick (1992), Ross (1992) and Berwick and Ross (1993) relate the notion of accommodation to the oral testing of language. Ross and Berwick (1992) studied the talk produced in oral proficiency interviews (OPIs) as a product of native–non-native discourse in terms of features of control such as topic nomination and abandonment and reformulations, as well as features of accommodation such as clarification requests, display questions and simplifications. In another study, Ross (1992) argues that the nature of language produced by a candidate in an oral interview context is to some extent dependent on the nature of the accommodation of the interviewer. As a result he suggests that the final ratings assigned to candidates in an oral interview should take into account the amount of accommodation that occurred, so that the role of the interviewer in the interaction is included.

Just as candidates bring a complex set of characteristics to the oral testing environment, so do examiners. Their background knowledge and affective schemata are also likely to impact on the outcome of an oral interview. In a recently conducted study, Berwick and Ross (1993) look at the cross-cultural pragmatics in oral proficiency interview strategies. They found that a Japanese and an American interviewer had quite different communicative styles. Berwick and Ross claim that the Japanese style represented authority through attention to form and careful adherence to instructions. Whereas the American style focused on control through attention to content and the interviewees willingness to engage with issues. They suggest that if such differences are a function of culture, then the language generated in the interview context will differ dramatically, and thus the score a candidate gets will be based on a different sample of language.

Lazaraton (1993) examines the instructional language and verbal
prompts of examiners for the Cambridge Assessment of Spoken English (CASE). This study was carried out in order to develop a template for the assessment of oral examiner performance across time in relation to a highly specified interlocutor frame. This will assist CASE test developers and trainers to determine quickly and efficiently how well examiners individually and as a group are performing during training, certification, and live encounters. The template can provide feedback to examiners on how they might improve their interviewing.

58 CASE assessments were studied. They had been gathered at a live trial conducted in Tokyo in 1992. 10 examiners were involved in the study. All 58 assessments were transcribed according to conversational analysis conventions.

CASE is a paired assessment in three phases. The first phase, which was analysed in most depth in this study, is based on an interlocutor frame which is intended to act as a guide to assessors and provide candidates with the same amount of input and support. The analysis of the tapes and transcripts indicated that certain prompts were not used consistently or were worded in ways that deviated from the prescribed wording of the frame. The deviations from the interlocutor frame in the second phase of the interview were much less frequent and more tolerable. Gross deviations appeared to be a function of personality of individuals than a problem for all examiners.

The analysis of examiner language indicated that many features of natural conversation were present. This is a positive finding in that it implies that conversation of some sort is tapped by CASE. However, much of the speech modification that occurred was not at all consistent. This might mean that candidates are not given the same opportunity to perform, which affect the final score that they receive. Such a conclusion is, however, premature since it is not known that unequal opportunities, even if they do exist, affect score. Like Ross (1992) Lazaraton (1993) suggests that at this stage, it would be advisable to devote more time to examiner training. She suggests a focus on points like: refraining from distracting, intrusive or teacher-like behaviours that are inappropriate to this context e.g., laughing excessively, using evaluative markers, correcting responses; relying less on speech modifications to assist the candidate, that is excessive accommodation to the level of proficiency of the candidate.
So far I have looked at some of the considerations and research work carried out in relation to candidates and examiners. Attention has also been paid recently to the nature of rating instruments, that is, the number of scales, the length of scales and the nature of scales.

Opinions as to the number of scales to use in oral assessment have varied. Interviews like the FSI and ACTFL use a holistic approach. A single scale covers all the components of communicative competence implying perhaps that skill development has a flat profile. However, research has indicated that an individual's development of different facets of communicative competence might well not be at the same pace. For example, Schmidt (1983) followed the language development of Wes, a Japanese artist, over a four year period. He was able to demonstrate that Wes's sociolinguistic and strategic competence developed substantially over that period, but that his grammatical competence remained poor. A holistic assessment scale cannot readily cope with this type of situation. Similarly, Beebe and Takahashi (1989) show how pragmatic competence does not necessarily develop at the same rate as other interactional skills. Clearly there is a question in oral assessment about whether we adopt a holistic or analytic approach.

Once a scale, or series of scales, has been adopted, however, the question of how long they should be needs to be addressed. In general, scale length has been addressed intuitively until relatively recently. An empirical validation of scale length was undertaken by Milanovic et al. (1992) in relation to CASE.

CASE rating scales were developed along side the structure of CASE and in light of the model of communicative competence that was to be assessed. Essentially the scales break down into three main areas. These reflect grammatical, discourse and strategic competence. There are three grammatical scales: grammar, vocabulary and pronunciation. Each of these has five points. One scale is related to discourse competence (organisation) and another to strategic competence (communications skill and strategies). These scales have five points. Four scales focus on the success with which the four major tasks in CASE are achieved. These scales have four points each. Another scale, called interlocutor support, not dissimilar to Ross's (1992) notion of accommodation, is also included. This scale has four points. There is also a seven point scale which represents overall ability. This overall ability scale is an independent assessment and
not an arithmetic manipulation of the component scales. There are therefore eleven scales in all.

Data for the analysis was collected at a number of administrations of CASE. In all there were 70 candidates and 10 assessors involved. Inter-rater reliability was found to be high. In addition, we were able to accumulate evidence that the scales were indeed measuring different components of communicative language ability as represented by the model adopted. However, we were interested in the extent to which examiners were able to effectively distinguish between the different scale points. To this end we used Rasch partial credit analysis (Wright and Masters, 1982) which provides a means for the empirical validation of rating scales in a way that feedback from raters does not. The model describes the relationship between a person's ability and the difficulty of the scale in terms of the probability of an individual providing an adequate language sample to get a certain score.

In the study described here we were first interested in seeing whether it was equally difficult to get the same score in different scales. The table below shows clearly that a person is most likely to get a band 2 on the vocabulary and grammar scales at an ability level of -3 logits and below. It appears easier to get a band 3 on the pronunciation scale, and three band on that scale is rather wide.

CASE Analysis: Feb. 1992 R MODEL "BIGSTEPS" RASCH
INPUT: 70 PERSONS 11 ITEMS
ANALYSED: 68 PERSONS 11 ITEMS 47 CATEGORIES
MOST PROBABLE RESPONSES (BETWEEN "0" AND "1" IS "0", ETC.)

<table>
<thead>
<tr>
<th>ITEM</th>
<th>-7</th>
<th>-5</th>
<th>-3</th>
<th>-1</th>
<th>1</th>
<th>3</th>
<th>5</th>
<th>7</th>
<th>9</th>
</tr>
</thead>
<tbody>
<tr>
<td>GRAMMAR</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>VOCABULARY</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>PRONUNCIATION</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
<td>22</td>
</tr>
<tr>
<td>ORGANISATION</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>COMMUNICATION</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>SUPPORT</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>INTERVIEW</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>PRESENTATION</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>QUESTIONS</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>DISCUSSION</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
<td>111</td>
</tr>
<tr>
<td>OVERALL</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
<td>333</td>
</tr>
<tr>
<td>PERSON</td>
<td>-7</td>
<td>-5</td>
<td>-3</td>
<td>-1</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>7</td>
<td>9</td>
</tr>
</tbody>
</table>

Q 41352 231 224 22264 413 21 2 3 1 22
S M S Q
This may indicate that raters are hedging their bets to some extent when it comes to pronunciation. This is not a counter intuitive finding. We all know from experience that there is a tendency for pronunciation to fossilize. Perhaps the scale is a representation of this tendency. Interestingly, the bottom point on the grammar and vocabulary scales does not come into play with this sample. This may be an indication that the scales are too long, and that raters cannot differentiate at five points. It may mean that no one was at the one point in terms of these particular scales. Or, it may mean that the scale descriptions need clarification at the lowest ability levels. In any case, the analysis has provided us with a series of additional questions that we can now address.

Not only does this methodology provide a means of investigating scale length, as demonstrated above, it also provides a way of looking at the extent to which raters can effectively differentiate scale points. A graphic way of doing this is to examine the item characteristic curves (ICC's) for different scales.

ICC's show how the modelled probability of getting a particular point on...
the scale varies as a function of the ability of the people in question.

The above table illustrates the overall ability scale. It has six points ranging from 3–8. As can be seen from the table, the five points, from 3–7 are fairly clearly differentiated.

The 8 point was used but the two candidates it applied to were excluded from the analysis due to fit. We can clearly see from the overall ability scale that there is always a range of ability where a particular scale point is the most probable score. The degree of overlap between the five points is quite limited and the overlap is rarely more than one point on the scale.

Points 4 and 5 on the scale are the least discrete, while points 3, 6 and 7 are reasonably distinct. Point 6 seems to have the greatest spread, howev-

R MODEL "BIGSTEPS" RASCH ANALYSIS VER. 2.12
INPUT: 70 PERSONS 11 ITEMS
ANALYZED: 68 PERSONS 10 ITEMS 47 CATEGORIES
FOR GROUP "0", MODEL TYPE "R", ITEM NUMBERS: 2
PLOT CENTERED ON ITEM DIFFICULTY OF 2.38

GRAMMAR
er, the probability of an individual getting a 4 instead of a 6 is low. This particular ICC is reassuring in that it demonstrates raters' ability to distinguish the levels fairly clearly. We can feel confident that the scale is about the right length. It will be noted however, that the bottom points on the grammar and pronunciation scales have not been used at all. It may well be the case that only four meaningful differentiations can be made on these scales.

The peakedness of the curves is also an important feature of ICCs. The greater the peakedness, the higher the probability that an individual with an ability in line with the peak is being correctly assessed. In conjunction with the peakedness, it is important to consider the spread of the ICCs. The wider the spread, the greater the possibility of inappropriate scoring. The ICC related to grammar seems to indicate that the scale points are well differentiated.

However, the organization scale shows that the probability of achieving the second point on the scale is relatively low. This may be due to a lack of differentiation in the scale descriptor, it may indicate that a five point scale is too long, or it may suggest that the abilities measured by this scale has a rather strange step action. By this we mean that the ability may be either not present (1) or present to some reasonable degree (3) with little in between.

ORGANIZATION
Studies of this type demonstrate how techniques for the analysis of rating scales can be used in the validation of scales and oral tests. There has been relatively limited work of this sort in the development of operational measures to date. They can provide test developers with many useful insights and ammunition for continued work. It has not been the tradition in language testing to devote much time to the empirical validation of rating scales prior to tests being widely used. This needs to change in the future.

So far we have considered research that has been carried out on a number of facets in the assessment of spoken language: candidates, examiners, and rating scales. In addition to these, we might also look at the tasks used in oral assessment, how well they are defined, the extent to which they generate comparable language and so on.

However, we have not considered the interaction between the facets outlined above nor how we may investigate that. The ability of a candidate is mediated by the difficulty of a task and the severity of an examiner. Work on investigating these interactions empirically is really only just underway. Perhaps the most promising technique available to researchers at present is multi-faceted Rasch measurement (Linacre 1989). This type of analysis can be carried out through use of the programme Facets. Success on any given task is related to the performance setting, which is subdivided into a series of facets which would include the ability of the candidate and the difficulty of the task. In addition, characteristics of the rater and environment can also be considered. Facets has been used to address questions such as the stability of ratings over time (Lumley and McNamara 1993), the effects of interlocutor competence and audibility on rater performance (McNamara and Lumley 1993).

References


nual Language Testing Research Colloquium, Cambridge.


ABSTRACT

The Testing of Oral Interaction: A Research Perspective

Michael Milanovic

This paper reviews some of the research that has taken place in relation to the validation of tests of speaking. It considers four facets in some detail. These are the candidate, the examiner, the rating scales and the language elicited.

A number of research projects are reviewed in the paper. Work in progress by Cambridge and UCLA is focusing on the characteristics of the test taker with a view to investigating relationships between these and test performance and content. The language of the candidate is considered from a discourse perspective based on work by Young and Milanovic (1992). Lazaraton's work on examiner language (1993) in relation to CASE is considered and the development and analysis of rating scales is examined by Milanovic et al. (1992).

UCLES
Syndicate Buildings
1 Hills Road
Cambridge, CB12E
U. K.