This paper describes efforts to devise a more effective method of assessing oral proficiency in English using a computer-based testing and computer-based test result management system. This new test would be designed to directly operationalize the second language constructs it measured, and would be developed with reference to the best practices revealed by recent research in the field. Scores are automatically sent to databases at the completion and evaluation of each task. Student performance is then used to drive an adaptive algorithm, determining the difficulty of successive tasks. This computerized testing would aid in assigning skills or constructs to individual tasks, apportion the skills required for each task, and identify the features of interaction elicited in individual tasks. The study results revealed not only which tasks elicited most assessor/interlocutor interaction, but also that a great deal of the interaction took place outside the task boundaries. There has been a substantial move towards acknowledging the importance of interaction in oral proficiency and, therefore, oral proficiency testing featuring discourse and conversation analysis. The research detailed in this paper indicates that greater interaction does seem to be associated with greater oral proficiency, yet there are nagging issues of test validity still to be addressed in future research. (Contains 47 references.) (KFT)
This talk was offered at the CTELT Conference as it was thought to touch a number of poignant concerns. It was relevant to the themes and foci of the conference: 'effective use of technology', and 'computer-based testing'. It is presented here as two separate papers. The first deals with identifying the processes of NS-NNS interaction that take place during a locally-developed proficiency test which has a strong oral component. It also highlights the issues that have to be addressed if interactional features are to form a part of second language models which we are able to assess. The second paper dodges most of these issues, assuming with the general trend that OPIs are here to stay, and examines one specific aspect of nonverbal interlocutor support.

**PAPER I: INTERACTION AS A CONSTRUCT OF ORAL PROFICIENCY**

John D.E. Pollard
Paper I: Saudi Development and Training’s Five Star Proficiency Test Project.

This talk was offered at the CTELT Conference as it was thought to touch a number of poignant concerns. It was relevant to the themes and foci of the conference: ‘effective use of technology’, and ‘computer-based testing’. It is presented here as two separate papers. The first deals with identifying the processes of NS-NNS interaction that take place during a locally-developed proficiency test which has a strong oral component. It also highlights the issues that have to be addressed if interactional features are to form a part of second language models which we are able to assess. The second paper dodges most of these issues, assuming with the general trend that OPIs are here to stay, and examines one specific aspect of nonverbal interlocutor support.

Paper I: Interaction as a Construct of Oral Proficiency

The test development project in focus is an initiative to address the increasing pressures of localisation in the employment market, primarily in Saudi Arabia, but also in the Gulf area in general. A comprehensive, effective and reliable proficiency test was required as:

i. part of an 'assessment centre' approach to job recruitment

ii. part of a job-profiling tool to specify EFL entry and training requirements, and

iii. a preliminary placement test ahead of ELT and English medium training programmes.

The project was to take account of dissatisfaction clients had expressed with the results of indirect test formats such as those dominated by literacy-based (pencil-and-paper) discrete-point (multiple-choice) items.

The new test would be designed to directly operationalise the second-language constructs it measured, and would be developed – though within commercial constraints – with reference to principles of theory and research which represented best practice.

Initial surveys identified six relevant constructs, including Listening, Speaking, Reading and Writing. The fifth was Study Skills, defined as a sub-set of Reading which dealt with numeracy and the interpretation of lists, spreadsheets, graphs, and charts. The sixth was Interaction. This was included because it was seen to have a high prevalence in the target language use domain where one-to-one (but not necessarily face-to-face) encounters appeared to be the most common and highly valued format of NS-NNS events.

With the exception of Writing, the test was to be integrated into the single ‘event’ of a one-to-one oral proficiency interview-cum-discussion (OPI/D). The idea of integrating other language skills into OPI formats had first (to my knowledge) been muted in the 1980s - in Europe, by Nic Underhill, and in the USA by Leo van Lier:

...a well designed oral test which incorporates a number of different test techniques will give a quick and quite accurate measure of general proficiency. If desired, written or comprehension tasks can easily be built into such a test. (Underhill, 1987:12)

...different subparts of test batteries (Reading, Listening, Study Skills, etc) can all be included in a modular face-to-face session of no more than 30 minutes. (van Lier, 1989:505)

From the outset of the project in 1993 we considered computer-resourcing the tasks for such a process. Scores could be automatically sent to databases at the completion and evaluation of each task. Candidate performance could be used to drive an adaptive algorithm, determining the difficulty of successive tasks.

After preparing the way with surveys and task-trialing exercises, the prototype computer-resourced test was ready by the summer of 1994. (Pollard, 1994)
Just at this time a very large company requested a batch of proficiency assessments as part of a Saudisation selection process. From the perspective of a test developer, this was a premature move into high-stake assessment, where actual life-chances were being determined. However, the commercial pressures were overwhelming, and I could only urge caution and recommend procedures to ensure maximum reliability, as my decision-makers proposed the 'Five Star Test' to its client.

On the positive side, this expansion of use provided opportunities for piloting the test in an authentic environment. It has recently been pointed out that there are 'aspects of the validity of performance tests which can only be investigated once a test has become operational' (McNamara, 1996: 21). The company in question has main departments for Finance, Planning, Contracts, Construction, Manpower Resources, Personnel, Training and Staff Development, as well as support departments dealing with Staff Movements, Communications, Computer & Network Services, Maintenance, Security and Administration. All of these has a multinational workforce and clientele, so that English is the 'lingua franca'. This is typical of large commercial organisations in the Middle East, and as such provided an excellent site for initial research. The contractual agreement for reliable test results, however, meant that this had to be carried out with due caution.

Early requests to train in-company ELT staff in the use of the test were resisted, and most of the assessments were conducted by myself and a colleague who, though not from an EFL background, had been well inducted while programming the computer and working on the numerical mechanism to drive the scoring and reporting systems. He was a trained occupational psychologist familiar with counselling work and had all the necessary interpersonal skills. His thorough familiarity with the test and sympathy with the philosophy behind it was a big help.

Once sufficient tests had been conducted, a tentative enquiry was made into test-retest and inter-rater reliability. Under these very favourable conditions, it is not surprising that high point biserial correlations were obtained. These ran out at between 0.88 and 0.98, which, on the purely statistical basis of Pearson’s ‘r’, this indicated a minimal chance occurrence of 0.005. However, as we could only counterbalance our design adequately for 20 of the tests, which is too limited a sample to make robust claims, these results are best viewed with some caution. Variables would have multiplied if a more diverse group of assessors had been used, and reliability would have become a more complex issue, as recorded in the literature. (Barnwell, 1989; Ross & Berwick, 1990; Ross, 1992; Wigglesworth, 1993; Chalhoub-Deville, 1995). Reliability of assessments and consistency of interlocutor behaviour are notoriously difficult considerations where the roles of assessor and interlocutor are combined. They are, however, of enormous importance, and are considered in Paper II below.

By the middle of 1995, the test was demonstrating huge face validity for test-takers and test users. This, as we know, is an inadequate criterion for validity. However, the positive field-feedback helped us to obtain the funding for the more extensive research project described below.

Even with long-established proficiency tests predictive validity is difficult to demonstrate. For example, in a study carried out with candidates of the British ELTS test in the eighties, (now the IELTS test) scores were demonstrated to account for only 10% of the variance in later academic achievement. (Criper & Davies, 1988: 63) With a test such as Five Star which was being used very restrictively at this stage, no post hoc population with any statistical or sampling adequacy could have been provided.

For this reason, it was decided that an à priori enquiry into task constructs would be the most feasible method of gaining insight into validity at this stage. It would involve exposing the tasks to the judgement of a panel of independent experts. Although precedents for this can be found in the language testing literature, (Lumley, 1993) there have been cautions that expert opinion can be unreliable (Alderson, et al, 1995).

In order to eliminate the peer-group pressures and bandwagoning of open panel discussions, we therefore adopted a process known as a Delphi which allowed our experts to act as a
panel while retaining their anonymity. This research project was carried out at Sheffield Hallam University, UK and was co-ordinated by Nic Underhill. The panel consisted of twelve TEFL expert teachers working at SHU, all of whom had experience in the use of other OPI tests, including Cambridge UCLES FCE and the British Council IELTS. A special Delphi design was drawn up for the purpose by Dr Bunny La Roue; procedures to counterbalance for the order of task acquaintance and ensure equal task coverage were designed by Nic and myself. The research based on video data and interaction, was designed and piloted by myself in Riyadh. The project was split into three phases:

**Phase I:** Assigning skills or constructs to individual tasks

**Phase II:** Apportioning the skills required for each task  
(Reported in Pollard & Underhill, 1996), and

**Phase III:** Identifying the features of interaction elicited in individual tasks

For the present paper, I would like to focus on Phase III. However interesting the issues concerning methods of assessing 'unassisted' Listening and Speaking, Reading and Study Skills, there is a prevalent view that *interaction* is somehow fundamental to second language proficiency and its inseparable correlate, second language acquisition. This has recently been examined in a number of related branches of research, including:

- **Second Language Acquisition** (e.g. Færch & Kasper, 1984; Kramsch, 1986; Ellis, 1991)
- **Second Language Classroom Research** (e.g.; Chaudron, 1988; Long, 1983; Pica, et al 1989-1996; Johnson, 1995.)
- **Conversation Analysis** (e.g. Sacks, Schegloff, et al 1974-1995; Atkinson & Heritage, 1984; Jacoby & Ochs, 1995; Eggins & Slade, 1997)
- **Second Language Testing Research** (e.g. Shohamy, 1983-93; van Lier, 1989; Ross, 1992 & 1994; Ross & Berwick, 1992; Young & Milanovic, 1992; Zeungler, 1993; Young, 1994; Wigglesworth, 1994; Lazaraton, 1992 & 1996)

If we are to break away from idealised models of second language proficiency, it seems that the construct will have to include ability in the dynamic processes of real language encounters. The strongest expression of this view comes from the analysts of conversation, who claim that interaction is 'the primordial locus for the development of language, culture, and sense-making' (Jacoby & Ochs, 1994: 187)

Our working definition of the *Interaction* at the outset of this study was *'a learner's ability to facilitate participation in a one-to-one discussion through the employment of negotiation devices such as confirming understanding, requesting repetition and seeking clarification.'* This was derived from second language classroom interaction, as revealed in the work of Hatch, Long, Pica, et al cited above.

The construct, however, was omitted from the first two phases of the SHU research, as for some panelists the working definition was inadequate. They felt that interaction overlapped with Speaking and Listening and was therefore 'much harder to define' than these 'core skills'. Including it at this stage would have jeopardised the outset consensus between panel members, and hence the research methodology, which 'theoretically demanded independence between skills'. This is a reminder of the need to make compromises in order to further our understanding, but also echoes warnings that we may lose sight of the object of inquiry to 'preserve the integrity of the tools' we use in research designs (Lantolf & Frawley, 1985). If 'Interaction' necessarily overlaps with 'Listening' and 'Speaking', then it follows that 'Listening' and 'Speaking' necessarily overlap with 'Interaction'. The fact that interactional behaviour is difficult to separate from other areas should not, in itself, exclude it from our models of proficiency. However, this brief intra-panel debate highlighted a very important area of obscurity in our treatment of oral proficiency, and both of these papers reflect an attempt to better understand this hugely complex issue.
An additional reason for excluding Interaction from the early part of the inquiry was that in Phases I & II the panel had only examined test tasks. By the time Phase III got under way more than 500 assessments had been completed, and the panel were able to view a video-recorded samples. While they did so they completed observation sheets following a procedure which had been piloted with a group of EFL teachers in Riyadh. (The key and a sample of the observation matrix appear in Appendices I & II). The object of this was to find out (i) if a construct domain of interaction was salient to professional observers who might be typical of trainee assessors, (ii) if there were any patterns regarding the frequency and density of the specified interactional features within and between tasks, and (iii) if there was a significant contribution to the completion of test tasks by these features. No attempt was made to establish validity beyond these modest enquiries.

The huge questions raised about the generalisability of interaction in OPIs (van Lier, 1989) is arguably the biggest global validity question of all concerning this type of test (Messick, 1994). Such questions have only recently begun to be addressed in the case of widely used and long-established proficiency tests, as in the studies conducted by Young and Milanovic (1992), Young (1994) and Lazaraton (1996).

The first two of our questions were affirmed by the raw data, and the consensus at the end of the exercise was that ‘the Five Star Test can be seen centrally as a test of direct interaction between interlocutor and participant’. (Underhill, 1996)

The results revealed not only which tasks elicited most interaction, but that a great deal of interaction took place outside the ‘task boundaries’. For example, although pre-recorded Arabic instructions were used for the earlier, less challenging tasks, (to eliminate the ‘listening’ component when other constructs were in focus), later ones relied on the assessor/interlocutor explaining what had to be done. This is of great interest, as there is a sense in which these explanations represent the most authentic use of target language in the whole event. They were often sections of the test where the interactional features on the matrix had a high density of occurrence.

This not only identified sections of the test worthy of further analysis, but also influenced test and task design in the upgrade version which has now been developed. For example, some tasks are now ‘split’ so that the task explanation is offered in English as a Listening and Interaction task in its own right. An Arabic explanation back-up is available where the task procedure cannot be negotiated with the candidate’s English.

It has also led to split evaluations where the candidate has to explain an Arabic task instruction in order for the test to proceed – thus creating a quite natural ‘information gap’. This innovation not only achieves high levels of interactivity, but also reverses the roles of assessor and candidate in terms of topic and goal orientation. (Young & Milanovic, 1992)

The third question posed – the extent to which interaction contributed to the completion of tasks – has proven to be much more complex. It remains to be seen what transformations can be performed on the data to shed light on this. When means and standard deviations are applied to derive Z and T scores for criterion instances per unit of time, turn of specified dimensions, t-unit, etc. a clearer picture may emerge of the relationships between interactivity, task, and evaluation criterion. Procedures and processes for this have been explored in the context of observing second language classroom interaction (Chaudron, 1988: 17-24). However, this brings us close to huge questions yet to be answered by anyone. There has been a substantial move towards acknowledging the importance of interaction in oral proficiency and, therefore, oral proficiency testing. As indicated in the above citations, this started in the eighties. Since van Lier’s (1989) seminal article and the research it has generated, this has moved increasingly towards the areas of discourse and conversation analysis. The overriding impetus behind this is embedded in the interrelated issues of sampling and generalisability which are the fundamentals of validity. For test developers this opens up whole areas which will need to be re-assessed, ranging from theoretical justifications to actual methods and procedures for quantifying, measuring and reporting second language proficiency.

John Pollard, Riyadh, 19/01/1998
References:


Chaudron C, 1988: Second language classrooms: research on teaching and learning. CUP.


Young R, 1994: *Conversational styles in language proficiency interviews*. Language Learning 45/1: 3-42

**APPENDIX I**

**KEY**

◆ **CONFIRMS UNDERSTANDING**
  1. Offers appropriate response.
  2. Says "I understand" or =.
  3. Says "Yeh", "Yeah", "u-uh", etc.
  4. Agrees.
  5. Disagrees.

◇ **SEEKS CONFIRMATION**
  7. Asks "Do you mean + item?" or =.
  8. Repeats interviewer words/segments with questioning intonation.
  9. Refers with deixis - "You?", "Me?", "Here?", "This?"

■ **SEEKS CLARIFICATION**
  10. Says "I don't understand".
  11. Says "I'm sorry?", "Excuse me?" or =.
  12. Says "Please repeat" or =.
  13. Repeats part of words/segments from interviewer's turn with obvious uncertainty.

□ **INDICATES NEED FOR CLARIFICATION**
  14. Fails to respond/extended silence.
  15. Responds with disfluencies such as "errrrr.....", "ermmm.....".

○ **CONFIRMS OWN PREVIOUS TURN**
  16. Says "Yes" or =.
  17. Says "That's right" or =.
  18. "Yes, + item".
  19. Repeats item.

☆ **RE-FORMS OWN PREVIOUS TURN**
  20. Says "No, what I meant was (or =) + repeats/rephrases item".
  21. Rephrases item.
APPENDIX II

INTERACTION - SAMPLE OF COMPLETED OBSERVATION MATRIX (Numbers in cells represent the instances of occurrence recorded by observers):

Identify the features of verbal interaction used by the learner. For each occurrence of a listed feature, put a tick in the cell.

<table>
<thead>
<tr>
<th>TASK</th>
<th>VERBAL INTERACTION (facilitating self-understanding)</th>
<th>VERBAL INTERACTION (facilitating other-understanding)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1(</td>
<td>\text{CONFIRMS UNDERSTANDING})</td>
</tr>
<tr>
<td>4 Names</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>11 Numeracy</td>
<td>8</td>
<td>5</td>
</tr>
<tr>
<td>13 Al Harbis</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>15 Student Grades I</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>19 Student Grades II</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>23 Vehicles</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>24 Footballers</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>26 Kettle</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>28 Signs I</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>29 Fridge</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>50 Signs II</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>55 Kuwait City</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>62 Car ownership</td>
<td>8</td>
<td>1</td>
</tr>
</tbody>
</table>
I. DOCUMENT IDENTIFICATION:

Title: Interaction as a Construct of Oral Proficiency

Author(s): John D.E. Pollard

Publication Date: 19/01/1998

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, Resources in Education (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents

Permission to reproduce and disseminate this material has been granted by

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 1

Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) and paper copy.

The sample sticker shown below will be affixed to all Level 2A documents

Permission to reproduce and disseminate this material in microfiche, and in electronic media, for ERIC Collection subscribers only, has been granted by

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2A

Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only.

The sample sticker shown below will be affixed to all Level 2B documents

Permission to reproduce and disseminate this material in microfiche only has been granted by

Sample

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

Level 2B

Check here for Level 2B release, permitting reproduction and dissemination in microfiche only.

Documents will be processed as indicated provided reproduction quality permits.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Signature:

Printed Name/Position/Title: EFL Advisor, MR

Organization/Address: Saudi Development & Training
PO. Box 67275 Riyadh 11517 K.S.A

Telephone: 09/1702-490304
Fax:

E-Mail Address:

Date: 09/04/2000
III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

<table>
<thead>
<tr>
<th>Publisher/Distributor:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td>Address:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Price:</td>
</tr>
</tbody>
</table>

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

<table>
<thead>
<tr>
<th>Name:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Address:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

OUR NEW ADDRESS AS OF SEPTEMBER 1, 1998
Center for Applied Linguistics
4646 40th Street NW
Washington DC 20016-1859

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3595

Telephone: 301-497-4080
Toll Free: 800-499-3742
FAX: 301-953-0263
e-mail: ericfac@internet.com
WWW: http://ericfac.piccard.csc.com

(REv. 9/97)
PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.